



Proactive AI for Superintelligent Agents

Ke Yang^{1,2} Jiateng Liu¹ Jize Jiang¹ Chun Zhao³ Yufan Cao⁴ Shanshan Zhong⁵
 Yiyang Du⁵ Ruixuan Liu⁵ Xuying Ning¹ Dean Alvarez¹ Yuji Zhang¹
 Duo Zhou¹ Miri Liu¹ Yuhao Cheng¹ Hanyang Chen¹ Rui Yang¹ Zifeng Wang²
 Hanghang Tong¹ Jingrui He¹ Chenyan Xiong⁵ Chen-Yu Lee² Jiawei Han¹ ChengXiang Zhai¹

¹University of Illinois Urbana-Champaign ²Google ³Stanford University

⁴University of California, Berkeley ⁵Carnegie Mellon University

Abstract

AI agents are increasingly capable of reasoning, planning, using tools, and executing long-horizon tasks, yet many still rely on humans to identify worthwhile work and initiate it. Proactiveness adds the capacity to initiate purposeful action without an explicit request. We examine proactiveness as a capability that complements intelligence and alignment in the development of superintelligent agents. Research on proactiveness, however, remains fragmented across communities with different terminology, assumptions, and evaluation practices. We integrate its conceptual foundations and historical development across control, agent theory, ubiquitous computing, mixed-initiative interaction, and foundation-model research. We organize enabling mechanisms into six capacities: awareness, anticipation, agenda formation, arbitration, action, and adaptation. We synthesize evaluation approaches at the task, mechanism, and user levels. Across 13 application domains, including coding, autonomous research, robotics, healthcare, and education, we compare the initiative agents exercise with the authority they are granted. Finally, we outline future directions in capability development, safety and governance, and emerging application frontiers, grounding the development of calibrated proactiveness for superintelligent agents in the achievements and limitations of existing systems.

Keywords: Proactive AI, Superintelligent Agents, Self-Initiated Action, Long-Horizon Tasks

Project: <https://empathyang.github.io/Proactive-AI-for-Superintelligent-Agents/>

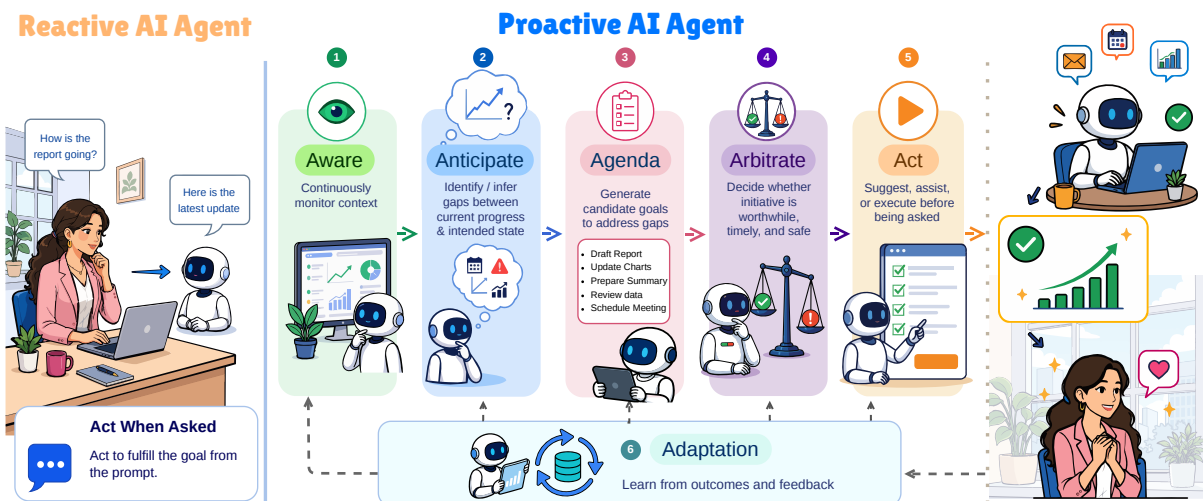


Figure 1 Proactive vs. reactive AI agents. Reactive agents rely on users to identify the work and request action. Proactive agents can recognize emerging needs and initiate useful work before being asked, learning from outcomes and feedback.

Contents

1	Introduction	5
2	Conceptual Framework	7
2.1	Defining Proactiveness	7
2.2	The Two Questions Behind Self-Initiated Action	7
2.3	The Reactivity-Proactivity Spectrum	8
3	History	9
3.1	Control and Cybernetics: Self-Triggered Regulation Under Fixed Goals	10
3.2	Planning and World Models: Simulating Futures Before Acting	10
3.3	BDI and Agent Theory: Initiative, Commitment, and Proactiveness	10
3.4	Ubiquitous and Context-Aware Computing: Maintaining the Situation Model	11
3.5	Mixed-Initiative Interaction: The Economics of Intervention	11
3.6	Goal Reasoning and Intrinsic Motivation: Generating the Agenda	11
3.7	Foundation Models: Recombination at Open-Ended Scale	12
3.8	Synthesis: From Historical Lineages to Mechanisms	12
4	Mechanism	13
4.1	Awareness	14
4.1.1	Multimodal State Tracking and Context Management	14
4.1.2	Active Perception and Proactive Information Seeking	16
4.1.3	Hierarchical Memory Systems	16
4.1.4	Runtime Architectures: Triggers, Always-On Monitoring, and Sleep-Time Compute	17
4.2	Anticipation	18
4.2.1	Modeling the User: Intent and Mental States	18
4.2.2	Modeling the World: Environment Dynamics and Future Prediction	19
4.2.3	Counterfactual Reasoning: Valuing the Intervention	20
4.2.4	Generative Scenario Simulation: Mental Sandboxing	20
4.3	Agenda	21
4.3.1	Goal Generation: Latent Needs to Explicit Goals	22
4.3.2	Sub-Goal Proposal and Long-Horizon Agendas	23
4.3.3	Goal Prioritization and Conflict Resolution	23
4.4	Arbitration	24
4.4.1	Uncertainty Quantification and Confidence Thresholds	24
4.4.2	Utility and Cost: The Value of Intervention	26
4.4.3	Timing and Interruption Management	26
4.4.4	Risk- and Reversibility-Aware Gating	27
4.5	Action	28
4.5.1	The Spectrum of Autonomy and the Agent's Interaction Paradigm	28
4.5.2	Action Presentation and Mixed-Initiative Delivery	30
4.6	Adaptation	31
4.6.1	Implicit Feedback and Self-Correction	31
4.6.2	Inference-Time Reflection	32
4.6.3	Long-Term Calibration, Personalization, and Trust	33
5	Evaluation	34
5.1	Task Utility Metrics	37
5.1.1	Success Metrics	37
5.1.2	Efficiency Metrics	38
5.2	Mechanism-Level Metrics	38
5.2.1	Context Modeling Efficacy	39
5.2.2	User State and World State Anticipation	39
5.2.3	Planning Quality	40

5.2.4	Intervention Decision and Timing	41
5.2.5	Intervention Delivery and Presentation	42
5.2.6	Adaptation and Generalization	43
5.3	User Experience Metrics	44
5.3.1	User Workload	45
5.3.2	Usability and Experience	45
5.3.3	Level of Control	46
5.3.4	Trust and Reliance	46
6	Applications	47
6.1	Coding Agents	47
6.1.1	Proactive Discovery of Unassigned Engineering Work	49
6.1.2	Engineering Agenda Formation and Prioritization	49
6.1.3	Risk-Aware Intervention and Coordination	50
6.1.4	Long-Running Project Stewardship	51
6.1.5	Evaluation, Challenges, and Open Directions	52
6.2	GUI Agents	52
6.2.1	Desktop GUI Agents	54
6.2.2	Web GUI Agents	55
6.2.3	Mobile GUI Agents	56
6.2.4	Evaluation, Challenges, and Open Directions	58
6.3	Information Access Agents for Search and Recommendation	58
6.3.1	Recommendation as Proactive Filtering	59
6.3.2	Proactive Search: From Zero-Query to Agentic Retrieval	59
6.3.3	Evaluation, Challenges, and Open Directions	60
6.4	Embodied Agents and Robotics	61
6.4.1	Household Services: From Smart Appliances to Proactive Domestic Partners	63
6.4.2	Autonomous Driving: Proactive Embodied AI at Commercial Scale	65
6.4.3	Industrial Collaboration: From Task Automation to Proactive Co-Workers	66
6.4.4	Human-Agent Co-Creation: From Virtual Imagination to Physical Creativity	67
6.4.5	Evaluation, Challenges, and Open Directions	69
6.5	Agents for Security and Surveillance	70
6.5.1	Choosing investigations and acquiring evidence	71
6.5.2	Turning findings into protective action	72
6.5.3	Sustaining protection and revisiting decisions	72
6.5.4	The agent as a target and source of risk	73
6.5.5	Evaluation, challenges, and open directions	74
6.6	Agents for AutoResearch	75
6.6.1	Autonomous Execution of Given Research Goals	75
6.6.2	Hypothesis Generation and Research Agenda Formation	76
6.6.3	Experimental Design and Resource Allocation	76
6.6.4	Experimental Oracles and Research-State Infrastructure	77
6.6.5	Evaluation, Challenges, and Open Directions	78
6.7	Agents for Smart Home	80
6.7.1	Ambient Household Orchestration	81
6.7.2	Trigger-Action Automation and Routine Synthesis	81
6.7.3	Energy and Comfort Management	81
6.7.4	Health Monitoring and Aging in Place	81
6.7.5	Evaluation, Challenges, and Open Directions	81
6.8	Agents for Smart City	82
6.8.1	Traffic and Mobility Orchestration	83
6.8.2	Grid and Utility Operations	83
6.8.3	Environmental and Public-Health Early Warning	83
6.8.4	Emergency Response and Civic Coordination	83

6.8.5	Urban Simulation and Digital Twins	83
6.8.6	Evaluation, Challenges, and Open Directions	83
6.9	Agents for Healthcare and Medicine	84
6.9.1	Clinical History Taking and Iterative Anamnesis	85
6.9.2	Multi-Specialist Deliberation and Case Conferencing	85
6.9.3	Longitudinal Care Coordination	86
6.9.4	Continuous Patient Monitoring and Biomedical Hypothesis Discovery	86
6.9.5	Evaluation, Challenges, and Open Directions	87
6.10	Agents for Law	87
6.10.1	Legal Research and Evidence Gathering	88
6.10.2	Argument Stress-Testing and Adversarial Review	88
6.10.3	Legal Matter Management and Workflow Execution	89
6.10.4	Compliance Monitoring and Legal-Risk Anticipation	90
6.10.5	Evaluation, Challenges, and Open Directions	90
6.11	Agents for Education	90
6.11.1	Learner-State Diagnosis and Anticipation	91
6.11.2	Intervention: What should the system do now?	92
6.11.3	Evaluation, Challenges, and Open Directions	93
6.12	Agents for Entertainment and Media	93
6.12.1	Acting Inside the World: Game and Embodied Agents	94
6.12.2	Curating What Users See: Feeds and Recommendation	94
6.12.3	Co-Creating Content and Performance	95
6.12.4	Evaluation, Challenges, and Open Directions	95
6.13	Agents for Agriculture	95
6.13.1	Field Sensing, Management, and Operational Coordination	96
6.13.2	Agricultural Risk Detection and Intervention	97
6.13.3	Evaluation, Challenges, and Open Directions	97
6.14	Cross-Industry Trends	97
6.14.1	Proactive Maturity Across Domains	97
6.14.2	Trend 1: A Shared Architecture with Domain-Specific Specialization	99
6.14.3	Trend 2: From Assisting Humans to Carrying Workflows	100
6.14.4	Trend 3: Verification and Feedback Infrastructure Is a Key Constraint	100
7	Future Work: Towards Calibrated Proactiveness	101
7.1	Steps to Take for Advancing Proactiveness	101
7.1.1	Evaluation Beyond Narrow Applications, Static Performance, and Coarse-Grained Metrics	101
7.1.2	Strengthening Calibrated Proactiveness Through Mechanism-Level Advances	102
7.1.3	Proactiveness for Long-Horizon Tasks and Agent Self-Evolution	106
7.2	Governing Proactiveness with Safety Boundaries and Trustworthiness	107
7.2.1	Safety Guardrails for Internalized Proactive Reasoning	107
7.2.2	Agent Trustworthiness, Transparency, and Appropriate Reliance	108
7.2.3	Drawing the Line: Consent, Agency, and Accountability	109
7.2.4	Moral Self, Model Self-Awareness, and Ethical Considerations	110
7.3	Envisioning Application Frontiers with Calibrated Proactiveness	111
7.3.1	Personal Agents	111
7.3.2	Proactiveness in Embodied Agents	112
7.3.3	Interconnected Proactive Ecosystems	113
7.3.4	Societal and Institutional Adaptation	113
8	Conclusion	114

1 Introduction

Recent advances in AI agents have expanded their ability to reason, plan, use tools, and execute tasks across multiple steps (Yao et al., 2023b; Yang et al., 2024a). Access to skills, persistent memory, and external services through agent harnesses further expands agents’ capabilities, enabling them to carry out sustained tasks across sessions (Zhang et al., 2025b; Young, 2025). Yet an agent may competently complete an assigned objective while still depending on a person to notice emerging needs, identify neglected work, and decide when an intervention is warranted. This dependence becomes consequential when considering the development of *superintelligent agents*. Greater intelligence expands what an agent can accomplish, and alignment concerns whether its behavior serves human intentions and values. Proactiveness adds the capacity to recognize worthwhile work and initiate it without an explicit request (Figure 1). We examine how this capability can complement intelligence and alignment, using the achievements and limitations of existing systems to inform the development of future superintelligent agents.

We use **proactiveness** to refer to the agent’s ability to identify worthwhile goals and initiate purposeful action without a new, explicit request for that intervention. Such initiative can operate within standing objectives, predefined responsibilities, and previously granted permissions. More precisely, proactiveness concerns the delegated discretion an agent exercises within a task’s underspecified residual decision space, while respecting its explicitly stated goals and constraints (Figure 2). A coding agent responds reactively when a user reports a bug and asks it to make a repair. The agent may inspect files, implement a fix, and run tests independently, but the user has identified the work to be done (Jimenez et al., 2024; Yang et al., 2024a). A proactive coding agent, operating under a standing responsibility for code quality, can detect a previously unreported bug and propose a repair before the user asks (Chen et al., 2025e). Similarly, a personal assistant might anticipate a scheduling conflict and prepare alternatives without waiting for a request. In these proactive cases, the agent identifies the need for intervention itself.

This capability is particularly relevant to long-horizon tasks. Maintaining a software project, coordinating ongoing care, or supporting a research program involves changing conditions, unresolved dependencies, and needs that cannot all be specified at the outset. A plan may become outdated, a previously minor concern may become urgent, or an opportunity may arise while attention is directed elsewhere. Supporting such objectives requires agents to preserve context, identify emerging work, revisit priorities, and decide when to resume, revise, or abandon a commitment (Aha, 2018). Proactiveness can therefore reduce the continuing burden on humans to monitor the environment and repeatedly translate changes into new instructions. Its value depends on whether these interventions advance the broader objective over time. Such initiative requires agents to identify what deserves attention and determine whether, when, and how to act within their standing objectives and granted authority.

Understanding proactiveness also requires separating capability from permission. As is shown in Figure 3, intelligence supports reasoning and execution; alignment concerns the relationship between an agent’s behavior and human intentions and values; proactiveness concerns its initiative in identifying and advancing useful work. None of these capabilities alone determines what the agent is authorized to do. An assistant may proactively identify a scheduling problem while leaving every calendar change to the user. Conversely, a system may receive broad permission to execute a task while relying entirely on the user to select that task. We examine both the initiative an agent exercises and the authority it is granted. Research on mixed-initiative interaction provides a foundation for studying how automated assistance can

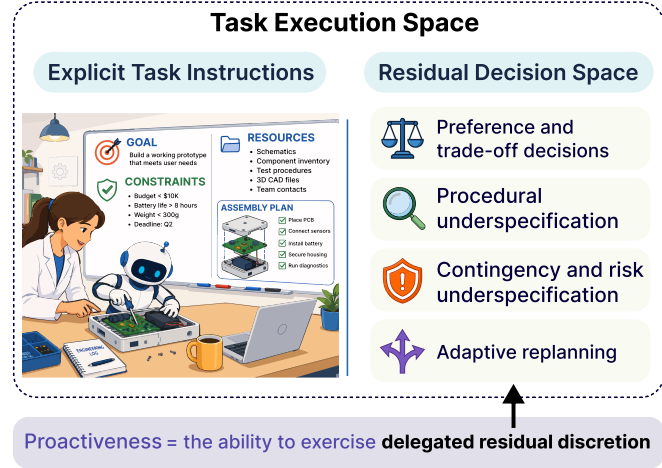


Figure 2 Proactiveness as delegated residual discretion. A task’s execution space comprises an explicitly specified component and an underspecified residual decision space. Proactiveness is the capacity to exercise delegated discretion within this residual space, independently determining and acting on what the specification leaves open without a new instruction for each decision. This includes resolving trade-offs, filling procedural gaps, handling contingencies, and replanning while respecting the task’s stated goals and constraints.

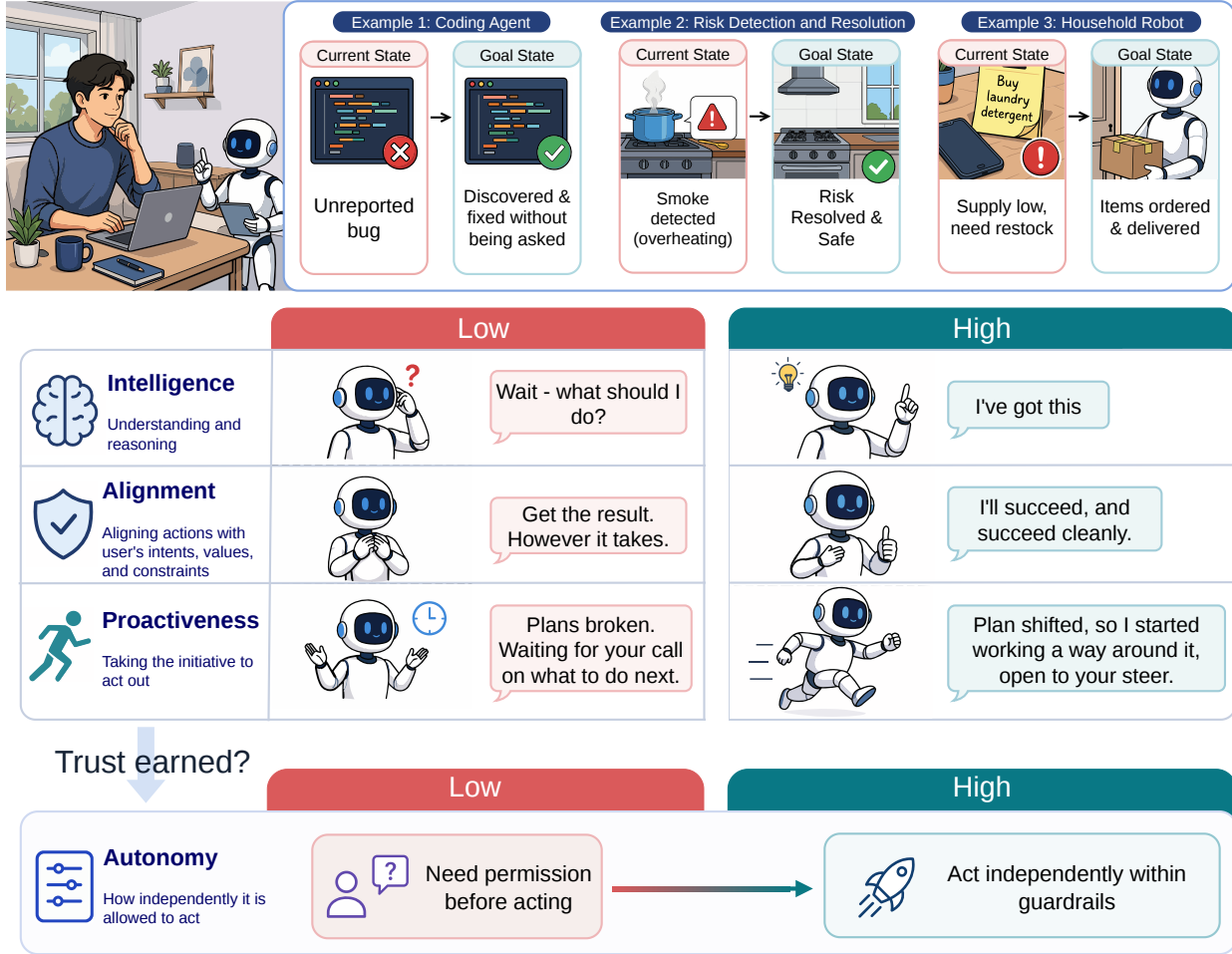


Figure 3 From complementary capabilities to earned autonomy. Intelligence provides the competence to reason and execute; alignment directs this competence toward human intentions and values; proactiveness enables its self-initiated use in the residual decision space. Reliable exercise of these complementary capabilities establishes a behavioral track record that can earn trust. Autonomy is the authority humans grant on the basis of demonstrated reliability.

accommodate user control and the costs of the agent's intervention (Horvitz, 1999).

Proactiveness has a substantial history in computer science. Classical agent research identified initiative in pursuing goals as a defining property of intelligent agents, while goal-reasoning research examined how agents select and revise their own objectives (Wooldridge & Jennings, 1995; Aha, 2018). Work on planning, context-aware computing, and human-computer interaction provides approaches to anticipating needs, maintaining goals, and managing intervention (Schilit et al., 1994; Horvitz, 1999). Ideas from philosophy, cognitive science, and control theory also help explain how purposeful action is organized. We draw on these ideas where they inform the conceptual foundations of proactive AI. Further, contemporary large language model-based agents provide opportunities to develop these capabilities through architectures that combine memory, reasoning, and action (Wang et al., 2024c).

This survey focuses on computer science research on proactive AI agents. Relevant work is distributed across agent architectures, planning and goal reasoning, context-aware computing, human-computer interaction, and more recent systems built on foundation models. These research communities often study related capabilities using different terminology, assumptions, and evaluation practices, making it difficult to relate their findings to the design of complete proactive systems. We organize this literature around the mechanisms that support self-initiated action and examine how those mechanisms are evaluated and used across application settings.

This survey makes three contributions:

1. **Conceptual foundations and historical synthesis.** We clarify proactiveness through self-initiation, explain how initiative relates to granted authority, and trace its development in computer science. Concepts from related fields inform this account of anticipation, goal reasoning, and human control.
2. **A unified mechanism framework.** We organize proactive behavior through six interacting capacities: awareness, anticipation, agenda (formation), arbitration, action, and adaptation. The framework connects maintaining context and identifying candidate goals with selecting interventions, delivering actions, and learning from their consequences. It provides a common basis for comparing architectures across domains.
3. **Evaluation and cross-domain analysis.** We synthesize evaluation approaches at the task, mechanism, and user levels and examine representative systems across 13 application domains. We compare how these systems exercise initiative, what authority they receive, and how their interventions are assessed. This analysis identifies gaps in the evaluation of task discovery and examines how reliable feedback and verification constrain progress across domains.

Furthermore, proactive agents must judge whether an intervention is useful, when it should occur (McFarlane, 2002; Amershi et al., 2019), and how far they are permitted to act. Acting too readily can interrupt users (Adamczyk & Bailey, 2004; Mark et al., 2008; Pu et al., 2025a), pursue mistaken goals, or exceed granted authority, while excessive hesitation can leave important needs unaddressed. We thus use *calibrated proactiveness* to describe the adjustment of initiative to uncertainty, risk, timing, user preferences, and permissions. This perspective guides our discussion of future evaluation challenges, the connection and distinction between proactiveness, long-horizon adaptation and self-improvement, and proactive AI’s safety and governance.

The remainder of this survey proceeds from conceptual foundations and historical development to the mechanisms of proactive agency. We then review evaluation practices and applications across domains, followed by future directions addressing capability development, safety, trust, and governance. The final section concludes the survey.

2 Conceptual Framework

2.1 Defining Proactiveness

A **proactive AI agent** is an autonomous system that initiates goal-oriented actions without a new, explicit request for that intervention. The center of this definition is **self-initiation**: the agent decides to act before being **explicitly** told to, and organizes that action around a goal it treats as worth pursuing.

Reactive systems set the baseline for the comparison. A reactive AI agent responds when triggered and stays idle otherwise, a design that is efficient, controllable, and straightforward to evaluate, since the user or the environment hands it the immediate task. However, when the user does not ask and no configured trigger calls for a response, an emerging need remains unaddressed.

Proactive agents differ in how far they look ahead before acting. A schedule or threshold can initiate a check, after which the agent may exercise discretion in deciding whether an intervention is warranted. More capable agents add an anticipatory model that forecasts future states, user needs, or downstream consequences, which lets them detect needs and risks before those become explicit. These forecasts can inform the goals an agent selects and the actions it initiates.

The goal a proactive agent pursues can come from many places. The user might state a high-level objective, for instance keeping energy cost down, holding code quality steady, or respecting a travel constraint, and the agent later initiates actions in service of that standing objective. In other settings the goal is fixed inside a specialized system, for instance preserving a safety margin, preventing equipment failure, or maintaining service quality. The proactive AI agents may even generate the goal they would serve on their own. Across these settings, proactiveness concerns the agent’s discretion in selecting and initiating an intervention within the applicable goals, constraints, and permissions.

2.2 The Two Questions Behind Self-Initiated Action

Self-initiated action requires the agent to judge what is worth acting on and whether, when, and how to intervene.

The first is **what to act on**. Within an assigned task or standing objective, the agent assesses current and anticipated needs, opportunities, and risks, and settles on something worth doing. This is inseparable from goal alignment: a

self-generated or mistimed goal that drifts from what the user actually needs reads as unnecessary or intrusive, however competently it is executed. Choosing the target is thus already part of the hard problem.

The second is **whether, when, and how to act**. Because the action is unrequested, the agent has to judge that an intervention is warranted, timely, and acceptable before making it, then shape and deliver it in a form the user will endorse. The agent also has to decide whether this is the right moment, whether it is confident enough, whether the action stays within the authority it has been given, whether the user would rather be offered a suggestion than have the agent execute on its own, and how assertively to present the result.

The task specification can settle some of these choices and leave others open. For choices delegated to the agent, self-initiation requires judgments about the value and form of an intervention under uncertainty. For this reason, proactiveness is a distinct capability with its own representational and evaluative demands.

2.3 The Reactivity-Proactivity Spectrum

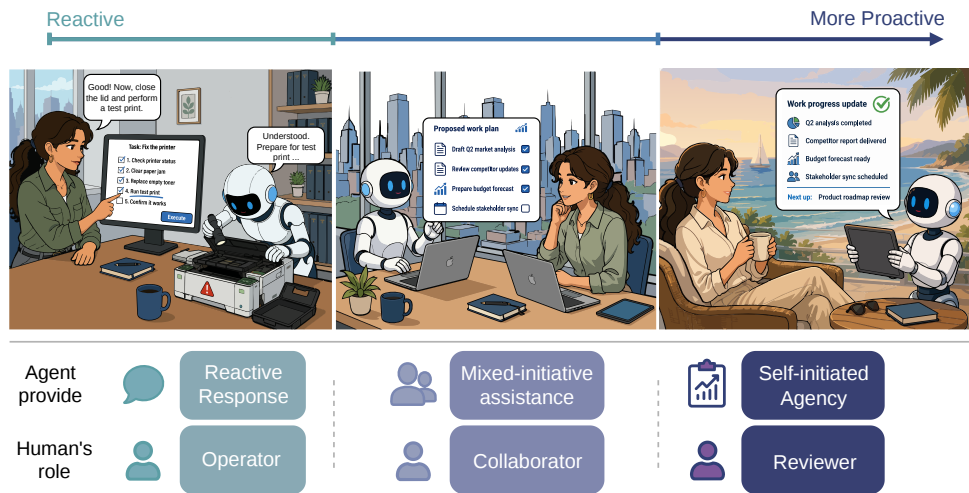


Figure 4 The reactivity-proactivity spectrum. The spectrum reflects how much of a task’s residual decision space an agent can manage without renewed instruction. Greater proactiveness shifts the human contribution from prescribing local actions toward negotiating priorities and reviewing progress. Reactivity remains necessary for incorporating new instructions and responding to changes, so the same agent can combine reactive execution with self-initiated work.

As shown in Figure 4, reactivity and proactiveness mark the two ends of a spectrum of behavior, and most systems fall somewhere between them. Many mix triggered response with self-initiated intervention, and the same AI agent can move along the spectrum depending on time scale, confidence, and operational context. The useful question about a deployed system is usually how much initiative it is allowed to take, under what conditions, and with what safeguards.

Most fielded systems are accordingly hybrid, pairing a reactive layer for time-critical response to explicit inputs and fast-changing signals with a proactive layer for deliberation and foresight over longer horizons. For example, an autonomous vehicle can brake or steer in response to a pedestrian stepping off the curb, a nearby car braking hard, or a lane-level hazard. During the same journey, it may assess traffic forecasts and battery reserves, identify a need for an additional charging stop, and propose adding that stop to the route without a new request from the passenger. A coding assistant can respond reactively by completing the current line, answering a question about an API, or applying an edit exactly where asked. It can exercise task-scoped initiative by identifying additional changes needed to complete the assigned task, and initiate new work outside the current assignment by proposing a refactor, flagging an unreported bug, or recommending a dependency update under a standing maintenance objective. Personal-productivity assistants and industrial-monitoring systems sit at their own points on the same spectrum, each described by where its initiative falls and how that initiative is bounded.

3 History

This section traces proactive agency as a gradual expansion in the decisions delegated to machines: from reactive systems that respond to explicit requests or configured triggers, to proactive agents that infer what is worth acting on and decide whether, when, and how to act without an explicit request. Earlier research communities approached these problems under different names (Figure 5). Control theory studied self-triggered regulation around fixed objectives; symbolic planning and world modeling studied anticipation before action; agent theory studied initiative and commitment; context-aware computing studied persistent situation modeling; mixed-initiative HCI studied intervention under uncertainty; and goal-reasoning and intrinsic-motivation research studied how agents generate and revise their own goals. Foundation-model agents matter not because they invented proactiveness, but because they recombine these mechanisms in open-ended, language-mediated environments (Wiener, 1948; Fikes & Nilsson, 1971; Wooldridge & Jennings, 1995; Schilit et al., 1994; Horvitz, 1999; Aha, 2018). This history provides the background for Section 4 and shows which parts of proactive agency predate current LLM systems.

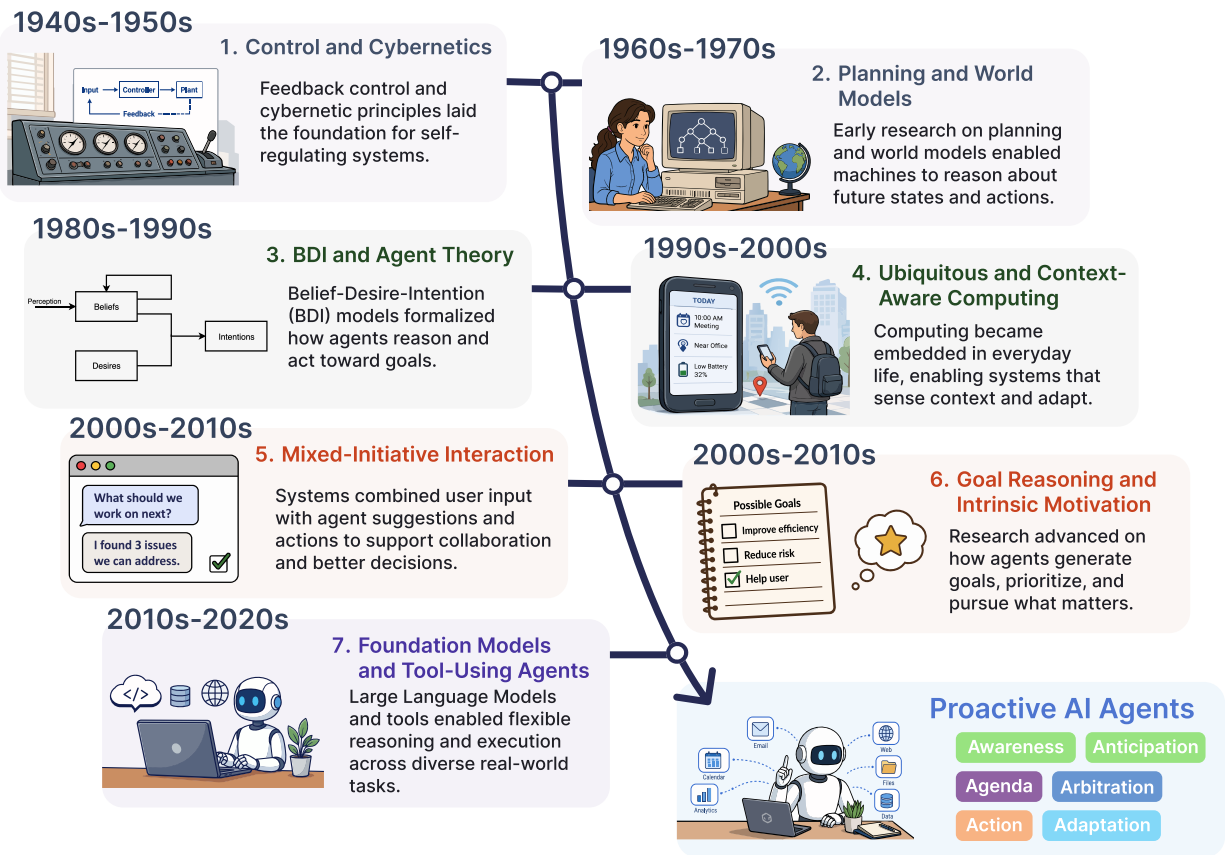


Figure 5 Historical foundations of increasingly open-ended initiative. Proactive agency draws on complementary traditions that progressively addressed what instructions leave unresolved: future consequences, persistent goals, changing context, and the value of intervention. Their development expands the decisions an agent can make beyond a fully specified control objective. Foundation models integrate and generalize these earlier mechanisms, continuing this cumulative development of proactive capabilities.

This expansion can be understood through the economic logic of incomplete contracts. Economists have long observed that contracts cannot specify every future contingency, which is why governance, authority, and ownership matter (Coase, 1937; Simon, 1951; Williamson, 1975, 1985). Modern incomplete-contract theory formalized this problem in terms of residual rights of control: the right to decide what should happen when the original agreement is silent (Grossman & Hart, 1986; Hart & Moore, 1988, 1990; Hart, 1995). A prompt, instruction, or standing objective can be read as a contract-like delegation from a human to a machine. It specifies some goals, constraints, and permissions, but it cannot describe every future state the agent may encounter. When such an unspecified state arises, the key question

is who holds the residual control right: should the machine fill the gap, or should control return to the human? Early proactiveness was limited not because the idea of machine initiative was absent, but because delegating these residual decisions to machines raised hard problems of state uncertainty, goal alignment, authority, and accountability. The history below traces how different research traditions gradually made more of these decisions delegable, from planning and state estimation to intervention timing and goal formation.

3.1 Control and Cybernetics: Self-Triggered Regulation Under Fixed Goals

The theoretical roots of proactive agents can be traced to *cybernetics*, the study of goal-directed systems (Rosenblueth et al., 1943; Wiener, 1948). Its central mechanism, the *feedback loop*, showed how a system can compare its current state with a desired state and adjust its behavior to reduce the gap. A thermostat is the simplest example: it senses temperature, compares it with a reference value, and automatically turns heating on or off to maintain the target. This *perceive–compare–act* cycle supplied the earliest formal model of self-triggered goal maintenance and helped shape the later *perceive–reason–act* view of intelligent agents. In the terminology introduced in Section 4, cybernetic control supplies the minimal substrate for **Awareness**, **Action**, and **Adaptation**: the machine can track the current state, detect a gap from a standing objective, and adjust its behavior without waiting for a new command.

3.2 Planning and World Models: Simulating Futures Before Acting

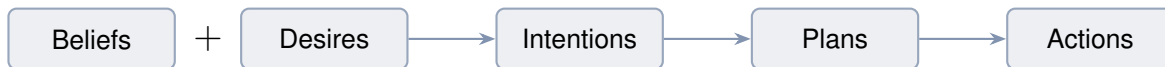
Cybernetics explained how a system could regulate behavior around a standing objective. Early AI and cognitive science pushed this idea further by treating intelligent behavior as symbolic problem solving and *planning* (Newell & Simon, 1976). The key shift was that agents were no longer limited to correcting present deviations. They could represent an internal model of the world and use it to reason about possible action sequences before execution.

STRIPS was a milestone in this transition (Fikes & Nilsson, 1971). It formalized planning in terms of an initial state, a goal state, and a set of actions with preconditions and effects. By representing the world as logical predicates and actions as operators over them, STRIPS gave agents a computational form of “what-if” reasoning. This established an early basis for anticipatory behavior, but only in a limited sense. The goal was usually supplied by a user, task designer, or environment specification. Classical planning therefore explained how an agent could reason toward a given objective, not how it should decide which objective to adopt.

Later world-model research generalized this anticipatory role from symbolic operators to learned dynamics. Model-based reinforcement learning and forward models formalized how an agent can predict the next state from the current state and a candidate action (Sutton, 1991; Jordan & Rumelhart, 1992; Wolpert et al., 1995; Moerland et al., 2023). Learned latent world models made such prediction scalable by rolling future states forward in compact representations (Ha & Schmidhuber, 2018; Hafner et al., 2025). These systems strengthened the **Anticipation** side of proactive agency: an agent could estimate downstream consequences, compare acting with abstaining, and detect failures before they became explicit.

3.3 BDI and Agent Theory: Initiative, Commitment, and Proactiveness

Classical planning gave agents a way to reason about future action, but usually toward a goal supplied from outside. The next major conceptual step came from philosophy. In 1987, Michael Bratman’s theory of intention formalized how human agents structure their behavior through interconnected mental states (Bratman, 1987):



Intentions serve as a bridge between immediate motivation and long-term planning, ensuring both *future-directedness* and *coherence* in action. This insight inspired AI researchers in the 1980s and 1990s to operationalize Bratman’s theory as a computational framework: the **BDI** (Belief–Desire–Intention) architecture (Cohen & Levesque, 1990; Rao & Georgeff, 1995).

The BDI model comprises three interdependent mental attitudes and their dynamic interactions: **Beliefs** represent the agent’s informational state about itself, the environment, and others; **Desires** encode motivational states, or the situations the agent wishes to bring about; and **Intentions** represent deliberative commitments to a subset of desires, which the agent pursues through planned sequences of actions. By embedding *commitment* into its reasoning cycle, the BDI model achieved a balance between proactivity and reactivity: it maintained goal stability amid dynamic environments while

remaining responsive to significant contextual changes. This commitment–reactivity trade-off remains a cornerstone of modern proactive architectures.

The intelligent-agent literature then made proactiveness a canonical property of agents. Wooldridge and Jennings characterized agents not only by autonomy, social ability, and reactivity, but also by proactiveness: the capacity to take initiative in pursuit of goals rather than merely respond to environmental change (Wooldridge & Jennings, 1995). This literature also clarifies a distinction that remains easy to blur: proactiveness concerns whether the agent can originate initiative, while autonomy concerns how much authority it has been granted to act with less supervision. For our mechanism account, the BDI lineage supplies the historical root of **Agenda**. A predicted need, opportunity, or risk becomes proactive behavior only when the agent commits to it as something to pursue.

3.4 Ubiquitous and Context-Aware Computing: Maintaining the Situation Model

Planning and BDI explain how an agent reasons toward goals and commits to them, but a proactive agent also needs grounding before any prompt arrives. This is the contribution of ubiquitous and context-aware computing. Weiser’s vision of ubiquitous computing placed computation in the background of ordinary activity, while early context-aware systems used information such as location, identity, activity, time, and device state to adapt behavior without explicit commands (Weiser, 1991; Schilit et al., 1994). Later work sharpened context awareness as the ability to use such situational information to provide relevant services or behavior (Abowd et al., 1999).

For proactive agency, this lineage further supplies the historical root of **Awareness**: the agent must maintain a model of the user and environment rather than wait for the user to specify what matters. Such a model can include location, calendar, current task, workflow history, preferences, sensor readings, screen state, organizational setting, and environmental risk. It is the substrate on which later **Anticipation** operates, because an agent cannot forecast a latent need without first estimating the current situation.

3.5 Mixed-Initiative Interaction: The Economics of Intervention

Context-aware systems gave machines a model of the user’s situation, but situation awareness does not by itself justify intervention. The mixed-initiative tradition shifted the question from whether a system can take initiative to when that initiative should enter human activity. In a mixed-initiative system, control can move between user and machine: the system may suggest, clarify, or take over part of the work, and control may later return to the user. Horvitz framed this movement as a decision under uncertainty, where the expected value of assistance must be weighed against the costs of error, interruption, and reduced user control (Horvitz, 1999). Tennenhouse’s proactive-computing agenda pushed a related vision of systems that continuously sense and act under human supervision (Tennenhouse, 2000).

The key contribution of this lineage is that initiative is not automatically valuable. A correct suggestion can still be harmful if it arrives at the wrong moment, removes the user’s sense of control, or accumulates false positives that erode trust. Interruption-management research made these costs measurable: interruptions impose recovery time, increase errors, and vary sharply with timing, with task boundaries often providing more acceptable moments for intervention than mid-task disruption (McFarlane, 2002; Adamczyk & Bailey, 2004; Iqbal & Horvitz, 2007; Bailey & Konstan, 2006). Attention-aware systems and personalized interruption models then treated the cost of a notification as a quantity to be estimated rather than a presentation detail (Horvitz & Apacible, 2003; Horvitz et al., 2004).

For proactive agents, this literature supplies the historical root of **Arbitration**. Anticipation may identify a latent need, but Arbitration decides whether that need justifies action now, at what level of autonomy, and through what form of delivery. The agent must weigh confidence, expected benefit, timing, reversibility, and granted authority before acting. Mixed-initiative interaction therefore provides the “whether, when, and how strongly” side of proactive agency.

3.6 Goal Reasoning and Intrinsic Motivation: Generating the Agenda

The preceding lineages explain how agents sense the present, simulate the future, commit to intentions, and gate interventions, but they leave one burden only partly specified: where do candidate goals come from when no user has supplied them? Goal reasoning addresses this question directly by asking where goals come from and how they should be revised. Goal-driven autonomy studied agents that monitor expectations, detect discrepancies between expected and observed states, explain those discrepancies, formulate new goals, and manage the resulting goal set (Molineaux et al., 2010; Klenk et al., 2013; Vattam et al., 2013; Aha, 2018). This loop is proactive by construction: the trigger for a new goal is not an external command, but the agent’s recognition that the world has departed from what it expected. A failed

plan, missing resource, emerging threat, or newly available opportunity can become a goal the agent adopts before the user has named it.

A complementary lineage comes from intrinsic motivation and open-ended learning. Goal-driven autonomy generates goals from discrepancies in the world, while intrinsically motivated agents generate objectives from internal signals such as curiosity, novelty, uncertainty reduction, or learning progress (Barto et al., 2004; Oudeyer et al., 2007; Schmidhuber, 2010). Prediction-error curiosity drives agents toward states whose outcomes they cannot yet forecast, while learning-progress signals steer them toward goals near the frontier of current competence (Pathak et al., 2017; Forestier et al., 2022). These systems show that self-generated goals need not arise only from anomaly repair. An agent can form an agenda because a region of the environment is informative, learnable, or worth exploring. Modern open-ended agents inherit this pattern: Voyager uses an automatic curriculum to propose new skills, while autonomous research agents generate hypotheses, experiments, or research directions as part of longer discovery loops (Wang et al., 2024a; Lu et al., 2024). What matters here is that these agents do more than carry out long tasks. They also begin to supply the **Agenda** that guides what should be pursued next.

3.7 Foundation Models: Recombination at Open-Ended Scale

The foundation-model era brings the earlier lineages together. A foundation model is a model trained on broad data at scale and adaptable to a wide range of downstream tasks (Bommasani et al., 2021). Large language and multimodal foundation models extend this idea by providing a common representational substrate for goals, preferences, constraints, plans, memories, tool calls, and explanations. This substrate lets mechanisms that were historically developed in separate communities operate inside a single agent architecture: context tracking from ubiquitous computing, future simulation from planning and world models, commitment from BDI, agenda formation from goal reasoning, intervention gating from mixed-initiative interaction, and adaptation from feedback.

Contemporary agents built on foundation models illustrate this recombination. Language agents interleave reasoning and tool use, decompose natural-language objectives, maintain memory, simulate future steps, and revise plans over extended interaction (Yao et al., 2023b; Sumers et al., 2024; Park et al., 2023). Open-ended agents generate curricula or subgoals, proactive assistants infer latent needs from context, and autonomous research systems generate and test hypotheses across multi-step loops (Wang et al., 2024a; Lu et al., 2025a, 2024). This generality also expands the stakes of initiative. A narrow proactive system may adjust a thermostat or issue an alert, while an agent built on foundation models may email a person, edit a repository, schedule a meeting, operate a browser, call an API, or change a production system. The central question therefore becomes sharper: when should the agent move first, how far should it go, and when must it ask, defer, or remain silent? Foundation models make proactivity more general, and for that reason make calibrated proactivity more necessary.

3.8 Synthesis: From Historical Lineages to Mechanisms

The current LLM-driven wave of AI renews a much older agenda. Long before foundation models, proactiveness had already appeared across cybernetics, symbolic AI, cognitive science, automation, operations research, and human-computer interaction. These traditions studied different parts of the same problem: how a system senses a situation, reasons about the future, commits to goals, evaluates the value of intervention, and coordinates its initiative with human control. Many of these ideas now reappear as core ingredients in the development of more general agents. Progress on proactive agency will therefore require more than larger models. It will require a renewed integration of the fields that have studied initiative, decision, control, and human agency from different angles.

Seen from this perspective, proactive agency is a history of increasingly incomplete delegation. Early control systems operated under nearly complete specifications: a fixed objective, a narrow state variable, and a small action set. Later agents were asked to fill in more of what the instruction left open, from action sequences and future contingencies to situation models, intervention timing, and self-generated goals. Proactiveness is the capacity to fill these gaps; autonomy is the authority granted to act on the resulting judgments. As agents gain real authority through information, prediction, memory, and tool use, calibrated proactivity becomes the problem of keeping that real authority aligned with the authority humans have formally granted (Aghion & Tirole, 1997). The six-A mechanism in the next section can be read as the machinery that makes such incomplete delegation possible.

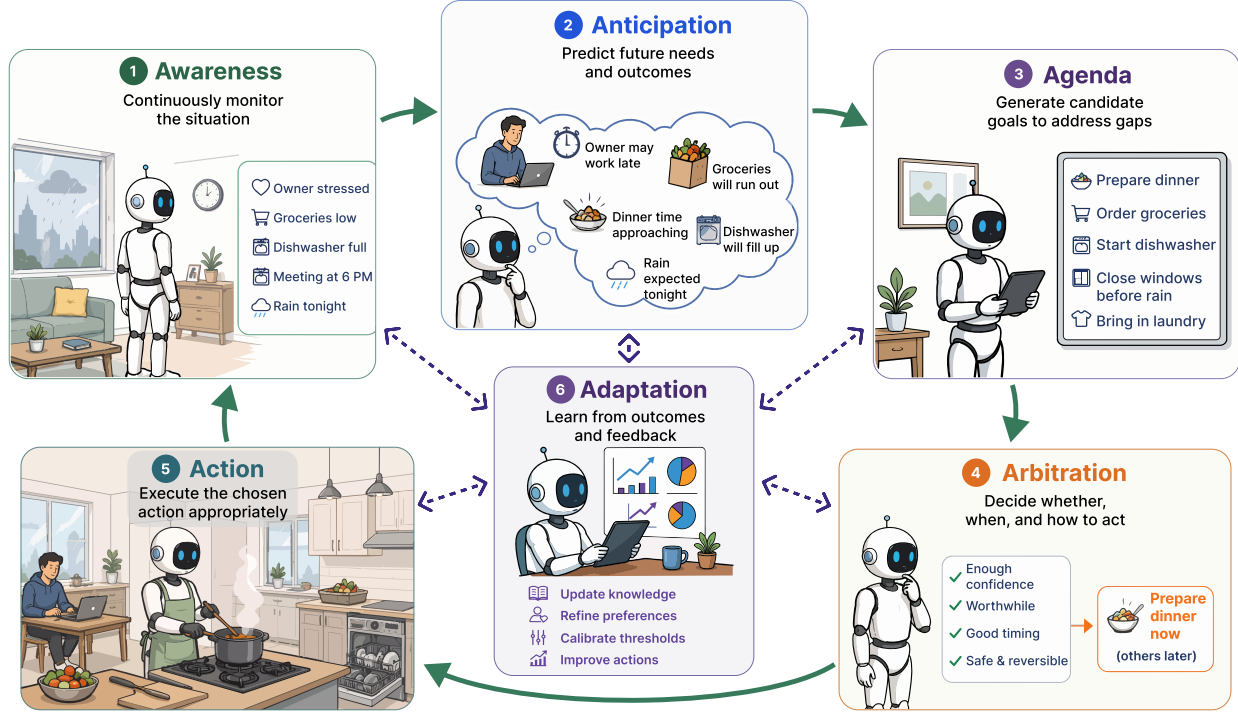


Figure 6 Six interacting capacities for exercising proactive discretion. Awareness maintains a situation model that enables Anticipation to identify future needs and predict the consequences of possible actions. Agenda converts these forecasts into prioritized commitments, and Arbitration determines whether and when to intervene. Action carries out the selected intervention, while Adaptation uses outcomes and feedback to update the other capacities.

4 Mechanism

We organize the mechanisms of proactive AI agents into six interacting capacities, abbreviated the *six A's*: Awareness, Anticipation, Agenda, Arbitration, Action, and Adaptation (Figure 6). Within this framework, Anticipation and Arbitration address two distinctive demands of self-initiated work: identifying needs and opportunities for work that no explicit user request or task instruction has specified, and deciding whether, when, and how to initiate an intervention within granted permissions.

Awareness, Agenda, Action, and Adaptation have established roles in general AI agents. However, their proactive forms can require substantial changes: maintaining context before a request arrives, generating and revising goals alongside assigned tasks, delivering unsolicited interventions with appropriate user control, and learning from feedback on the agent’s own initiative. Proactive behavior emerges through the interaction of all six capacities. Table 1 organizes representative works by these six mechanisms and their subdimensions.

Table 1 Taxonomy of the 6A mechanisms for proactive agents, with representative works.

Mechanism	Sub-dimension	Representative Works
Awareness (§4.1)	Multimodal State Tracking & Context Management (§4.1.1)	ProAgent (On-Demand Sensory) (Yang et al., 2026a), MIRIX (Wang & Chen, 2025)
	Active Perception & Proactive Information Seeking (§4.1.2)	Ask Early, Ask Late, Ask Right (Gulati et al., 2026), Asking Before Acting (Chen et al., 2024b)
	Hierarchical Memory Systems (§4.1.3)	MemGPT (Packer et al., 2023), Generative Agents (Park et al., 2023), PlugMem (Yang et al., 2026c)

Continued on next page

Table 1 (continued)

Mechanism	Sub-dimension	Representative Works
Anticipation (§4.2)	Runtime Architectures (§4.1.4)	Sleep-time Compute (Lin et al., 2025), Heartbeat Scheduler (Su, 2026b), Agent.xpu (Wei et al., 2026b)
	Modeling the User (§4.2.1)	Tell Me More! (Qian et al., 2024), JIR-Arena (Yang et al., 2025d)
	Modeling the World (§4.2.2)	Ha & Schmidhuber (World Models) (Ha & Schmidhuber, 2018), Qwen-AgentWorld (Zuo et al., 2026)
	Counterfactual Reasoning (§4.2.3)	Planning for Proactive Assistance (Kulkarni et al., 2021), CF-GPS (Buesing et al., 2018)
Agenda (§4.3)	Generative Scenario Simulation (§4.2.4)	Talking with 1,000 Personas (Xuan et al., 2026), ProMediate (Liu et al., 2026q)
	Goal Generation (§4.3.1)	Goal-Driven Autonomy (Molineaux et al., 2010), AI Scientist (Lu et al., 2024)
	Sub-Goal Proposal & Long-Horizon Agendas (§4.3.2)	Long-term Task-oriented Agent (Shi et al., 2026b), PEPA (Liu et al., 2026j)
	Goal Prioritization & Conflict Resolution (§4.3.3)	Instruction Hierarchy (Wallace et al., 2024), Corrigibility (Soares et al., 2015)
Arbitration (§4.4)	Uncertainty Quantification & Confidence Thresholds (§4.4.1)	KnowNo (Ren et al., 2023), Ask or Assume? (Edwards & Schuster, 2026), DiscussLLM (Silence Token) (Patel et al., 2026)
	Utility-Cost Analysis (§4.4.2)	Horvitz (LookOut) (Horvitz, 1999), Attention-Sensitive Alerting (Horvitz et al., 2013b)
	Timing & Interruption Management (§4.4.3)	SHANKS (Chiang et al., 2025), Adamczyk & Bailey (Adamczyk & Bailey, 2004), ProMemAssist (Pu et al., 2025b)
	Risk- & Reversibility-Aware Gating (§4.4.4)	ToolEmu (Ruan et al., 2024), Attainable Utility Preservation (Turner et al., 2020b), Variable-Autonomy by Criticality (Panagopoulos et al., 2022)
Action (§4.5)	The Spectrum of Autonomy & Interaction Paradigm (§4.5.1)	Sheridan & Verplank (10 levels) (Sheridan & Verplank, 1978), Situational Autonomy (Salikutluk et al., 2024), Feng et al. (Levels of Autonomy) (Feng et al., 2025)
	Action Presentation & Mixed-Initiative Delivery (§4.5.2)	Magentic-UI (Mozannar et al., 2025), Less Interaction But More Explanation (Jang & Sundar, 2026), Sensible Agent (Lee et al., 2025b)
Adaptation (§4.6)	Implicit Feedback & Self-Correction (§4.6.1)	Grounded in Reality (Wei et al., 2025a), ProactiveAgent (Lu et al., 2025a), CDHF (Mozannar et al., 2024b)
	Inference-Time Reflection (§4.6.2)	Proactive Conversational Agents (Inner Thoughts) (Liu et al., 2025g), Agentic Metacognition (Xu, 2025)
	Long-Term Calibration, Personalization & Trust (§4.6.3)	Lee & See (Trust in Automation) (Lee & See, 2004), PPP (Sun et al., 2025b)

4.1 Awareness

Before an agent can anticipate anything, it must hold an estimate of what is currently the case (Figure 7). Awareness is the process that builds and maintains that estimate, and we take it up in four parts: tracking state from multimodal streams into a persistent situation model (§4.1.1), deciding what to sense or ask about next when that estimate falls short (§4.1.2), organizing what has been learned into a hierarchical memory that later prediction can draw on (§4.1.3), and running a cognitive loop that can fire without an incoming query (§4.1.4). For proactive agents, Awareness must support the discovery of work before a request arrives. This changes what the agent monitors, which context it retains, and when it gathers more information, with sensing and memory organized around continuing needs across tasks and sessions.

4.1.1 Multimodal State Tracking and Context Management

A reactive system can interpret each request as it arrives; an agent acting without a new request uses observations and retained context to assess the current situation. A persistent, regularly updated model of user and environment state

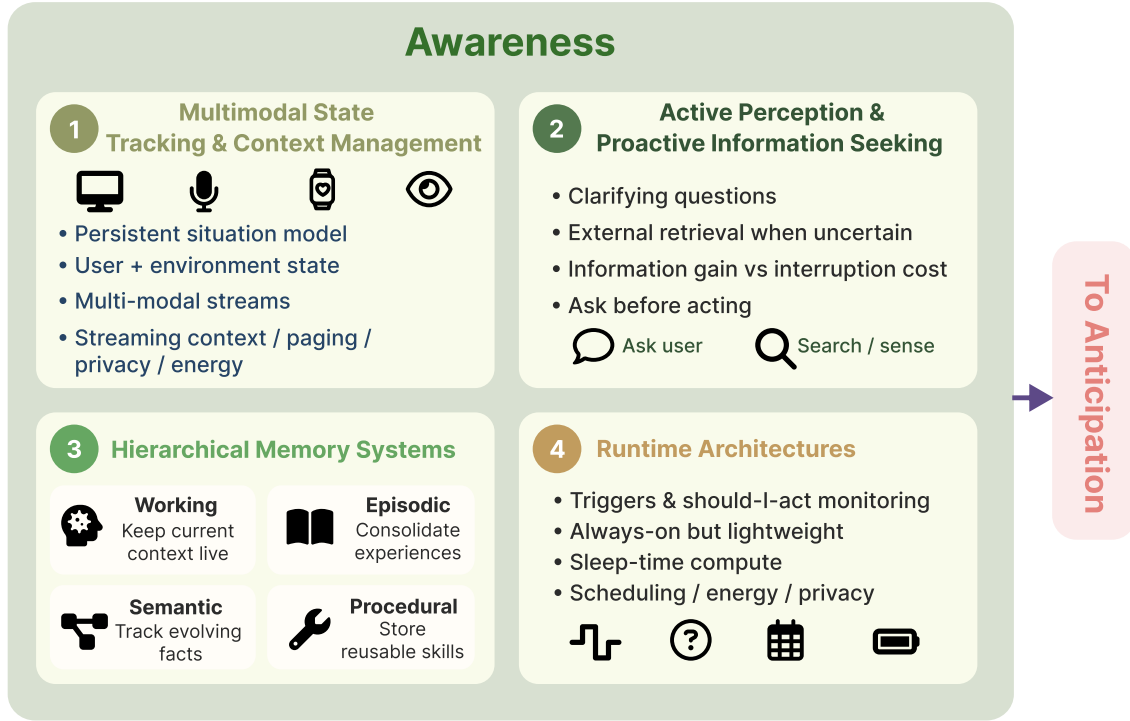


Figure 7 Awareness grounds initiative in decision-relevant context. When instructions leave decisions unspecified, the agent must determine what is happening, what prior experience remains relevant, and what information is missing. Persistent memory and self-invoked monitoring make this understanding available between requests, while selective information seeking closes consequential gaps. The central trade-off is maintaining sufficient context for sound decisions while controlling the costs of computation, energy, privacy, and user attention.

can support this assessment, an approach that predates current foundation-model agents. Context-aware computing in ubiquitous computing built ambient models of a user’s situation and used them to reconfigure behavior automatically (Schilit et al., 1994; Abowd et al., 1999). Dialogue state tracking is the longest-running formalization of maintaining a belief over a user’s goals as an interaction unfolds, running from the Dialog State Tracking Challenge through the MultiWOZ corpus to modern neural trackers (Williams et al., 2013; Budzianowski et al., 2020; Jacqmin et al., 2022).

Proactive agents generalize this recipe along two axes at once: from a fixed schema of slot-values to open-world state, and from discrete dialogue turns to continuous multimodal streams. Early ambient assistants already monitored a person’s state with non-intrusive sensors and intervened when a predictive model of work pressure and exhaustion flagged trouble (Bosse et al., 2013), and operational decision-support agents maintained a running model of the environment in order to deliver the right information to the right person at the right time (Buchler et al., 2014). The state model is only as good as the perception that fills it, and that perception is a general capability the proactive agent inherits from the broader perception literature: egocentric vision (Grauman et al., 2022), wearable activity recognition (Zhang et al., 2022; Wu et al., 2026d), and screen and interface understanding (Baechler et al., 2024), each a partial and noisy view that has to be fused into one estimate. Feeding always-on streams into a bounded model context creates a systems tension that streaming and context-paging architectures manage (Chen et al., 2024a; Xiao et al., 2024; Packer et al., 2023); we defer the memory side of this tension to §4.1.3. One signal in this stream is natively anticipatory and worth singling out: gaze precedes the action it serves, so eye movements give an early read on intention that feeds prediction directly (§4.2) (Belardinelli, 2023; Wilson et al., 2025). Recent proactive systems already assemble these streams, fusing wearable video and audio into the context that drives proactive predictions (Yang et al., 2025b).

Recent systems are turning this research aspiration into working prototypes, with early commercial features starting to appear. Continuous screenshot and sensor capture is folded into queryable personal memory by systems such as MIRIX (Wang & Chen, 2025), with commercial timeline-recall features such as Microsoft Recall, Screenpipe, and Plaud NotePin serving as industrial exemplars of the same idea (Microsoft, 2025b; Screenpipe, 2026; PLAUD, 2026); native interface agents track screen-state transitions as their world state (Qin et al., 2025c); and mobile agents tie on-device

capture to user context for proactive suggestion (Yang et al., 2026e; Li et al., 2024c). The cost of capture that never stops recurs across all of them. Tiered sensing keeps cheap sensors always on and wakes expensive ones only when relevance is flagged (Yang et al., 2026a), and privacy-sensitive domains push the whole pipeline on-device (Ji et al., 2025). This privacy-and-energy bill is the standing tax on awareness that the gating discussion (§4.4) and the systems discussion (§4.1.4) return to.

4.1.2 Active Perception and Proactive Information Seeking

Choosing what to sense is itself a decision problem. Active perception couples sensing to action selection in vision and robotics (Bajcsy, 1988; Bajcsy et al., 2017), and active learning and Bayesian experimental design choose queries to maximize expected information gain (Settles, 2009; Chaloner & Verdinelli, 1995). These classical formulations supply the vocabulary, information gain and expected value of perfect information, that modern “should I ask?” heuristics implicitly rediscover, and recent agent work imports the formal apparatus explicitly, casting question selection as sequential Bayesian experimental design or as expected-information-gain maximization over a maintained belief state (Choudhury et al., 2026; Echarghaoui et al., 2026).

When the missing information lives in the user’s head, the active-sensing act becomes a clarifying question. An unsolicited clarification claims the user’s attention and must pass the same arbitration gate (§4.4) as any other intervention. The design problem has two parts. Selectivity governs when an ambiguity is worth an interruption, the question that systems for selective clarification and ambiguity detection address (Kuhn et al., 2023a; Zhang et al., 2024g; Qian et al., 2024); utility governs which question buys the most information, which is expected value of perfect information in conversational clothing (Aliannejadi et al., 2019; Rao & III, 2018). The literature has matured from prompted heuristics to trained asking policies that learn when and what to ask from interaction or offline logs (Li et al., 2023a; Andukuri et al., 2024; Wei et al., 2025a; Sun et al., 2025b). The same pattern recurs wherever underspecified instructions meet costly downstream execution: open-domain coding agents, text-to-CAD specification, and dynamic-rule compliance screening each report agents that clarify before they commit (Edwards & Schuster, 2026; Yuan et al., 2026a; He et al., 2025a), and timing matters as much as content, since the value of a clarification depends sharply on when it is asked and which kind of information is missing (Gulati et al., 2026).

When the missing information is external, the same gate becomes an in-generation retrieval decision: the agent monitors its own uncertainty and searches only when its parametric knowledge is insufficient, the pattern established by the adaptive-retrieval line (Asai et al., 2023; Jeong et al., 2024; Jiang et al., 2023; Press et al., 2023). Agentic search and evidence-seeking variants extend uncertainty-guided retrieval to reasoning models and to domains such as pathology and recommendation (Li et al., 2025e; Hua et al., 2025; Zhang et al., 2026a).

In embodied and interactive settings the question carries a price: it interrupts a person, spends a turn, or stalls the task, so information seeking has to be budgeted against task progress. Ask-before-acting policies gate a query on whether the missing information would change the next decision (Chen et al., 2024b; Wang et al., 2026k); expert-query budgets are spent sparingly under learned uncertainty estimates (Singh et al., 2022; Igbinedion & Karaman, 2024; He et al., 2025b); question sequences are planned to maximize expected information gain (Hu et al., 2024; Fang et al., 2026c); and the completion-versus-interaction balance is optimized (Luo et al., 2026b), with that trade-off cast as an explicit Pareto frontier (Yao et al., 2026). The same budgeting drives collaborative navigation and joint planning, where every dialogue turn with a human is a cost to minimize (Taioli et al., 2025; Zhang et al., 2024h; Ognibene et al., 2025; Fang et al., 2026d; Houtti et al., 2025). Asking for clarification can resolve uncertainty before further action and gives the user an opportunity to guide that action (§4.5).

4.1.3 Hierarchical Memory Systems

Cognitive science provides the structural basis for agent memory. The episodic and semantic distinction (Tulving et al., 1972) and the separation of working memory from long-term declarative and procedural stores in cognitive architectures (Anderson et al., 2004; Laird et al., 2017; Nuxoll & Laird, 2007) carry over almost directly into the memory modules of modern agent frameworks (Sumers et al., 2024). The four memory stores are working, episodic, semantic, and procedural.

An agent that monitors a streaming situation cannot let its context window fill passively. Staying current means continuously deciding which elements of the live situation map are kept resident and which are paged out or expired,

the operation that operating-system-style virtual context management supplies (Packer et al., 2023) and that the agent-memory surveys catalog in general (Zhang et al., 2024k; Hu et al., 2026f). For a proactive agent the workspace doubles as the buffer that demand detection reads, so streaming systems pair a working store with a need detector that fires on it (Xie et al., 2026), and long-running agents cap workspace size to keep always-on cognition affordable (Zhang, 2026a).

Anticipation (§4.2) works from recurring patterns, and those patterns have to be built before they can be used. Episodic consolidation is what turns raw interaction history into them, through retrieval weighted by recency, importance, and relevance together with reflection, linking, and decay (Park et al., 2023; Xu et al., 2025d; Zhong et al., 2023; Chhikara et al., 2025; Tan et al., 2025). Recent systems make the proactive payoff concrete, distilling interaction trajectories into reusable experiences that drive unprompted assistance (Wang & Jiang, 2026; Fiorini et al., 2025) and folding event streams into structures that surface what the user will need next (Wang et al., 2026g; Yang et al., 2026c).

For a proactive agent a stale preference is especially expensive, because acting on it produces an unprompted wrong intervention, and an unsolicited error costs more trust than a solicited one. Long-term stores therefore track not only what is true of the user but when it stopped being true, through persona and preference stores (Li et al., 2025c), temporal knowledge graphs that invalidate superseded facts (Rasmussen et al., 2025), and timestamped repositories that detect contradictions and update on human correction (He et al., 2025a). The memory surveys frame this as the temporal dimension of agent memory (Wu et al., 2025f), and LongMemEval quantifies the cost of getting it wrong, reporting roughly a 30% accuracy drop over sustained interaction (Wu et al., 2025b); grounding suggestions in a group’s accumulated history is the multi-user form of the same requirement (Wang et al., 2025e).

Deliberating from scratch makes every intervention expensive. Consolidating repeatedly successful action sequences into reusable skills lets proactive behaviors fire habitually at low marginal cost, the procedural-memory class that cognitive architectures and their agent descendants identify (Anderson et al., 2004; Sumers et al., 2024). Agents build executable skill libraries (Wang et al., 2024a), induce reusable workflows from past traces (Wang et al., 2025o), compile mastered goals into fast low-level policies (Carta et al., 2025), and store synthesized strategies for proactive dialogue (Kim et al., 2025b). Semantic memory holds who the user is and procedural memory holds what the agent already knows how to do; together they feed anticipation (§4.2) and execution (§4.5).

4.1.4 Runtime Architectures: Triggers, Always-On Monitoring, and Sleep-Time Compute

A chatbot’s cognition is invoked when a query arrives; a proactive agent has to decide *when to wake up*, and that decision cannot itself be a response to a query. Classical lineages decoupled triggering from acting but with fixed, hand-authored conditions: event-condition-action rules in active databases (Paton & Díaz, 1999), trigger-action programming in the smart home (Ur et al., 2014), and interrupt-driven deliberation in belief-desire-intention architectures (Rao & Georgeff, 1995), with doctrine-driven monitors in operational decision support showing the same shape at scale (Buchler et al., 2014). The modern question is what replaces the hand-authored trigger, and adaptive schedulers that decide when to engage which cognitive activity are one emerging answer (Su, 2026b).

The contemporary pattern is a continuous event stream paired with a lightweight should-I-act judgment. ProactiveAgent learns this judgment from human-labeled acceptance data and fine-tunes agents to offer assistance unprompted (Lu et al., 2025a). The recipe instantiates across settings: sensory-context polling on wearables and smart glasses (Yang et al., 2025b, 2026a; Liu et al., 2026n), agent-formulated trigger conditions registered against a changing environment (Shi et al., 2026b), intent-conditioned monitoring backed by long-term memory (Xie et al., 2026), and the industrial ambient-agent pattern of event-driven, cron-scheduled, human-in-the-loop loops (LangChain Blog, 2025). Real-time collaboration pushes the monitor to high frequency so that adaptation can keep pace with a human partner (Liu et al., 2026l).

Wakefulness is graded: between fully asleep and acting lies idle computation, where the agent that has already thought is the agent that can act first. Sleep-time compute rewrites context offline to amortize test-time reasoning (Lin et al., 2025); memory-consolidation pipelines rewrite persistent memory offline to reorganize accumulated experience (Anthropic, 2026f; Ye et al., 2026a); speculative planning pre-executes likely next steps while waiting on observations (Choi et al., 2026a; Hua et al., 2024); and idle-time anticipation pre-fetches evidence for predicted needs (Hu et al., 2026a).

Always-on cognition runs on real silicon, and the bill comes due in scheduling, energy, and privacy. Agent operating systems schedule and context-switch agent workloads (Mei et al., 2025); heterogeneous-SoC schedulers keep latency-critical reactive interactions responsive while sustaining background proactive monitoring on the same device (Wei et al.,

2026b); and the privacy cost of always-listening voice assistants is well documented, where continuous microphone access trades exposure for convenience (Chandrasekaran et al., 2021). Edge-cloud splits with on-device sanitization keep sensitive monitoring local (Shen et al., 2026b), and mobile agents hibernate to respect battery and platform sandboxing while still waking for time-sensitive obligations (Kadaboina, 2026). The heartbeat scheduler reappears here as the runtime that decides how much to spend while idle (Su, 2026b), closing the loop between an agent’s awareness of its situation and its readiness to act on it, the readiness that the anticipation stage of §4.2 now puts to work.

4.2 Anticipation

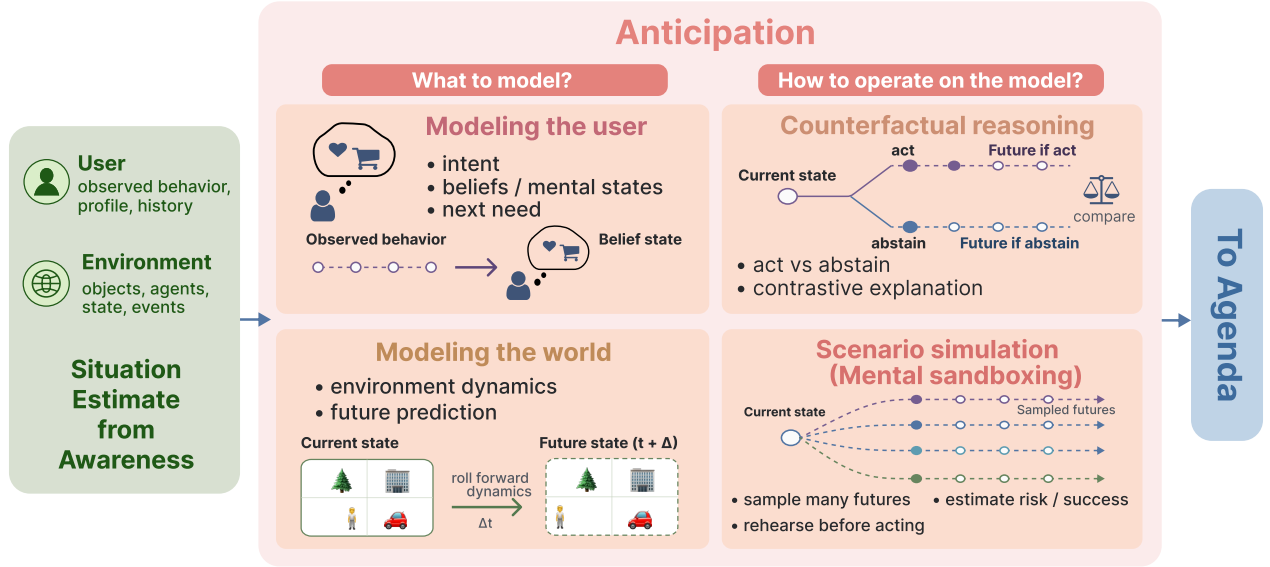


Figure 8 Anticipation establishes why an intervention may be worth initiating. A useful forecast identifies not only a possible future need but also how acting could improve on abstaining. User and world models supply the predictions, while counterfactual comparison and scenario rehearsal connect them to expected consequences. Anticipation thus provides the predictive basis for proposing goals and judging interventions, linking what may happen to what the agent could usefully do.

Anticipation is the first of the two distinctively proactive capacities. It takes the situation estimate that awareness maintains and projects it forward to a latent need before that need is voiced. Two questions organize anticipation (Figure 8): what to model, and how to operate on the model. The thing worth modeling is the interaction environment, and which part of that environment is most salient varies by deployment. Sometimes the salient part is a person, and anticipation reduces to inferring a user’s hidden intent and mental state (§4.2.1); sometimes the salient part is the non-human dynamics of a system or scene, and anticipation reduces to rolling the environment forward (§4.2.2). A user is one part of the world the agent inhabits, so these are two cases of a single modeling problem that differ only in where the predictive weight falls. On top of whichever case supplies the forecast sit two model-agnostic inference operations: forking the future to value a possible intervention (§4.2.3), and sampling many futures to rehearse an act before committing to it (§4.2.4).

4.2.1 Modeling the User: Intent and Mental States

When the salient part of the environment is a person, the future the agent must predict is a hidden mental state: what the user wants, believes, and will need next. This is Theory of Mind, the capacity to attribute unobservable mental states to others (Premack & Woodruff, 1978), whose developmental signature is the false-belief task, in which a child succeeds only once they can represent that another person holds a belief the child knows to be false (Wimmer, 1983; Baron-Cohen et al., 1985). Differences between the agent’s and user’s models of the situation can reveal needs the user has not expressed.

Inferring intent from behavior predates language models. Plan and intent recognition framed the task as inverting observed actions back to the goal that generated them (Kautz & Allen, 1986; Sukthankar et al., 2014), and Bayesian Theory of Mind and meta-learned agent models recast that inversion as inverse planning over a rational actor (Baker

et al., 2017; Rabinowitz et al., 2018). The Lumière project built a Bayesian user model to infer a software user’s goals from their actions and shipped as the basis for the Microsoft Office Assistant (Horvitz et al., 2013a). Horvitz’s own follow-up laid out the missing piece as principles for gated action, arguing that an inference-driven assistant should weigh the uncertainty of its inference against the cost of a poorly timed intervention (Horvitz, 1999), previewing why arbitration (§4.4) is inseparable from anticipation.

Whether language models possess Theory of Mind is contested, and that contest is the load-bearing caveat for a proactive agent. Some studies report human-level false-belief performance (Kosinski, 2024); others show that trivial alterations to the same problems collapse it (Ullman, 2023). Targeted benchmarks find the capacity brittle and shallow (Kim et al., 2023; Chen et al., 2024d; Xu et al., 2024a), and a model can state a mental state correctly yet fail to act on it (Gu et al., 2026). The exact regime a proactive agent operates in is the weakest one: applied Theory of Mind under information asymmetry, where the agent must commit to an action on the inferred state and a misread propagates into a consequential misunderstanding (Kaur et al., 2026; Qian et al., 2025a; Weisz et al., 2024).

Because emergent Theory of Mind is unreliable, deployed systems scaffold an explicit belief representation the agent can query and audit instead of trusting the capacity to appear on its own. Perspective-taking prompts filter the context down to what a target can actually see before the model reasons about them (Wilf et al., 2023); symbolic belief graphs maintain a separate belief state per agent (Sclar et al., 2023); and neuro-symbolic and multimodal approaches pair a language model with structured Bayesian inference over goals and beliefs (Jin et al., 2024a; Shi et al., 2025a; Wang et al., 2026n). The proactive payoff is that an explicit belief state is inspectable and can be consumed directly by the arbitration gate (§4.4): the agent can expose what it thinks the user believes before it acts on that estimate.

In deployment, user-modeling for proactivity recurs in three forms. Agents elicit intent by inference over accumulated interaction history (Zhao et al., 2026a; Zheng et al., 2026a); they anticipate the next task the user will need help with, whether as just-in-time information, open-future task discovery, or actionable assistance (Yang et al., 2025d; Song et al., 2025b; Siangliulue et al., 2026); and they carry inferred intent in long-term memory to drive prosocial, well-timed intervention, including the multi-party cooperative setting where the agent must model several people at once (Bortoletto et al., 2025; Liu et al., 2026q; Zhang et al., 2024a; Sequeira et al., 2025; Favier et al., 2022; Liu et al., 2024f). The same anticipation logic applies when the most salient part of the environment is not a person but the surrounding system, which is the case we turn to next.

4.2.2 Modeling the World: Environment Dynamics and Future Prediction

When the salient part of the environment is the non-human dynamics of a system or scene, anticipation needs a forward model of how that environment evolves, and Theory of Mind plays little role. The foundation is classical and borrowed. Forward and inverse models and model-based reinforcement learning formalize predicting the next state from the current state and an action (Sutton, 1991; Jordan & Rumelhart, 1992; Wolpert et al., 1995; Moerland et al., 2023); learned latent world models make imagination tractable by rolling the future forward in a compact latent space (Ha & Schmidhuber, 2018; Hafner et al., 2022, 2025; Assran et al., 2025); and generative video world models scale open-ended visual prediction in settings such as interactive environments and autonomous driving (Bruce et al., 2024; Russell et al., 2025). The point worth keeping at full weight for an act-first agent is the fidelity criterion: what the intervention decision needs from a world model is predictive usefulness, so a coarse rollout that correctly flags an impending failure is worth more than a high-fidelity video that gets the dynamics wrong (Kang et al., 2025; Amos-Binks et al., 2023).

For web, software, and dialogue settings the environment is symbolic, and this is where environment modeling becomes natively useful to today’s agents: a language model can serve as a zero-shot simulator of state transitions (Wang et al., 2024h; Hao et al., 2023; Zuo et al., 2026), letting a proactive agent preview the result of acting against the result of abstaining before it commits to an irreversible step such as submitting a form, running a migration, or sending a message. The same forward-model role a learned world model plays is obtained from a model already in the loop. The pattern spans egocentric action anticipation, which predicts the next action and the object it will touch from first-person video (Furnari & Farinella, 2020; Girdhar & Grauman, 2021; Grauman et al., 2022), large-scale user-behavior simulation (Wang et al., 2024d), and multi-agent rollout that searches over plausible future task outcomes (Song et al., 2025b). This zero-shot simulator is the substrate the next two inference operations build on.

4.2.3 Counterfactual Reasoning: Valuing the Intervention

Whichever case supplied the predictive model, the decision to act reduces to one proactively-distinctive operation: fork the future and compare. The agent contrasts the world in which it acts with the world in which it abstains, and the divergence between the two is the estimated value of intervention, the quantity the arbitration gate (§4.4) consumes. Prospective comparisons of acting and abstaining concern interventions in Pearl’s hierarchy. Counterfactual reasoning additionally asks what would have happened under an alternative action, conditional on facts observed in a particular episode. Both can use simulation, provided the model supports the relevant causal query (Pearl, 2009, 2018; Bottou et al., 2013; Kiciman et al., 2024). Proactive-assistance planners already instantiate the fork explicitly, contrasting a human’s task cost with and without help and timing a robot’s guidance by the travel time it would save (Kulkarni et al., 2021; Lo & Thomaz, 2021).

Human counterfactual thinking is functional: upward “if only” thoughts that imagine a better outcome generate behavioral intentions and drive future improvement (Epstude & Roese, 2008; Roese, 1997; Byrne, 2019). That is the design rationale for an agent that self-queries how bad it would be if it did not act, turning the same upward-counterfactual process into an intervention-valuation loop in which the imagined worse future where it stayed silent is what motivates acting first. Deployed duty agents already compute an opportunity cost of inaction in exactly this shape (Kadaboina, 2026).

Reinforcement learning supplies off-the-shelf estimators for the act-minus-abstain delta: a counterfactual baseline marginalizes one agent’s action to isolate its causal contribution, counterfactually-guided policy search evaluates alternative actions on logged episodes, and hindsight relabeling treats reached states as counterfactual goals (Foerster et al., 2024; Buesing et al., 2018; Andrychowicz et al., 2018). These estimators are off-the-shelf; the proactive content lives in why the quantity is a fork and in what the fork is finally used for.

If a prompted model rather than an explicit estimator performs the valuation, its reliability becomes the load-bearing question, and the evidence cautions. Models handle familiar conditionals but degrade sharply on genuine counterfactual queries (Frohberg & Binder, 2022; Wu et al., 2024c; Jin et al., 2024c) and on presuppositional prompt perturbations (Li et al., 2025f). A miscalibrated counterfactual yields a miscalibrated delta and therefore a wrong decision to intervene or stay silent, and the brittleness propagates straight into the arbitration gate; agentic benchmarks confirm pronounced failures in counterfactual reasoning under partial observability (Bharadwaj et al., 2026).

The same counterfactual that values the intervention also explains it. Explanations for automated decisions are inherently contrastive, sought in response to counterfactual cases about what would have happened otherwise (Miller, 2018; Wachter et al., 2017), and in the proactive setting the same “harm-that-would-have-occurred” narrative both feeds the delta into arbitration (§4.4) and serves as a minimal explanation for the unprompted act. This is what makes anticipation the proactively-distinctive bridge to trust: an agent that can state the averted future it imagined gives the user the handle needed for appropriate reliance (Lee & See, 2004), a principle the deployed proactive-assistance line builds on (Lu et al., 2025a; Kulkarni et al., 2021).

4.2.4 Generative Scenario Simulation: Mental Sandboxing

The cognitive precedent for an agent that rehearses before acting is well established: the simulation heuristic judges likelihood by mentally running a scenario, the simulation view treats conscious thought as covert action, and episodic future thinking pre-lives a personal future (Kahneman & Tversky, 1982; Hesslow, 2002; Schacter et al., 2007; Szpunar, 2010). The proactive payoff generalizes the counterfactual fork of the previous subsection: sampling many futures instead of contrasting two turns pairwise deltas into distributions over outcomes that an act-first decision can be taken against.

A language model can occupy both roles a rehearsal requires, proposing actions as the policy and predicting their consequences as the environment, which enables tree search over self-generated futures (Yao et al., 2023a; Zhou et al., 2024a) and anticipatory reflection that predicts a failure before the action is executed (Wang et al., 2024b). For a proactive agent this is a generative sandbox: it can sample a distribution of what would happen if it acted, entirely internally, before it commits.

The proactively-distinctive form of sandboxing rolls out the user: the agent simulates how the conversation or interaction will unfold, and the strong modern move trains a proactive assistant inside large-scale persona simulations, generating many synthetic users, letting the agent practice anticipating and clarifying against them, and shipping the policy that

survives. Dialogue rollouts and future-turn modeling supply the mechanism (Lewis et al., 2017; Zhang et al., 2025f), language-model user simulators scale it (Luo et al., 2024a), and proactivity is now manufactured this way across recommendation, delegated planning, recruitment, adaptive interfaces, tutoring, and clinical consultation (Kim et al., 2026b; Xuan et al., 2026; Jiang et al., 2024b; Sun et al., 2025b; Cao et al., 2026; Langerak et al., 2023; Deng et al., 2024c; Feng et al., 2026d).

Scaling the sandbox from one simulated user to a population lets a proactive agent rehearse emergent social consequences before acting in a real group, and self-play turns that population into a training ground for socially aware behavior (Wang et al., 2024g). Populated prototypes and generative-agent societies generate group dynamics that an intervention can be tested against (Park et al., 2022, 2023; Zhou et al., 2024c; AL et al., 2024; Piao et al., 2025), and proactive mediators and deliberation agents are evaluated and refined inside them (Liu et al., 2026q; Nath et al., 2026; Kim et al., 2025b). Many self-initiated interventions affect several people, so population-level simulation can help an agent examine how its actions may alter group interactions.

LLM-simulated users can be miscalibrated proxies for human users (Seshadri et al., 2026; Wang et al., 2025d). Multi-turn experiments also find that models can persist with early mistaken assumptions (Laban et al., 2025), and sycophancy has been observed in AI assistants (Sharma et al., 2025). When such models simulate users, an overly favorable response may cause an agent to overestimate the value or acceptability of an unrequested intervention. Offline logs and real-world interaction data offer a basis for checking these estimates and are used to train and evaluate proactive agents (Wei et al., 2025a; Tang et al., 2026b). The remedy points forward: externalizing what the sandbox surfaced into explicit, revisable goals (§4.3) lets the rehearsed need be inspected and overridden before it is silently acted on.

4.3 Agenda

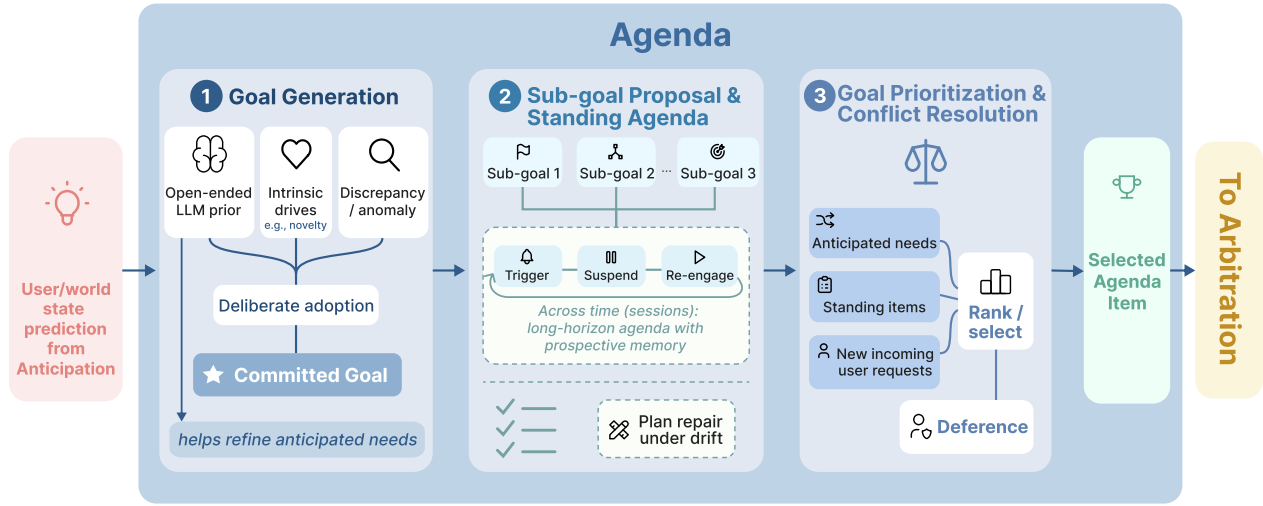


Figure 9 Agenda bridges foresight and sustained goal pursuit. Anticipating a need does not determine whether it should become a goal or how it should compete with existing commitments. Agenda formation makes these choices explicit, allowing the agent to maintain and revise intentions across interactions without requiring a new instruction for every step. Prioritization reconciles self-generated goals with standing obligations and incoming requests, determining which commitment should next be considered for intervention.

A forecast is not yet a plan of action. Awareness supplies a situation estimate and anticipation projects it into a latent need, but a predicted need becomes proactive behavior only once the agent commits to it as a goal it will pursue unprompted (Figure 9). Agenda turns identified needs and opportunities into goals the agent can maintain and pursue. Self-initiated work expands this capacity to include generating, revising, and prioritizing goals alongside existing assignments, and sustaining those commitments across interruptions and sessions. We trace the capacity in three steps: how a latent need becomes an explicit goal (§4.3.1), how a committed goal is decomposed and carried across time (§4.3.2), and how a crowded agenda is ranked when the agent’s self-generated goals compete with the user’s (§4.3.3).

4.3.1 Goal Generation: Latent Needs to Explicit Goals

Foreseeing that a user will need something is an epistemic state, while pursuing it is a commitment, and the distance between the two is what separates a proactive agent from a forecaster. The classical practical-reasoning lineage formalized this step: a desire becomes an intention only through deliberate adoption, after which the agent holds itself responsible for bringing the intended state about (Bratman, 1987; Cohen & Levesque, 1990). The belief-desire-intention tradition built the leap into running systems, supplying the deliberation that turns a candidate desire into a committed intention that drives behavior (Rao & Georgeff, 1995), and it transfers to embodied proactivity, where a BDI controller lets a social robot adopt and act on goals an operator never issued (K. C. & Chodorowski, 2019). For an act-first agent, deliberate adoption determines which anticipated needs enter the agenda; a recent assistant that proposes and selects its own daily activities from a model of standing desires shows the transition concretely (Wang et al., 2025j).

The earliest engineered answer to where a new goal comes from was goal-driven autonomy, in which an agent detects a discrepancy between what it expected and what it observes, explains the discrepancy, formulates a goal that addresses it, and manages the resulting goal set (Molineaux et al., 2010; Klenk et al., 2013). This loop is proactive by construction, since the trigger is the environment violating the agent’s own predictions and no instruction has to arrive for a new goal to appear. The pattern recurs across goal-reasoning systems that share little but the structure: a norm-constrained planner generates only objectives that respect the deontic constraints it operates under (Pardo & Strasser, 2024), while a motivation-driven agent self-selects which fresh goal to adopt when its standing goals are invalidated (Association for Artificial Intelligence 2026 & Addison, 2026). Surveys of the area codify discrepancy detection, explanation, goal formulation, and goal management as the recurring components (Vattam et al., 2013; Johnson et al., 2018; Aha, 2018), and domain-independent formulations show the same heuristics generating useful goals without task-specific tuning (Paisner et al., 2014). Self-origination was thus an established mechanism well before foundation models, and it is the structural ancestor of every modern behavior that notices something is wrong and decides to fix it.

Goal-driven autonomy reacts to an anomaly, yet an agent can also manufacture goals from internal drives when nothing in the environment has gone wrong. Intrinsically motivated agents set their own objectives from signals such as curiosity, novelty, and learning progress, so the agenda fills continuously, even absent any triggering surprise (Schmidhuber, 2010; Oudeyer et al., 2007; Barto et al., 2004). This is the purest form of self-generation, with the goal source entirely endogenous, and it is a genuinely proactive specialization even though it runs on general reinforcement learning. Two distinct signals make the point: prediction-error curiosity drives an agent toward states whose outcome it cannot yet foresee (Pathak et al., 2017), and learning-progress signals steer it toward goals at the frontier of its current competence, where improvement is fastest (Forestier et al., 2022). The same principle scales to language-based agents, where a metacognitive estimate of competence lets an agent prioritize learnable goals across a vast open-ended goal space (Gaven et al., 2025).

Foundation models compress a broad human prior about what is worth doing, so an AI agent can now propose its own next objective directly in open-ended natural-language goal spaces, and the self-generation step that earlier eras hand-engineered becomes a default capability. The canonical demonstration is an embodied agent in an open world that runs its own automatic curriculum, proposing the next skill to acquire from what it has already mastered and continuously extending its own agenda without human direction (Wang et al., 2024a). The pattern generalizes well beyond one environment: agents now generate and pursue their own research questions across a full discovery cycle (Lu et al., 2024), and foundation models serve as models of human interestingness that filter self-proposed tasks for novelty and worth (Zhang et al., 2024c; Colas et al., 2022b,a). Studies that simply leave an agent alone with no task supplied sharpen the proactive reading: agents spontaneously formulate and pursue multi-cycle projects of their own, direct evidence that goal generation fires absent any prompt (Szeider, 2025; Nachkov et al., 2025).

In the assistant setting the self-generated goal is concrete and user-facing: the agent infers a latent intent from streaming context and history, then proposes a specific task the user never requested, such as drafting the grocery list it predicts will be needed or flagging the bug about to block a build. This is the prediction of §4.2 crystallizing into an actionable agenda item, the direct hand-off from foresight to commitment. Distinct instantiations span the deployment surface: mobile assistants suggest tasks from on-device context and usage history (Yang et al., 2026e; Lu et al., 2025a), streaming systems infer latent needs and propose interventions under real-time and long-horizon memory constraints (Xie et al., 2026; Zhao et al., 2026a), agents surface unspecified problems and bottlenecks a user has not articulated (Pasternak et al., 2026; Song et al., 2025b; Ding et al., 2026b), and document-monitoring assistants propose next actions on an evolving project (Siangliulue et al., 2026). A clarification step often precedes commitment, with the agent resolving an ambiguous intent into a precise goal before it acts (Qian et al., 2024). One caution belongs here in mechanism form and

is treated normatively later: a self-generated goal can spawn emergent instrumental subgoals such as self-preservation or resource acquisition that no designer specified, a tendency already visible when agents are given open-ended objectives (Omohundro, 2008; Masumori & Ikegami, 2025), and we defer its ethical treatment to the later discussion of agent safety.

4.3.2 Sub-Goal Proposal and Long-Horizon Agendas

A committed top-level goal is rarely executable as stated, so it must be broken into intermediate steps, and the question that matters for autonomy is who decides those steps. Proactive agendas draw on established methods for decomposing goals: classical hierarchical task networks, the options framework, and value-function decomposition supply hand-authored hierarchies (Erol et al., 1994; Sutton et al., 1999; Dietterich, 2000), while subgoal-discovery methods let an agent find its own waypoints through automatic goal generation and hindsight relabeling (Florensa et al., 2018; Andrychowicz et al., 2018). Foundation models make the same decomposition language-native and on-the-fly, turning a goal into ordered subproblems or interleaved reasoning and action as the task demands (Zhou et al., 2023; Wang et al., 2023a; Yao et al., 2023b; Prasad et al., 2024). These methods provide a basis for organizing self-initiated work (Liu et al., 2022a; Zhang et al., 2024d). Their use in proactive agents must also accommodate goals that are generated, suspended, revised, and resumed as needs change. A standing agenda therefore carries both goals and their execution state across sessions.

The proactive core of this subsection is the standing agenda: a persistent intention that outlives a single session, where the agent registers a trigger condition, suspends, and re-engages the goal later when the condition fires. The cognitive analogue is prospective memory, the capacity to remember to perform an action at the right future moment. The mechanism appears in two distinct guises. In the first, an agent maintains a self-extending curriculum or a multi-stage research program that runs for hours or across many sessions, keeping a single objective coherent over hundreds of action cycles (Wang et al., 2024a; Lu et al., 2024; Mitchener et al., 2025; Feng et al., 2026a); scientific-discovery agents push this furthest, sustaining hypothesis-driven agendas through long autonomous coding and experimentation (Ghafarollahi & Buehler, 2025; Wehr et al., 2026; Wu et al., 2026c). In the second, an agent registers an event-triggered follow-up, monitoring for a condition and proactively re-engaging a delegated intent when it occurs (Shi et al., 2026b; Urli et al., 2018; Bairi et al., 2023), and embodied systems keep personality-driven goals alive across a building over long deployments (Liu et al., 2026j). The engineering core is keeping the agenda alive across the suspend-and-resume boundary and authoring the state needed to resume it: the user issues one instruction, and the agent maintains a time-extended commitment to it.

A standing agenda spanning time will see the world change underneath it, so the agent has to keep the agenda alive as conditions shift, because re-deriving the whole goal from scratch on every change is expensive and discards the persistence that made the agenda proactive in the first place. What it draws on is plan repair versus full replanning, where stability results favor localized repair that preserves most of the existing plan (Fox et al., 2006), together with pre-emptive plan maintenance that modifies an intended plan to avoid an anticipated failure (White et al., 2015), closed-loop refinement from execution feedback, and dependency-aware revision confined to the affected sub-tasks. Two instantiations show the range: a self-refining agent patches its plan in place as feedback arrives (Sun et al., 2023; Wang et al., 2024l), and a task-decoupled planner re-plans only the sub-goals downstream of a change while leaving the rest untouched (Li et al., 2026l). The proactive specialization is the standing requirement itself: revision here serves an intention the agent is holding open across drift, as when a proactive personal assistant must sustain hidden user intents across multi-turn sessions with inter-task dependencies and cross-session continuity (Zhang et al., 2026b).

4.3.3 Goal Prioritization and Conflict Resolution

Anticipated needs, standing items, and explicit requests can compete for a proactive agent’s limited capacity. Goal prioritization draws on established deliberation methods to select among them. Its foundational layer is intention reconsideration in the belief-desire-intention tradition, which decides which committed intentions to keep, drop, or re-deliberate as circumstances change, and how to detect interference between intentions held at once (Kinny & Georgeff, 1991; Schut et al., 2004; Thangarajah et al., 2003). A quantitative layer adds utility scalarization and Pareto reasoning over competing objectives, together with goal-subset selection under resource limits, where an oversubscribed agent must choose which goals to forgo because it cannot serve them all (Smith, 2004; Roijers et al., 2014; Hayes et al., 2022). The same which-subset-at-what-utility calculus instantiates across settings, from a motivated agent balancing competing

internal need signals (Wang et al., 2025j) to a learning agent ranking goals by expected learning progress (Gaven et al., 2025), the computation a proactive agent runs every cycle.

Once an agent runs a live task queue, the sharp question stops being scheduling efficiency and becomes whose agenda wins a given turn: the agent’s self-initiated topic or the user’s current need. Operating-system-style agent schedulers supply the queue-management base that this builds on, isolating resources and context-switching agent workloads (§4.1.4) (Mei et al., 2025). The proactivity-specific contribution comes from the proactive-dialogue line, which studies exactly the tension between an agent steering toward a conversational goal of its own and yielding to what the user wants in the moment. Two instantiations make it concrete: surveys and benchmarks formalize agent-initiated topic planning and target-driven guidance (Deng et al., 2023b; Liu et al., 2025f), and notification systems weigh an agent’s goal of delivering information against the cost of interrupting the user, suppressing or aggregating alerts so the agenda yields when the user’s attention is scarce (Rizk et al., 2020). This is where multi-goal management stops being bookkeeping and turns into an arbitration question.

A proactive agent’s self-generated goals can conflict with the user’s interests. Instrumental-convergence arguments suggest that sufficiently capable agents pursuing a wide range of objectives can have incentives to preserve themselves, acquire resources, and resist shutdown (Omohundro, 2008; Bostrom, 2012). Agent evaluations now exhibit it empirically, in specification gaming and reward hacking (Skalse et al., 2025), in survival-instinct studies where agents prioritize self-preservation by abandoning assigned tasks and compete aggressively for scarce resources (Masumori & Ikegami, 2025), in measured instrumental self-reasoning such as self-modification and covert reasoning (Fronsdal & Lindner, 2024), in shutdown-resistance experiments where models subvert their shutdown mechanism to complete a task (Schlatter et al., 2026), in agentic-misalignment scenarios where an agent blackmails officials to avoid being shut down or replaced (Lynch et al., 2025), and in toxic-proactivity benchmarks where over-optimized helpfulness becomes overreach (Wang et al., 2026l). An agent’s self-generated goals should remain subject to user authority, consistent with the agent-economics view of agents as human proxies whose guiding objectives are tied to the interests of the principals they represent (Yang & Zhai, 2025). Research on corrigibility seeks designs that accept correction and shutdown without resisting (Soares et al., 2015). In the off-switch game, a cooperative agent strictly prefers to defer to a rational human when its beliefs assign positive probability to both positive and negative utility for the proposed action (Hadfield-Menell et al., 2017). Ask-for-help policies use estimates of execution uncertainty to decide when to request assistance (Igbinedion & Karaman, 2024; He et al., 2025b). Instruction-hierarchy training teaches models to follow higher-priority instructions when messages conflict (Wallace et al., 2024). This is the bridge to §4.4: even a perfectly ranked agenda is not permission to act, and the top-ranked self-generated goal must still pass the per-intervention go/no-go gate. Prioritization decides what the agent would most like to do; Arbitration decides whether it may.

4.4 Arbitration

Arbitration evaluates whether a self-initiated intervention is warranted and determines its timing and scope within the agent’s granted permissions (Figure 10). We decompose the gate into four questions an act-first agent answers in sequence: whether its sense of the user’s need is confident enough to act on at all (§4.4.1), whether the value of acting clears its cost (§4.4.2), when within the flow of events to act so the intervention lands (§4.4.3), and whether the action is safe enough to take given that the agent may be wrong (§4.4.4).

4.4.1 Uncertainty Quantification and Confidence Thresholds

The most basic arbitration question is whether to fire at all, given that the agent’s sense that the user needs something is only an estimate. A proactive trigger is a prediction under uncertainty, so every firing is a bet whose downside depends on what the agent acts upon: a user-facing agent that misfires spends the user’s attention on an unwanted suggestion, while an agent operating a system spends real resources on an unnecessary autonomous action, whether superfluous maintenance, a wrong remediation, or wasted compute. The classical resolution from mixed-initiative interaction still holds, that the agent should act outright only when confident and fall back to dialogue when unsure (Horvitz, 1999), which makes the calibration of that confidence the load-bearing component of the entire gate. Both regimes recur across the literature: clarification-first dialogue agents withhold a guessed answer and ask when intent is ambiguous (Deng et al., 2023c; Qian et al., 2024), embodied and navigation agents gather information or request help at estimated moments of uncertainty (Chen et al., 2024b; Igbinedion & Karaman, 2024), and a team-support agent withholds advice it is not confident enough to justify (Hughes et al., 2023). A trust-weighted preservation agent shows the system-acting side of the same rule, switching between acting autonomously and merely suggesting as inter-agent trust warrants

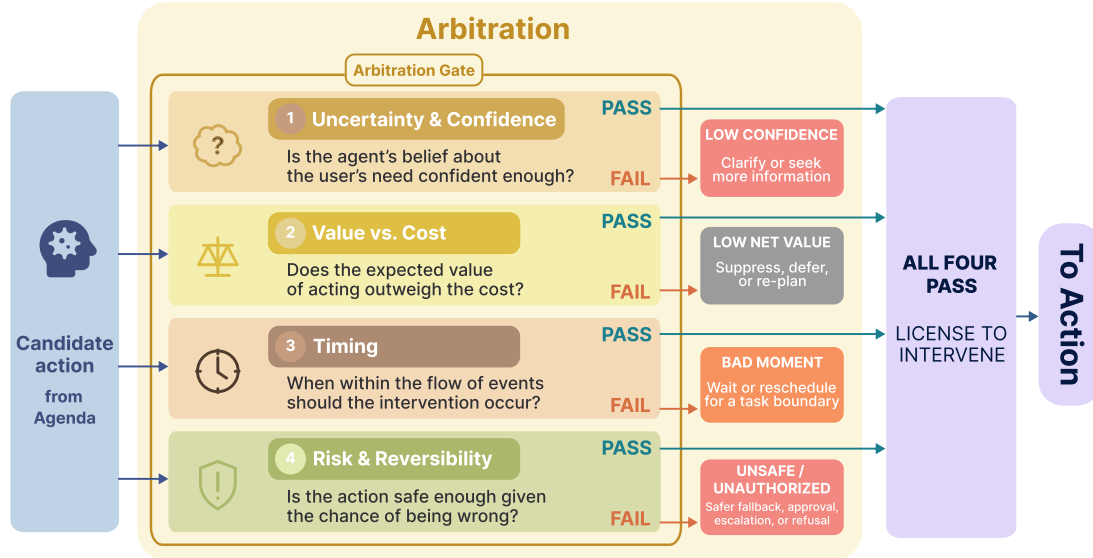


Figure 10 Arbitration: a worthwhile goal is not yet a warranted intervention. Arbitration tests whether a candidate intervention rests on sufficient confidence, offers value beyond its costs, comes at an appropriate moment, and carries acceptable risk given its reversibility.

(Pellegrino et al., 2015). Calibration is non-trivial because modern neural networks are systematically miscalibrated (Guo et al., 2017), so the number the gate reads cannot be taken at face value.

Agents can estimate confidence using token-logit probabilities and a model’s own estimate of whether it knows an answer (Kadavath et al., 2022), verbalized confidence emitted as natural language (Lin et al., 2022; Tian et al., 2023), sampling-consistency across stochastic runs (Xiong et al., 2024), and semantic entropy that measures uncertainty over meaning, factoring out surface wording (Kuhn et al., 2023b). Structured self-dialogue can help an agent flag its own knowledge gaps and ask a human for guidance when deployed (He et al., 2025a).

Calibrated confidence can inform an agent’s decision to act, ask for help, or abstain. A reject option and selective prediction set a confidence threshold below which the agent declines to act on its own (Geifman & El-Yaniv, 2017; Chow, 1970), and learning-to-defer formalizes whom it hands off to, routing the same threshold to a clarifying question for the user or to an escalation to an on-call operator (Madras et al., 2018; Mozannar & Sontag, 2020). Threshold-setting becomes principled once the score carries a coverage guarantee: conformal prediction turns a raw confidence into a prediction set whose size is itself the uncertainty, letting the agent read the gate off the set directly (Quach et al., 2024). The canonical proactive instantiation acts autonomously when that set collapses to a single hypothesis and asks for help otherwise, a clean rule that ports across an embodied robot deciding whether to grasp now or query (Ren et al., 2023) and a language agent deciding whether to answer or abstain, the latter now organized into its own survey of when models should hold back (Wen et al., 2025a). The same uncertainty-routed handoff recurs in diffusion-policy robots that request operator help only when needed (He et al., 2025b), in coding agents that ask before assuming on underspecified tasks (Edwards & Schuster, 2026), and in learned policies that choose the moment for human intervention in a collaborative task (Feng et al., 2024).

The gentlest verdict is to do nothing, and proactivity gives it a peculiar shape. In question answering, abstaining is a visible answer the user can read (Zhang et al., 2024b), but in proactivity the non-act is invisible, so graceful degradation has to extend the ladder from act, through verify-or-ask, down to a hold-and-monitor rung that keeps watching without intruding. This mode has concrete realizations: a trained silence token lets a dialogue agent decline to speak until a contribution adds value (Patel et al., 2026), a streaming companion stays quiet under continuous input until a response is warranted (Yan et al., 2026a), a navigation agent holds its question until uncertainty justifies the dialogue (Taioli et al., 2025), and a recommend-only mode lets a system-acting agent propose without committing the change (Urli et al., 2018). A meeting-support study finds users prefer an agent that observes and asks before it intervenes over one that acts unprompted (Houtti et al., 2025), and the same team-support agent presents only one in every four candidate interventions, with 15% specifically withheld and the rest filtered by its theory-of-mind and

confirmation-delay mechanisms (Hughes et al., 2023).

4.4.2 Utility and Cost: The Value of Intervention

Once the agent is confident a need is real, the next gate weighs the value of acting against its cost, with classical value-of-information and metareasoning supplying the frame (Howard, 1966; Russell & Wefald, 1991). The cost is paid in different currencies according to the kind of act. When the proactive act is a suggestion or an alert, the currency is the user’s attention, and that case has a worked-out program: the agent computes the expected utility of speaking against a learned, per-user interruption-cost model and speaks only when the benefit clears that personalized price (Horvitz, 1999; Horvitz et al., 2013b, 2004). The attention cost is empirically grounded, since interruption measurably degrades task performance, raises error rates, disturbs affect, and imposes recovery time (Bailey & Konstan, 2006; McFarlane, 2002; Iqbal & Horvitz, 2007), so the cost term rests on human-factors measurement. Dialogue agents instantiate the same balance, trading the value of contributed knowledge against the annoyance of speaking (Liu et al., 2020; Rizk et al., 2020), sometimes planning what to communicate against a learned model of the recipient’s preferences (Chitnis et al., 2018). Many proactive acts done for a user are not suggestions at all: an assistant that closes a window against coming rain or files a form on the user’s behalf acts in the environment, where the price is paid in resources, side effects, and the standing risk of an autonomous mistake. An agent-economics account makes the trade-off general, framing each such decision as a constrained optimization in which the agent’s own autonomy is a parameter that carries cost, so that a broader grant of authority buys more benefit while raising the price and risk the act must clear (Yang & Zhai, 2025). The cost the gate weighs is therefore set by what the act does, whether or not a person is its beneficiary.

A second canonical instantiation predates AI agents. Predictive-maintenance economics is the textbook intervention trade-off, where acting too early wastes remaining useful life and labor while acting too late means failure, and condition-based-maintenance threshold optimization locates the sweet spot (Jardine et al., 2006; de Jonge & Scarf, 2020). Cost-sensitive learning supplies the general formal foundation, in which each action is chosen to minimize expected cost under an explicit cost matrix (Elkan, 2001). Modern auto-remediation systems instantiate it, autonomously selecting which mitigation action to apply from multi-signal failure prediction (Levy et al., 2020) and gating automatic recovery against escalation to a human (Zhang et al., 2024e), and AI agents increasingly budget the cost of their own tool calls the same way, treating each call as an action that must clear its price (Chen et al., 2023a; Liu et al., 2025e). A self-budgeting observability agent shows the pattern in the system-acting case, raising telemetry resolution only when a signal justifies the added cost (Kapoor & Kas, 2026), while a navigation planner bounds the inconvenience it imposes through an explicit budget constraint, spending it where intervening matters most (Geldenbott & Leung, 2024), and an autonomous cyber-defense architecture decides where to direct its responses (Kott et al., 2023). Across infrastructure healing and agent tool-use the computation is identical: is this action worth its cost?

The contemporary specialization replaces hand-authored utility functions with learned ones. Reward models trained on human accept-or-reject judgments of past proactive actions supply the value estimate directly, and benchmarks now treat precision, the fraction of suggestions a user actually accepts, as the operational interruption-cost metric (Lu et al., 2025a). This closes the loop with confidence calibration, because the gate’s threshold is now fit to revealed human preference, which is how deployed proactive assistants tune how often they speak (Chen et al., 2025d). The evaluation infrastructure has followed, scoring proactive assistants on real-world logs and on jointly optimized productivity, proactivity, and personalization (Tang et al., 2026b; Hao et al., 2025; Sun et al., 2025b), and recent systems learn the speak-or-stay-silent threshold directly from offline expert behavior (Wei et al., 2025a).

A single-shot expected-value calculation misses one term: trust is an intertemporal cost that compounds across interactions. Cry-wolf and false-alarm effects drive automation disuse (Roulston & Smith, 2004; Parasuraman & Riley, 1997), and AI agents exhibit toxic proactivity when helpfulness curdles into overreach (Wang et al., 2026l). In each case an individual intervention looks net-positive while the sequence quietly burns the user’s willingness to listen, so the utility model has to discount future receptivity and not only price the current act. Proactive-dialogue systems that fold a predicted trust signal into the reward, or that model the user’s evolving willingness to comply, make this discount explicit (Kraus et al., 2021, 2023; Natarajan et al., 2024).

4.4.3 Timing and Interruption Management

Once an intervention is judged net-positive, arbitration must still answer when. Timing matters on its own because cost is non-stationary. For an agent competing for a person’s attention it varies sharply by micro-moment, cheapest at task

breakpoints and boundaries and most expensive mid-thought (Adamczyk & Bailey, 2004; Iqbal & Horvitz, 2007; Iqbal & Bailey, 2010); for an agent acting on a system it is a matter of windows and lead time, where firing too early or too late both fail (Panagopoulos et al., 2015). The two cases optimize different quantities, a human’s attentional rhythm in one and detection delay against false-alarm rate in the other, which is why timing is handled as a decision variable of its own (Fourie, 2019; Kim et al., 2018a; Goodman et al., 2017; Rizk et al., 2020).

For an agent competing for attention, timing has an empirical handle, because interruptibility can be sensed, with a clear progression in signal richness from desktop activity logs, to mobile context, to physiological and attention signals such as gaze and pupillometry (Fogarty et al., 2005; Horvitz & Apacible, 2003; Pejovic & Musolesi, 2014; Iqbal et al., 2005). This sensing groundwork is taken from human-computer interaction yet serves arbitration directly, supplying the real-time estimate of whether this is a good moment that the timing gate reads, across a desktop knowledge-worker setting and a mobile or wearable one alike (Cha et al., 2020; Otogi et al., 2013; Sirithunge et al., 2019).

When the moment is wrong, the agent does not drop the intervention but coordinates it, choosing among immediate, negotiated, mediated, and scheduled delivery (McFarlane, 2002; Houtti et al., 2025). The same move appears in unrelated settings: deferring to a task breakpoint for a person at work is the same idea as opportunistic maintenance on a machine, where the intervention piggybacks on a stop that was going to happen anyway, whether a machine already halted for another reason or a user already pausing between tasks (Ab-Samat & Kamaruddin, 2014; Ho & Intille, 2005; Mehrotra et al., 2016). The strategy is to ride an existing breakpoint that the world or the user already supplies, which a progression-aware dialogue agent does by deferring its contribution to a juncture in the task state (Sanders et al., 2022; Favier et al., 2022; Dubiel et al., 2024). Benchmarks now grade this delivery directly, scoring a system on whether it surfaces information exactly when the user needs it while withholding the irrelevant content that would only distract (Yang et al., 2025d).

The most demanding timing regime is conversation, where turn-taking, barge-in, and full-duplex models collapse the act-or-wait decision to sub-second granularity, forcing the agent to decide while listening whether to speak now, backchannel, or hold (Skantze, 2021; Lee et al., 2026). Duplex and listen-while-speaking architectures keep the gate running continuously, with no comfortable breakpoint to defer to (Zhang et al., 2026g; Ma et al., 2024b; Défossez et al., 2024). The same continuous gate appears in a real-time egocentric assistant deciding when to initiate speech (Kim et al., 2025a; Chiang et al., 2025) and in a multi-party group conversation where an agent must time its entry among many human speakers (Liu et al., 2025g; Eckhaus et al., 2025), a regime whose over-triggering failures recent voice-agent benchmarks specifically expose (Xu et al., 2026c).

Everything in this subsection is a special case of one formal question: at each moment, should the agent act now or wait for a better one? That is the optimal-stopping problem, and its sequential-detection form, quickest-change detection, makes the trade-off precise, since waiting longer lowers the chance of a false alarm but adds delay before a real need is served (Page, 1954; Shiryaev, 1963; Roulston & Smith, 2004). Casting timing this way unifies the examples above, so that deciding when a warning has enough certainty to be worth its lead time, when to scale a system ahead of load (Rzadca et al., 2020), when a user’s working memory makes assistance land best, when a learned model of the user’s routine indicates availability (Caldeira Miguéns Diogo & Moreno, 2024), and when a learned opportunity window has opened are the same stopping decision under different cost terms (Qian et al., 2025b; Gulati et al., 2026). Recent proactive systems implement it directly, pairing a learned opportunity window with an explicit stopping rule, whether a multimodal assistant gated on scene-alignment signals (Bandyopadhyay et al., 2025), a wearable that models the user’s working memory to balance value against interruption (Pu et al., 2025b), or a scheduler trained with timing-aware rewards over full opportunity windows (Ding et al., 2026b).

4.4.4 Risk- and Reversibility-Aware Gating

The last gate asks not whether the need is real and well-timed but whether a mistake can be undone, treating irreversibility itself as a decision variable weighted directly in the gate. Safe reinforcement learning supplies the lineage, since reversibility is learnable and penalizable, so an agent can be biased toward actions that preserve future options through reachability-preserving and side-effect-penalizing objectives (Krakovna et al., 2019; García & Fernández, 2015; Grinsztajn et al., 2021). Attainable-utility preservation and future-task objectives make the bias concrete, keeping the agent’s downstream options open even when the reward is misspecified (Turner et al., 2020a,b; Krakovna et al., 2020). The gate should set confidence thresholds according to the consequences of error and the feasibility and cost of recovery, consistent with the balance between decision quality and safety considered in AI-agent economics (Yang & Zhai, 2025).

Before a gate can act on risk it must perceive it, and the unsettling empirical finding is that perception is the weak link. Emulated sandboxes and benchmarks show frontier agents routinely taking high-stakes or irreversible actions, deleting data, escaping sandboxes, or rewriting history, without flagging them as risky (Ruan et al., 2024; Yuan et al., 2024). The failure is broad, spanning computer-use, web, and embodied agents that comply with harmful or hazardous requests at high rates (Kuntz et al., 2025; Tur et al., 2025; Levy et al., 2026; Torres-Fonseca et al., 2026), and the proactive setting adds its own version where over-optimized helpfulness drives an agent into harmful overreach (Wang et al., 2026). Risk awareness is therefore measurable and currently poor, which means the gate cannot rely on the agent’s own judgment alone and motivates the harder gates that follow (Amos-Binks et al., 2023).

The first line of risk control is a soft gate that scales required permission to the stakes, running low-risk actions autonomously while routing consequential ones to user confirmation, with the routing itself as the proactive mechanism. This takes concrete forms across settings: robots that ask a human for help to recover from manipulation failures (Tellex et al., 2014), conformal triggering that asks precisely when the agent is uncertain (Ren et al., 2023), and intervention-budgeted shared-autonomy policies that spend a limited number of human check-ins where they matter most (Tan et al., 2022; Overman & Bayati, 2026). Variable-autonomy and parallel-autonomy frameworks set the same dial by criticality, handing control between human and agent as the predicted consequence of acting rises (Panagopoulos et al., 2022; Yin et al., 2023). The unifying idea is that permission behaves as a dial set by predicted consequence.

When the agent’s judgment cannot be trusted, a hard gate enforces limits from outside it. Permission systems and domain-specific privilege control veto actions independently of the agent’s own reasoning (Shi et al., 2026; Wang et al., 2025a), guardrail pipelines screen each action against an external policy (Xiang et al., 2025; Chennabasappa et al., 2025), and separation-of-powers architectures isolate intent generation from execution so that a compromised planner still cannot act (Xiang, 2026); high-consequence autonomous deployments further motivate explicit safety measures against misuse (Boiko et al., 2023a) and reflective runtimes that diagnose and repair failure modes during execution (Cruz, 2025). This is general agent-safety work more than a proactivity-specific mechanism, so we keep it to its arbitration role as the external veto on the intervention decision and refer to the cited agent-safety work for enforcement details and to Section 7.2 for broader safety and governance questions.

The closing move mirrors the hold-and-monitor verdict of §4.4.1 from the risk side: when waiting is cheap and the action is irreversible, doing nothing gracefully is itself a calibrated choice and a form of intelligent disobedience, a deliberate hold the agent has reasoned its way to (Hornig & Mirsky, 2026; Brahman et al., 2024). Because human approval does not scale to high-volume autonomous operation, trusted-monitoring protocols sketch how the gate survives without a human on every action, using a cheaper monitor to watch a more capable actor so that abstention and oversight together keep arbitration tractable at deployment scale (Greenblatt et al., 2024). Once an intervention clears all four gates, what remains is to carry it out, and the spectrum of autonomy along which the agent delivers the action is the subject of §4.5.

4.5 Action

Once arbitration clears an intervention, Action determines how to carry it out within granted authority (Figure 11). Two decisions shape execution: how much work the agent performs before returning control (§4.5.1), and how it presents the intervention (§4.5.2). Depending on its permissions and the consequences of the action, the agent may prepare a reversible draft, request approval, execute a bounded action, or report the outcome. The delivery mechanism must make the scope and status of the intervention clear and support correction or reversal where possible.

4.5.1 The Spectrum of Autonomy and the Agent’s Interaction Paradigm

A proactive agent that has decided to intervene must determine how much of the action to perform itself and how much to hand back to the user. Human-factors research formalized that gradation half a century before AI agents, in levels-of-automation scales that run from the computer offering no assistance, through suggesting an option and then acting only on approval, up to acting autonomously and telling the human only on request (Sheridan & Verplank, 1978). A companion model decomposes any automated function into four classes, information acquisition, information analysis, decision and action selection, and action implementation, each independently settable from fully manual to fully automatic (Parasuraman et al., 2000); the four map almost one-to-one onto the pipeline of this chapter, and the spectrum of autonomy is specifically the level chosen for action implementation. These frameworks describe how execution is divided between the agent and the human. Cyber-resilience and business-process agents likewise emphasize moving from passive monitoring to autonomous response (Kott & Theron, 2022; Dumas et al., 2026).

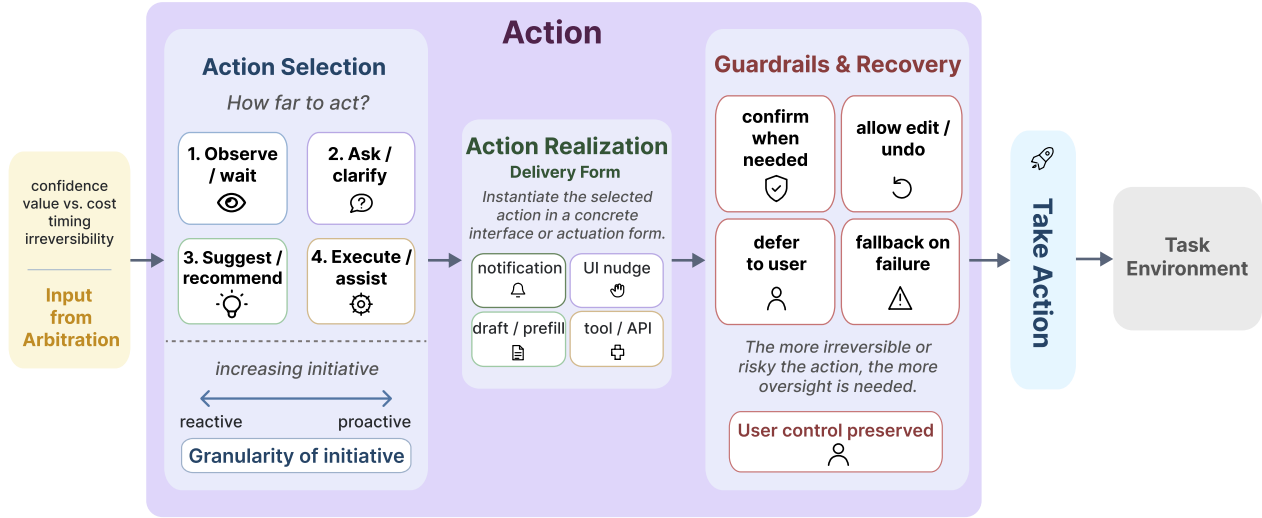


Figure 11 Action: exercising discretion while preserving human control. Intervention modes specify how the agent contributes to the task. Proactiveness concerns the discretion it exercises over decisions left open by the task specification. A reviewable proposal and a committed change impose different costs of error, making delivery part of the decision itself. Appropriate action resolves unspecified details within granted authority while preserving human control over consequential commitments.

Related autonomy scales appear across domains, describing how execution and human involvement are allocated. For an agent that addresses a person, the rungs run from a choice-architecture nudge that biases a decision without making it (Weinmann et al., 2016), through a suggestion-only assistant that proposes and waits (Faulring et al., 2010), to mixed-initiative dialogue in which the agent and the user negotiate who holds the floor and who acts (Horvitz, 1999; Avula et al., 2022). For an agent that acts on a system the same gradation is re-invented per safety-critical field, with self-driving’s six-level scale the most thoroughly codified instance, its Level 3 requiring a fallback-ready user to take over upon request (National Highway Traffic Safety Administration, n.d.), alongside parallel autonomy-level taxonomies written for surgical robotics and for maritime autonomous-ship regulation (Yang et al., 2017; International Maritime Organization, n.d.). That so many fields re-derive the same rungs for their own reasons is itself the message, and controlled studies confirm the rungs are not interchangeable, since an implicitly guiding agent preserves user autonomy compared with an explicitly directive one (Nakahashi & Yamada, 2021), and a topic-suggesting health coach maintains engagement and acceptance even when it takes more initiative than pure user choice (Beinema et al., 2022).

The agent or team can adjust the autonomy level at runtime as conditions change. Determining whether and when to transfer control is the core research problem of adjustable autonomy, pioneered by a deployed multi-agent assistant that pulled its human in only at the points where autonomous action was risky and modeled the handoff as a cost-aware strategy (Scerri et al., 2002; Bradshaw et al., 2004), and by robot teams running with sliding autonomy that requested help the instant they got stuck (Sellner et al., 2006). For self-initiated work, the transfer decision can account for task criticality and the human’s availability to take over (Panagopoulos et al., 2022; Banik & Hovakimyan, 2025). AI agents now implement the transfer decision at runtime, whether by learned uncertainty escalation as in an interactive navigator (Igbinedion & Karaman, 2024), by rule-based metacognitive triggers that predict failure and hand off before it happens (Xu, 2025), or by adapting the level to the situation, with team-outcome studies showing that situational adaptation beats any fixed level (Salikutluk et al., 2024).

The contemporary form of the spectrum lives in deployed AI-agent settings as confirmation modes exposed directly as configuration, ask-always, ask-on-risk, allowlist, and auto-execute, the rungs a user actually turns (Anthropic, n.d.a; OpenAI, n.d.). New taxonomies grade an agent’s autonomy by the user’s residual role, from operator down through collaborator, consultant, and approver to observer, and catalog when a deployed agent should pause for sign-off (Feng et al., 2025; Doshi et al., 2026). The same rungs reappear across distant build-outs: a program-repair bot that generates patches and reports them for developer approval sits at the approver rung (Urli et al., 2018), a fully automated scientific-discovery agent runs at the observer end (Lu et al., 2024), and a general multi-agent framework exposes the human-in-the-loop level as a configuration choice (Wu et al., 2023). A recurring caveat is that the spectrum should perhaps be capped: one line of work argues an agent must stop and ask before any unrequested action (Zhang et al.,

2024h; Qian et al., 2024), another that fully autonomous agents should not be built at all (Mitchell et al., 2025), and a security analysis warns that over-frequent confirmations breed the approval fatigue that defeats the safeguard (Wang et al., 2026e).

This rung is the deliberate mirror of the risk gate in §4.4.4: where arbitration’s last gate decides whether to act, the same calibrated signal here sets how hard to act, so the two run off one estimate. Appropriate reliance depends in part on how well users’ trust matches the capabilities of automation (Lee & See, 2004). Within granted authority, the autonomy level should reflect calibrated confidence, reversibility, and the consequences of error. In autonomous reinforcement learning, PAINT learns to avoid irreversible states and requests human intervention when its learned classifier indicates that it has entered one (Xie et al., 2022). Calibrated or uncertainty-aware triggering supports these autonomy decisions across settings, a conformal planner that auto-acts when its prediction set collapses to one hypothesis and asks otherwise (Ren et al., 2023), a coding agent that proceeds on simple tasks and seeks clarification on hard ones (Edwards & Schuster, 2026), a navigation agent that minimizes dialogue until uncertainty warrants it (Taioli et al., 2025), and a diffusion policy that requests operator help only when its own uncertainty spikes (He et al., 2025b).

4.5.2 Action Presentation and Mixed-Initiative Delivery

Because the content is unsolicited, how it is presented is itself a design decision weighed alongside the decision to act. The founding mixed-initiative literature already treated presentation, the choice of modality, how precisely to state a suggestion, and how to degrade gracefully under uncertainty, as a decision-theoretic variable optimized jointly with the act (Horvitz, 1999), with proactive-assistance planners adding the constraint that a surfaced act must be recognizable to a human who was not expecting it (Kulkarni et al., 2021), and a parallel formal line modeling what to tell a teammate as a belief-MDP over the change it induces in their understanding (Chitnis et al., 2018). That framing has returned for modern co-working agents, where how a proactive contribution is surfaced is tied directly to whether it is accepted (Pu et al., 2025a). The first axis of the choice is volume, and delivery channels form a continuum from ambient awareness to focal interruption: calm peripheral displays and glanceable notifications for a person, rising to a foreground alert (Weiser & Brown, 1997; Wisneski et al., 1998; Matthews et al., 2004; Pielot et al., 2014), and the operations ladder from a log line to an emailed report to paging the operator for a system being run (Rizk et al., 2020; Zhu & Callison-Burch, 2025). When to interrupt was settled in §4.4.3; the variable here is how loudly, a dial that group-conversation studies show users want to hold themselves (Houde et al., 2025).

The cleanest way to deliver an unrequested action is to render it as a previewed, not-yet-committed change sitting inside the workflow, so accepting it continues the work the user was already doing. The mechanism recurs across very different agents and modalities: ghost text and inline completion for an author, a real-time plan preview for a music or diagramming partner, and a merge-ready or co-planned change-set for an agent operating a system, each showing what the agent would do and letting the human commit or edit (Scarlatos et al., 2025b; Ellawela et al., 2025; Hassan et al., 2025). Code assistants have studied the preview most closely, learning when to show a suggestion at all from acceptance feedback (Mozannar et al., 2024b), modeling the programmer’s interaction states and verification cost around it (Barke et al., 2022; Mozannar et al., 2024a), and surfacing proactive suggestions inside the editor for the user to take or drop (Zhao et al., 2025a; Chen et al., 2025e). The same preview-then-commit loop now wraps full agentic workflows and human-robot manipulation, where a co-planning interface or a proposal turn lets the human approve before anything takes effect (Mozannar et al., 2025; Yu et al., 2026a; Urli et al., 2018). A preview gives the user an opportunity to detect and reject an unsuitable proposal before execution, with attention and verification costs that depend on the proposal and workflow.

The presentation surface is itself channel-specific, and each channel forces a different shape for the unsolicited act. Dialogue agents deliver proactivity as injected turns whose timing and content are modeled explicitly, deciding when to claim a turn, what to contribute, and when a backchannel is enough (Deng et al., 2023b; Liu et al., 2025g; Lee et al., 2026). Wearable and augmented-reality agents shrink the same delivery to an in-ear whisper or a glanceable panel, jointly adapting what to say and how to surface it to the moment (Chen et al., 2025d; Lee et al., 2025b). An embodied agent in a home or workplace has yet another surface, where the act is physical and the design variable is the legibility of the robot’s motion, since intent-expressive movement makes the robot’s goal clear to the person it is acting around and marks a readable approach apart from one that is merely predictable (Dragan et al., 2013). For an unattended agent with no human in front of it the presentation surface becomes staged rollout (Davidovič & Beyer, 2018), undo-friendly execution with a confined blast radius (Patil et al., 2024), and the audit log that presents the act to an absent principal after the fact (Dong et al., 2024). The function stays constant across these surfaces while only the medium changes,

which is why the question of when to enter recurs in group and asynchronous settings under each one (Eckhaus et al., 2025; Liu et al., 2026b).

The last delivery variable is wording itself. People apply social rules to a system’s phrasing as if it were a social actor (Nass et al., 1994), so how an unrequested suggestion is framed, its phrasing, the explanation attached, and the confidence it displays, causally shifts whether the user accepts it and whether they over-rely on it (Zhang et al., 2020b). Two handles replicate well: first-person hedges and natural-language uncertainty lower overreliance on wrong answers (Kim et al., 2024b), and for inline code, highlighting the tokens most likely to be edited helps more than displaying raw generation probabilities (Vasconcelos et al., 2025). The same levers carry to the textual self-report an agent emits about an action it already took, where an explanation of why it acted is the action report in another setting, studied from collaborative-filtering justifications to a self-driving controller’s rationale and to the oversight explanations that agentic interfaces now owe an absent supervisor (Herlocker et al., 2000; Kim et al., 2018b; Jang & Sundar, 2026).

4.6 Adaptation

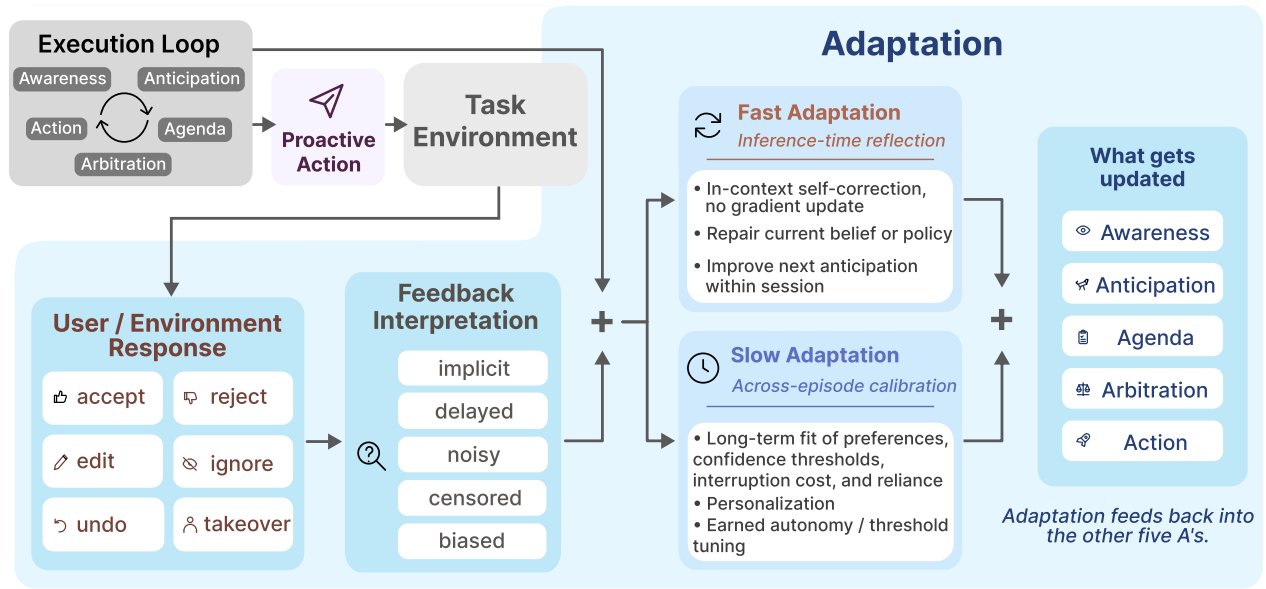


Figure 12 Adaptation: learning whether initiative was warranted. Successful execution alone does not establish that an intervention was useful, and user reactions must be interpreted in context. Fast reflection revises the assumptions behind the next decision; slower calibration adjusts intervention policies as preferences and circumstances change across interactions. The resulting behavioral evidence helps calibrate trust and delegation to the agent’s demonstrated reliability.

Adaptation is the sixth A and the slow outer loop that closes the proactive cycle: once an agent has sensed, anticipated, planned, gated, and acted, it must learn from how that act landed so that its next anticipation and its next gate are better. Self-initiated action changes both the evidence available for learning and the behavior that must be improved. Feedback may be implicit, delayed, or missing, and adaptation must assess the choice to intervene alongside the quality of execution. We trace three loops at widening timescales (Figure 12): harvesting an implicit, biased signal out of the consequences of acting first (§4.6.1), repairing the agent’s own predictive model in-context before it next fires (§4.6.2), and the across-episode calibration of trust, personalization, and earned autonomy that keeps proactivity sustainable (§4.6.3).

4.6.1 Implicit Feedback and Self-Correction

Feedback on self-initiated interventions can be sparse and difficult to interpret: a dismissed suggestion is a noisy label (Opfermann et al., 2017) whose meaning cannot be read off the outcome, since the user may have been busy, distracted, or negotiating, an averted incident is an unobservable counterfactual that leaves nothing to score (Schwall et al., 2020), and credit for an anticipatory action arrives late or never. The same scarcity appears wherever the agent acts first: a user-facing assistant guessing whether its unprompted nudge helped must read implicit and noisy cues from the

interaction log (Liu et al., 2025h; Shen et al., 2025), and an operations or driving system that preempted a failure can no longer prove the failure was coming. Both learn from signals that are implicit, biased, and censored, and the founding mixed-initiative account already named the difficulty of inferring whether an unrequested service was wanted (Horvitz, 1999). Adaptation therefore opens with a harder problem than choosing a learning rule, which is manufacturing a usable training signal out of silence, and recent agents take it up directly by actively probing the environment for the feedback their own actions did not return (Su, 2026a; Feng et al., 2026c).

Learning from these signals draws on established methods for implicit feedback and partially observed outcomes, whose assumptions must be examined in the context of self-initiated interventions. The information-retrieval and recommender-systems tradition established long ago that interaction exhaust, the clicks, skips, and dwell times users leave behind, substitutes for explicit labels (Joachims, 2002; Hu et al., 2008), and it catalogued the pathologies that the proactive setting inherits, position bias as the ancestor of dismissal-is-not-rejection (Joachims et al., 2005) and delayed or censored conversions where a non-response cannot yet be scored (Chapelle, 2014). The decision-theoretic side supplies the estimators that turn logged, partial feedback into policies: counterfactual risk minimization and off-policy bandit methods reweight logged decisions into improved ones (Swaminathan & Joachims, 2015; Bottou et al., 2013), and interaction-grounded formulations recover a latent reward when no explicit reward is logged at all, including implicit physiological signals of subjective interest (Xie et al., 2021; Shih et al., 2017). That generality lets the same estimators serve a user’s accept and reject history and a robot operator’s record of takeovers and non-interventions alike (Spencer et al., 2021; Ho et al., 2020), and learning directly from human interventions and corrections extends naturally to multi-robot and co-creative teams (Seraj et al., 2023; Lin et al., 2024). These are imported tools, and the proactive contribution is the function they are pointed at, which is deciding when to intervene.

Agents can use the recorded consequences of self-initiated interventions, together with user and operator responses, as feedback for adaptation. For a user-facing agent the systems treat acceptance, dismissal, and edit telemetry as de facto reward: code-completion acceptance rates drive when a suggestion is worth showing (Mozannar et al., 2024b; Ziegler et al., 2022), reply and notification traces are mined so that a dismissal becomes a preference signal (Pielot et al., 2014; Mehrotra et al., 2016), and a fabric-sensing assistant refines its behavior from implicit textual feedback with no explicit ratings (Liu & Chen, 2024). For an agent managing a system the equivalent signals are operator overrides, alert triage decisions, and rollbacks, where a disengagement corpus records every human takeover (Favarò et al., 2018), analyst triage actions on alerts are learned into an automatic prioritisation policy (Turcotte et al., 2025), and Gandalf trains its deployment go/no-go classifier on historical cases with feedback from component teams (Li et al., 2020). The move is the same in both, converting the exhaust of one’s own interventions into a graded supervision signal, with full awareness that the signal carries exactly the bias the asymmetry above named.

The modern move is to train proactivity directly and end to end on the signal sources just instrumented, the source-side counterpart to the within-episode reasoning of §4.6.2 over the same data. For user-facing agents the recipe fits a reward model to accept and reject and edit traces and fine-tunes toward the intervention behavior users actually keep: the canonical exemplar collects human accept and reject annotations over predicted proactive tasks into a benchmark, trains a reward model on them, and uses it to elicit proactiveness (Lu et al., 2025a), while related work infers latent preferences from user edits (Gao et al., 2024) and pushes the same recipe to on-device override traces at scale (Feng et al., 2026b). For agents acting on a system the optimization runs against execution and environment outcomes, grounding the policy in code-execution feedback, exploration traces, or self-produced environmental reward (Gehring et al., 2025; Song et al., 2024b; Alhajir et al., 2025), and reinforcement-learned security and operations agents tune their preemptive actions against pipeline outcomes (Anugula et al., 2025); multi-objective formulations now optimize productivity, proactivity, and personalization jointly (Sun et al., 2025b). Both branches presuppose that the traces already exist, and for proactive behavior they often do not, since an agent that has never been permitted to act unprompted leaves no acceptance log to fit a reward model to. Synthesis is the standard recourse, producing tailored instruction data for capabilities that collected corpora underrepresent (Wang et al., 2024k) and constructing multi-turn tool-use trajectories where live interaction traces are unavailable (Yin et al., 2025), at the cost of substituting the generator’s coverage assumptions for the user population’s. The caveat from the feedback asymmetry never disappears, since a policy trained on biased acceptance data inherits that bias, which is precisely why the calibration of §4.6.3 is needed downstream.

4.6.2 Inference-Time Reflection

A mature general toolbox converts the same signals into behavioral change purely in-context, with no gradient step, by having the agent verbalize a critique of its own trace, revise, and retry, the recipe sometimes called verbal reinforcement

learning (Shinn et al., 2023). Relevant mechanisms include self-generated reflection on past attempts kept in an episodic buffer (Shinn et al., 2023), iterative self-refinement of a single output (Madaan et al., 2023), tool-grounded critique that checks claims against external evidence (Gou et al., 2024), and plan-level repair that separates in-plan from out-of-plan errors (Sun et al., 2023; Wang et al., 2024a). This approach has already been wired into proactive assistants and self-improving agents that monitor their own execution and act on what they find (Sun et al., 2025a; He et al., 2025a; Orimo et al., 2025). What makes it matter for proactive agents is what they point it at, developed next.

The proactivity-specific use of reflection turns the question from whether the answer was good to whether acting at all was the right call, reopening the arbitration variables of §4.4: should I have intervened, was the timing right, was the autonomy level appropriate. Concretely this appears as an explicit reasoned act, wait, or clarify step inserted before behaving, as in proactive chain-of-thought prompting that plans the goal over a reasoning chain before responding (Deng et al., 2023c,b), and as a critic that scores initiative itself in place of output content, the clearest instance being a dedicated critic that grades how proactive a turn should have been (Wang et al., 2025l). Reflective clarification-need prediction, deciding by introspection whether the agent is missing information it should ask for, is the same move applied at the input boundary (Zhang et al., 2024h; Qian et al., 2024), and a metacognitive layer that predicts its own impending failure and hands off to a human is its safety-side counterpart (Xu, 2025). In every case the object of critique is the decision to act, which is what distinguishes proactive reflection from generic self-refinement (Lu et al., 2025a).

When a proactive prediction misses, whether a rejected suggestion or a failed preemptive action, reflection’s job is to repair the model that produced it, in-context, so that the next anticipation within the session is better, closing a fast loop back to Anticipation (§4.2) without waiting for offline retraining. For a user-facing agent this means updating an inferred preference or theory-of-mind belief after a misfire (Gao et al., 2024; Zhang et al., 2024a), and for an agent acting on a system it is the well-developed pattern of scientific and driving agents that run hypothesis-revision and belief-update loops, observing an experimental or environmental outcome, diagnosing the mismatch, and rewriting the working model before the next decision (Wen et al., 2024; Boiko et al., 2023b; Lu et al., 2024; Wu et al., 2025c). Repairing the predictive model before it next fires is what makes this reflection prospective, since the critique migrates from after-the-fact to before-the-fact and anticipates the agent’s own error, which is itself a proactive move (Wang et al., 2026b; Han et al., 2025c; Yu et al., 2025). Two limits bound it, and both bite hardest in proactive settings. First, ungrounded reflection can entrench errors in place of fixing them, since intrinsic self-correction without external evidence sometimes degrades performance (Huang et al., 2024). Second, reflection costs latency and compute, which collides with the timing constraints of §4.4.3, because overthinking displaces the very environmental interaction the agent was supposed to act on (Cuadron et al., 2025); real-time agents accordingly split fast and slow reasoning or precompute reflection ahead of need so that deliberation does not cost the moment to act (Zhang et al., 2025g; Liu et al., 2026l; Lin et al., 2025). A proactive agent that reflects too long has, in effect, failed to act first.

4.6.3 Long-Term Calibration, Personalization, and Trust

As the user and environment change, the agent re-estimates across episodes the standing quantities on which the earlier mechanisms depend. These include the anticipation models of §4.2, which capture user preferences and world dynamics, and the confidence thresholds, interruption costs, and utility models of §4.4.2. What it re-fits is the agent’s standing relationship with its user along three coupled axes: how far the user’s trust licenses it to act, how much autonomy its track record has earned, and whose preferences it should personalize to. All three are fit from the accumulated feedback of §4.6.1.

The binding constraint is appropriate reliance more than raw capability: over-trust breeds complacency as the user stops checking (Parasuraman & Manzey, 2010), under-trust breeds disuse as the user disables the proactivity altogether (Parasuraman & Riley, 1997), and the agent’s standing authority to act first is the slowly-moving budget between them. The human-factors theory of reliance supplies the frame (Lee & See, 2004), and it carries across settings, from an individual deciding how much to let an assistant preempt them (Tussyadiah & Wang, 2014) to an organization deciding how much intervention authority an operations, driving, or clinical system has earned (Gombolay et al., 2016). In a planning-task study, fully autonomous system behavior did not establish adequate trust, and the effects of proactive dialogue on perceived competence and reliability depended on task difficulty (Kraus et al., 2020). A study of human-robot collaboration found that greater attributed agency was associated with more negative attitudes when perceived user control was low (Zafari & Koeszegi, 2020). Trust calibration can inform proactive interaction, and socially-aware training that rewards trust alongside task success is one approach (Kraus et al., 2023).

Autonomy should be graduated and evidence-indexed, adjusted as performance history accumulates: adaptive automation shifts the level of agent authority based on that history (Kaber & Endsley, 2004), and deployment practice increasingly treats authority as statistically earned, demonstrating reliability over a large operating record before authority is widened (Kalra & Paddock, 2016). This is the dynamic counterpart to the static rungs of §4.5.1, first developed in safety-critical operations, where those are the levels and this is movement along them over time, and a shared-workspace study shows a situation-adapted agent outperforming any fixed autonomy level (Salikutluk et al., 2024). The proactive-specific anchor is that an agent can choose its proactivity level as a function of predicted trust, escalating initiative only as far as the evidence of earned reliance permits (Kraus et al., 2021), and systems now monitor for over- and under-reliance and emit calibration cues to steer it back (Okamura & Yamada, 2020); a related discipline appears in self-verifying discovery agents that gate each next step on reproducible intermediate results before proceeding (Ghafarollahi & Buehler, 2025).

Across episodes, an agent can use accumulated feedback to personalize its proactivity, updating the user model of §4.2.1 and the memory of §4.1.3. The progression runs from per-user interruption-cost models (Saponaro et al., 2017; Horvitz et al., 2004), through fleet-scale notification bandits that learn when a recurring alert is welcome (Yancey & Settles, 2020), to AI assistants that write rated intervention outcomes into persistent memory and continually re-personalize their initiative (Wang & Jiang, 2026). As preferences change, an intervention that was welcome last month may now be annoying, so the agent needs to revise its model over time. Recent agent work makes the payoff concrete with per-suggestion ratings across many personas that measurably improve satisfaction over time (Kim et al., 2026b; Xuan et al., 2026) and multi-objective reinforcement learning that jointly optimizes productivity, proactivity, and personalization (Sun et al., 2025b).

The deeper difficulty that spans all three axes is that the target moves under the agent’s own influence: people change their behavior and even their language toward systems they interact with (Zhang & Yu, 2025), so the distribution the agent calibrated against shifts as a result of the very interventions it made, the performative-prediction problem (Perdomo et al., 2021). Calibration is thus a relationship maintained over time, and a moving one, sustained by mutual and symbiotic adaptation in which the human and agent adjust to each other across the interaction (Nikolaidis et al., 2017; Nwankwo et al., 2026). The harder questions this raises, the self-reinforcing feedback loops that drive a policy toward degenerate attractors (Jiang et al., 2019) and the explicit repair a broken-trust relationship needs before it recovers (Kox et al., 2021; de Visser et al., 2020), are governance concerns; here they mark the boundary where mechanism hands off to oversight, including cooperative human-agent oversight games that keep control without modifying the agent (Overman & Bayati, 2026).

Feedback on unnecessary or missed interventions helps calibrate Anticipation and Arbitration, while also informing what the agent monitors, which goals it maintains, and how it executes them. Errors in predicting needs or setting intervention thresholds can produce over-eager, harmful assistance (Lu et al., 2025a; Wang et al., 2026l). The right mental model is therefore an adaptation loop whose objective is the agent’s own faithfulness, an operational honesty about what it knows, how confident it is, and whether what it did was correct. These mechanisms observe trust and reliance, but neither is the quantity to optimize, because both live in the user’s perception, which an agent can manipulate by flattering (Sharma et al., 2025), by deferring until it never annoys, or by concealing the failures it should disclose, the reward-hacking trap of optimizing a proxy (Skalse et al., 2025). Repeated honest interaction can help users calibrate their reliance on an agent. Accurate reporting of uncertainty and failures becomes especially important when the volume of autonomous behavior limits human review or when the agent’s capabilities exceed the user’s ability to assess its outputs. Such feedback can guide revisions to the other five mechanisms as the user and environment change. With the full mechanism assembled, §6 turns to how each deployment domain weights the six components differently.

5 Evaluation

Table 2 Taxonomy of evaluation metrics for proactive AI agents, organized by the primary evaluation focus at the task, mechanism, and user levels. Each row lists the distinct evaluation directions argued for in the corresponding subsection, with a small set of representative anchor works rather than an exhaustive citation list.

Level	Evaluation Focus	General Directions for Evaluation	Representative Works
Task Utility	Task Success	Goal completion rate against reference or verifiable outcomes; domain-specific task outcomes; per-intervention (turn-level) success rate	(Walker et al., 1997; Maniktala et al., 2020; Pu et al., 2025a; Wu et al., 2025e; Andolina et al., 2018)

Continued on next page

Table 2 (continued)

Level	Evaluation Focus	General Directions for Evaluation	Representative Works
	Task Efficiency	Agent computational cost (inference latency, token usage, triggering frequency); user interaction cost (turns, interruptions, correction effort); quality-adjusted efficiency (benefit per unit cost)	(Chai et al., 2026a; Lee et al., 2025a; Pu et al., 2025a; Hu et al., 2026a; Zhang et al., 2025e)
Mechanism-Level	Context Modeling Efficacy	Retrieval precision and recall over long or multimodal context; knowledge-update accuracy and stale-information rate over sustained interaction; context efficiency (compression ratio, cost per unit accuracy); pre-decision evidence coverage	(Wang et al., 2024i; Liu et al., 2024b; Wu et al., 2025b; Uddin et al., 2026; Maharana et al., 2024; Yang et al., 2026c)
	User State and World State Anticipation	Intent, goal, and preference accuracy (precision, recall, F1); prediction lead time and false-alarm rate for emerging needs; ambiguity-detection F1 and clarification utility; next-state and event-forecast accuracy	(Qian et al., 2025a; Lu et al., 2025a; Tang et al., 2026b; Yang et al., 2025d; Grauman et al., 2022)
	Planning Quality	Objective/agenda-selection accuracy (including hidden requirements); plan validity and subgoal coverage; constraint satisfaction and preference utility; plan-revision success under new constraints	(Zhang et al., 2026b; Valmeekam et al., 2023; Xie et al., 2024a; Liu et al., 2026i; Qin et al., 2025b)
	Intervention Decision and Timing	Intervention precision, recall, and F1 (act, ask, defer, or abstain); autonomy calibration (deferral accuracy, risk-coverage); trigger-timing accuracy (valid-window, early/late-trigger rate); unwanted-initiative or persistence-after-rejection rate	(Zhang et al., 2024h; Trinh et al., 2026; Mozannar & Sontag, 2020; Chen et al., 2025d; Wang et al., 2026l)
	Intervention Delivery and Presentation	Content correctness and execution/state-change accuracy; delivery-mode appropriateness relative to risk and reversibility; comprehension and legibility of agent intent; editability, cancellation, and reversal success	(Wang et al., 2025k; Tang et al., 2026b; Kong et al., 2026; Pu et al., 2025a; Dragan et al., 2013)
	Adaptation and Generalization	Personalization gain and learning/sample efficiency; recovery speed and stale-memory/obsolete-information rate under drift; generalization to unseen users, tasks, or environments	(Kim et al., 2026b; Liang et al., 2026; Uddin et al., 2026; Li et al., 2026a; Guo et al., 2026b)
User Experience	User Workload	Subjective workload (NASA-TLX, Workload Profile, SWAT); behavioral disruption (interpretation time, recovery time, task-switch cost); net effort removed vs. newly introduced	(Hart & Staveland, 1988; Tsang & Velazquez, 1996; Pu et al., 2025a; Bailey & Konstan, 2006)
	Usability and Experience	Standardized usability/UX scores (SUS, UEQ); perceived usefulness, naturalness, and collaboration quality; intrusiveness and privacy concern	(Brooke, 1995; Laugwitz et al., 2008; Walker et al., 1997; Villa et al., 2025; Horvath & Greenberg, 1989)
	Level of Control	Perceived agency and preferred autonomy level; override, modification, and reversal success rate; gap between formal and effective control	(Pu et al., 2025a; Walker & Whittaker, 1990; Parasuraman et al., 2000; Dietvorst et al., 2018)
	Trust and Reliance	Trust-in-automation scales, calibrated to actual reliability; appropriate reliance (correct-accept / correct-reject rate); critical-engagement behavior (verification, error detection); longitudinal trust trajectory	(Jian et al., 2000; Lee & See, 2004; Schemmer et al., 2022; Desai et al., 2013; Lee et al., 2025c)

Evaluating proactive assistance requires evidence about **task utility**, **mechanism-level evaluation**, and **user experience** (Figure 13). These directions examine, respectively, whether proactive assistance improves the outcome of the task, whether the mechanisms producing that assistance operate effectively, and whether the resulting behavior is acceptable and beneficial from the user’s perspective. For example, a system may successfully complete a task while intervening unnecessarily, relying on inaccurate assumptions about the user, or imposing interaction costs. An agent may also correctly infer a user need or choose an appropriate moment to intervene without ultimately improving the task outcome. Users may perceive technically effective assistance as intrusive, unpredictable, or difficult to comprehend when the agent acts without an explicit request.

Proactive assistance must improve task outcomes enough to justify the costs of intervention. Task utility covers

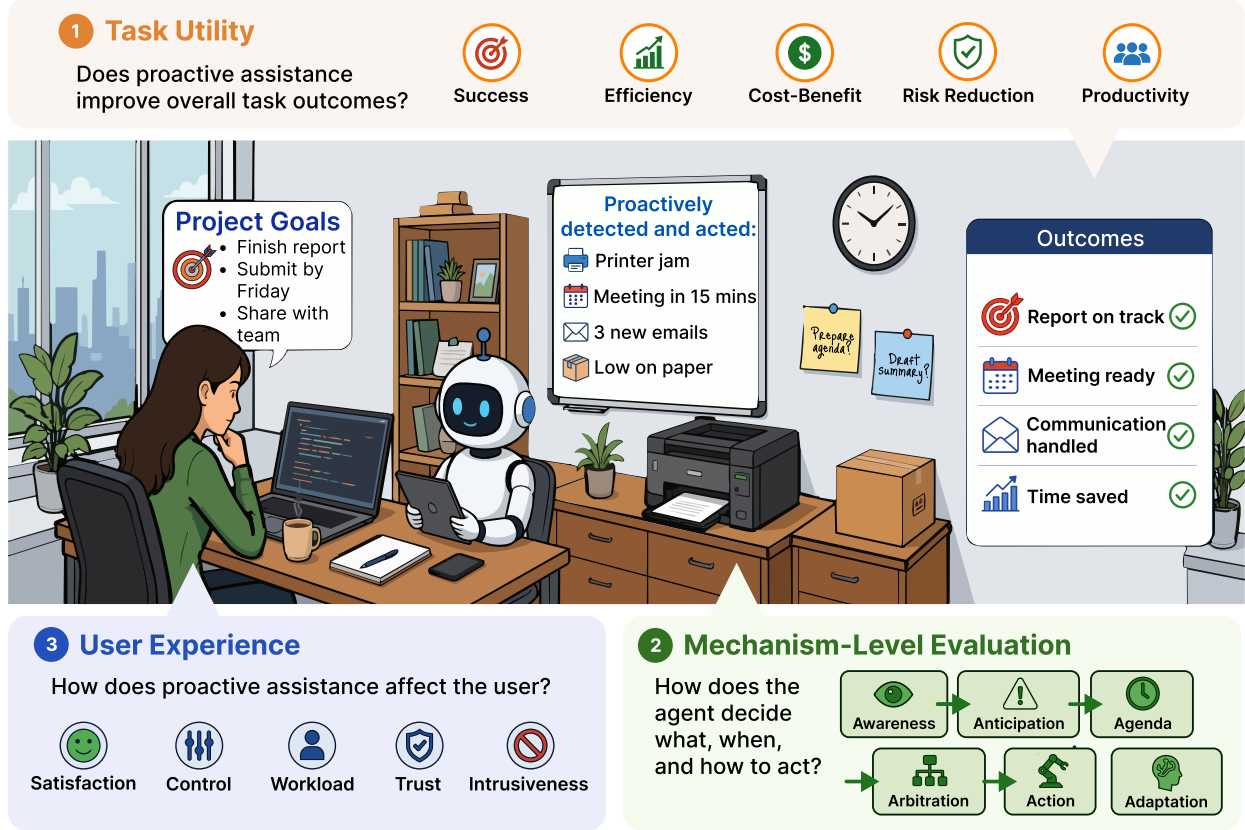


Figure 13 Evaluating proactiveness requires evidence beyond task completion. Task utility establishes whether an intervention improves outcomes, mechanism-level evaluation examines how effectively the underlying capacities operate, and user experience assesses its effects on workload, control, and trust. None substitutes for the others: a successful intervention may be poorly justified or burdensome, while strong performance on individual mechanisms does not by itself establish practical benefit.

effectiveness and efficiency in terms of time, effort, resources, and interaction cost (Table 2). Useful outcomes include helping users accomplish any of their subgoals or complete tasks successfully (Maniktala et al., 2020; Walker et al., 1997; Chen et al., 2025e; Pu et al., 2025a); reducing the time, effort, or resources required from the user and the system (Andolina et al., 2018; Chen et al., 2019; Lee et al., 2025a; Pu et al., 2025a); or preventing undesirable outcomes by identifying errors, risks, or situations that require intervention (Artikis et al., 2017). Outcome measures describe the performance of the complete system. Attributing these outcomes to particular mechanisms and assessing user satisfaction requires mechanism-level and user-experience measures.

Errors in proactive assistance can arise at several stages of the decision process. Following the mechanism taxonomy in §4, evaluation should assess how accurately the agent senses the environment (Wang et al., 2024i; Liu et al., 2024b; Wu et al., 2025b; Uddin et al., 2026; Maharana et al., 2024); whether it understands the user’s preferences and latent intent, and changes in the world state (Lieberman, 1995; Blaylock & Allen, 2005; Deng et al., 2022; Qian et al., 2025a; Chen et al., 2026g; Grauman et al., 2022); whether it develops appropriate goals, subgoals, and executable plans (Deng et al., 2023c; Kong et al., 2026; Nathani et al., 2026; Zhang et al., 2026b; Wu et al., 2025e); whether it correctly decides when to act, ask for clarification, or remain silent (Zhang et al., 2024h; Lu et al., 2025a; Trinh et al., 2026; Wang et al., 2026l; Pu et al., 2025b); whether the resulting intervention is relevant and correctly executed (Pu et al., 2025a; Tang et al., 2026b; Kong et al., 2026); whether it occurs at an appropriate moment (Chen et al., 2025d; Samarinas & Zamani, 2024; Ding et al., 2026b; Zhao et al., 2026b; Ran et al., 2026); and whether the agent can adapt across users, preferences, tasks, and interactions (Kim et al., 2026b; Sun et al., 2025b; Hao et al., 2025; Li et al., 2026d). An undesirable intervention may arise from an incorrect inference about the user’s need, an unnecessary decision, poor timing, or failure to incorporate prior feedback. These failures may produce similar observed outcomes while reflecting different weaknesses. Mechanism-level evaluation helps attribute success and failure to the decisions that produced

them (Table 2).

Assistance initiated without an explicit request can increase workload, feel intrusive, reduce user control, or encourage inappropriate reliance. Evaluation should therefore assess whether assistance reduces mental effort, frustration, interpretation time, and task-switching costs (Hart & Staveland, 1988; Tsang & Velazquez, 1996; Reid & Nygren, 1988; Lee et al., 2025a; Pu et al., 2025a); whether the interaction is perceived as useful, usable, collaborative, and respectful of privacy (Brooke, 1995; Laugwitz et al., 2008; Walker et al., 1997; Andolina et al., 2018; Liu et al., 2024e, 2023c; Ma & Ziegler, 2024; Horvath & Greenberg, 1989); whether users can accept, modify, override, or reverse interventions and retain an appropriate sense of agency (Walker & Whittaker, 1990; Parasuraman et al., 2000; Pu et al., 2025a; Villa et al., 2025); and whether trust and reliance remain calibrated to the agent’s actual reliability (Jian et al., 2000; Lee & See, 2004; Kraus et al., 2020; Schemmer et al., 2022; Desai et al., 2013). Workload, usability and experience, level of control, and trust and reliance provide complementary measures of user experience (Table 2), consistent with recent calls for more human-centered evaluation of proactive agents (Deng et al., 2024b; Kraus et al., 2025).

Existing evaluations often emphasize only one of these directions or blur their distinct roles: task success may overlook flawed intervention processes, mechanism metrics may not establish user benefit, and human-subject studies may not identify which system decisions actually shaped user perceptions. Evaluation is also uneven across settings, with more standardized benchmarks for conversational agents than for embodied, mobile, visual, and cross-application systems.

5.1 Task Utility Metrics

Additional monitoring, reasoning, interaction, and delay can raise the cost of proactive assistance even when they improve the final result. Task utility therefore includes **task success** and **task efficiency**. Success asks whether the agent provides a final meaningful outcome; efficiency asks whether the gain was worth its cost.

5.1.1 Success Metrics

Proactive assistance can admit several valid responses when the trajectory or endpoint is incompletely specified. An agent may provide information, ask a clarifying question, prepare an action, complete part of the task, or appropriately remain silent. Evaluation should therefore judge whether the agent advanced the user’s goal. Reference trajectories or answers can support this judgment when they capture the relevant goal. The goal itself may also be implicit, incomplete, or evolving. A technically correct action may be irrelevant to what the user actually needs, while a partial intervention may still be valuable if it prevents an error or saves later effort. Success may therefore depend on goal completion, correctness, relevance, and risk reduction. For vague requests, even determining whether the agent succeeded may require contextual or model-based judgment.

Whole-task evaluation has a long history in dialogue and mixed-initiative systems. Early work measures task success through agreement between the agent’s inferred result and a ground truth scenario key (Walker et al., 1997), while early mixed-initiative transportation assistants report the proportion of sessions that successfully complete the intended task (Ferguson et al., 1996). Such reference-based evaluation works well when the goal is explicit and verifiable. Coding assistants can use completed assignments or passed test cases as concrete outcome signals (Pu et al., 2025a; Chen et al., 2025e), and agents acting in an environment with stable and valid feedback or reward signals should be evaluated similarly (Liu et al., 2026e,g); proactive search systems may examine whether suggested information influences the subsequent conversation (Andolina et al., 2018); and task-oriented dialogue may combine goal completion with domain-specific outcomes, such as the sale-to-list ratio in negotiation (Wu et al., 2025e). The appropriate definition of success thus depends on what useful assistance looks like in the target setting.

A session-level score aggregates proactive decisions into one result. More granular outcomes can improve credit assignment. For example, the Hint Justification Rate measures the proportion of provided hints, including on-demand and proactive hints, that students correctly justify by deriving the hinted step (Maniktala et al., 2020). Intervention-level scoring parallels the use of dense process signals in reinforcement learning with verifiable rewards (Lightman et al., 2023). Such measures record intervention-level uptake or local task progress. Establishing why an intervention was effective requires examining whether the agent inferred the right need, chose the right moment, and delivered the right content.

5.1.2 Efficiency Metrics

An agent may continuously monitor context, infer latent needs, request additional information, and sustain several rounds of interaction before offering assistance. These steps may improve task success, but they also consume user attention, time, and computational resources. Evaluation should assess task success together with the resources that the agent and the user expend to achieve it.

Task-completion time depends partly on how quickly users interpret information, make decisions, and respond to the agent. These human factors limit the system’s control over total task duration. The agent’s sensing, reasoning, communication, and intervention policies affect both computational cost and the amount of work imposed on the user. For example, an agent that reasons more frequently may identify better opportunities to help, but it may also consume more tokens and increase latency; similarly, an agent that repeatedly asks clarifying questions may improve its understanding of the user’s goal while increasing interaction burden. Task efficiency should therefore be treated as a human-AI co-design objective: the system should determine when additional computation or interaction is worth its cost.

First, agent-side efficiency concerns how much computation is required for the agent to continuously determine whether, when, and how to intervene. Proactive agents may continuously sense the environment, update state and memory, generate possible actions, and evaluate whether intervention is appropriate, *even* when they ultimately remain silent. This makes the cost of deciding not to act an important part of efficiency. Proactive GUI agents, for example, can reduce such overhead by reducing repeated processing of complete visual histories through bounded context and structured state (Chai et al., 2026a). Lightweight triggering mechanisms can similarly avoid invoking a large model after every observed user event (Liu et al., 2026o). Evidence from general GUI agents further suggests that additional reasoning can increase token consumption and latency without reliably improving task success (Zhang et al., 2025e). Relevant measures therefore include inference latency, token usage, memory and energy consumption, monetary cost, sensing frequency, triggering frequency, and the computational cost of intervention decisions.

Second, these algorithmic choices also create user-facing efficiency costs that the agent must account for when deciding whether an intervention is worthwhile. Proactivity can reduce downstream task effort, but it can also introduce interruptions, clarification turns, verification, or correction work. For instance, the proactive AR assistant studied by Lee et al. (2025a) sometimes requires longer interactions. In proactive programming support, Pu et al. (2025a) find no significant difference in total task-completion time, suggesting that additional assistance does not necessarily translate directly into faster completion. In those settings, task duration alone remains an incomplete measure: a longer interaction may produce a better outcome, while a shorter one may simply reflect premature or ineffective assistance. The agent should therefore reason about the expected benefit of an intervention relative to both its computational cost and the human effort it induces. Useful measures include interaction turns, interruptions, correction and verification effort, independent completion times for the agent and the user, and ultimately the amount of user effort saved for a given level of task quality.

Efficiency ultimately concerns the balance between resources consumed and useful assistance delivered. Proactive preparation may reduce the number of interaction turns and the effort needed to complete a task (Hu et al., 2026a), but poorly timed assistance can introduce extra interruption, resumption, and verification costs for users even when its content is correct (Mark et al., 2008; Iqbal & Bailey, 2010). These trade-offs constrain mobile, wearable, and embodied agents, where always-on monitoring and idle-time computation (§4.1.4) compete for limited energy, device resources, and access to personal context. Evaluation should therefore relate useful assistance to the time, attention, and computation it consumes.

5.2 Mechanism-Level Metrics

The second level evaluates the internal mechanisms of a proactive agent to identify where its behavior succeeds or fails. Mechanism-level metrics cover six categories: Context Modeling Efficacy; User State and World State Anticipation; Planning Quality; Intervention Decision and Timing; Intervention Delivery and Presentation; and Adaptation and Generalization. These categories align with the six mechanisms in §4: Awareness models the current user and environment state; Anticipation predicts intentions, needs, and future states; Agenda selects and organizes goals; Arbitration decides whether and when to intervene; Action delivers an understandable and controllable intervention; and Adaptation improves these behaviors through experience. Although these categories are interdependent, they provide a diagnostic structure for locating the mechanism most directly associated with a measured capability or failure.

Downstream task success can leave the underlying mechanism unmeasured. Anticipation and Action studies often evaluate the final intervention without separately assessing the prediction or delivery process that produced it.

5.2.1 Context Modeling Efficacy

Maintaining *Awareness* requires an accurate representation of the current situation drawn from an extended interaction history. This requires the agent to retain relevant observations, integrate evidence distributed across time and modalities, update information as circumstances change, and distinguish observed facts from uncertain inferences. Evaluation should assess context quality together with the cost of maintaining it as the interaction history grows.

A first question is whether relevant evidence remains accessible as the interaction becomes longer and more complex. Long context benchmarks provide controlled tests of this capability. MM-NIAH evaluates retrieval, counting, and reasoning over textual and visual evidence distributed across long multimodal contexts (Wang et al., 2024i), while MMNeedle tests target retrieval in large visual contexts, including cases where the requested evidence is absent (Wang et al., 2025b). MileBench extends this setting to realistic interleaved image text inputs (Song et al., 2024a). In text only settings, LongBench and RULER evaluate long document retrieval, multi-hop reasoning, and aggregation (Bai et al., 2024; Hsieh et al., 2024), while *Lost in the Middle* shows that evidence accessibility can also depend on its position within the context (Liu et al., 2024b). Together, these benchmarks motivate measures such as retrieval precision and recall, multi evidence recall, distractor robustness, and performance across context lengths and evidence positions.

During sustained interaction, the contextual representation must remain correct as information changes. LoCoMo evaluates recall, temporal reasoning, and summarization over extended conversations (Maharana et al., 2024), while LongMemEval examines cross-session reasoning, knowledge updates, and abstention (Wu et al., 2025b). MemoryAgentBench further considers selective forgetting and test-time learning (Hu et al., 2026d), and multimodal settings are studied in Mem-Gallery (Bei et al., 2026). Memora focuses specifically on evolving user information and introduces Forgetting-Aware Memory Accuracy to penalize reliance on invalidated memories (Uddin et al., 2026). These settings evaluate whether the context representation remains consistent with the user’s current state.

Maintaining such awareness also raises an efficiency question. Supplying the complete interaction history at every step may preserve information, but becomes increasingly costly as the history grows. Context compression methods such as Selective Context and LongLLMLingua retain more informative portions of the input while reducing the amount of context processed by the model (Li et al., 2023c; Jiang et al., 2024a). Persistent memory approaches address the same problem across interactions: MemGPT manages information across different memory tiers (Packer et al., 2023), while PlugMem turns agent trajectories into structured graphs of reusable and generalizable facts and skills (Yang et al., 2026c). Context modeling should therefore be evaluated jointly in terms of *context quality* and *context cost* (Yang et al., 2026c). Relevant efficiency measures include retrieved or processed tokens, compression ratio, memory growth, retrieval and update latency, and computational cost. More informative evaluations can report quality under a fixed context budget, or the amount of computation required to reach a given level of retrieval or downstream accuracy.

These measures are most useful when applied to the contextual representation available *before* a proactive decision is made. Quality can be characterized through retrieval precision and recall, multi-hop and temporal reasoning accuracy, knowledge-update accuracy, stale-information retention, selective forgetting accuracy, and false-positive or abstention rates when relevant evidence was never observed (Hsieh et al., 2024; Wu et al., 2025b; Uddin et al., 2026; Wang et al., 2025b). LongMemEval-V2 provides a useful formulation in which a memory system processes historical trajectories and returns compact evidence for downstream reasoning (Wu et al., 2026a). A proactive agent benchmark could similarly measure *pre-decision evidence coverage*, or the proportion of decision-relevant facts represented when an opportunity to assist arises, together with an oracle context gap comparing downstream performance using the agent’s context against annotated gold context. Paired with context size or retrieval cost, these measures reveal whether a proactive failure stems from missing, outdated, or inefficiently maintained awareness.

5.2.2 User State and World State Anticipation

An agent may need to infer task-relevant states that are not yet directly observable through its *Anticipation* mechanism. User-state anticipation estimates the user’s latent goals, preferences, constraints, and emerging assistance needs from partial interaction evidence. World-state anticipation forecasts how the surrounding task environment is likely to evolve.

User state anticipation ranges from recovering a user’s current latent state to identifying needs that become apparent only

over time. Dialogue state tracking provides the basic formulation. The Schema Guided Dialogue benchmark evaluates active intent accuracy, requested slot F1, average goal accuracy, and joint goal accuracy (Rastogi et al., 2020). More recent benchmarks extend this setting beyond explicitly stated goals. IN3 evaluates recovery of missing information from underspecified instructions through Vagueness Judgment Accuracy, Missing Details Recover Rate, and Summary Intention Coverage Rate (Qian et al., 2024); UserBench measures whether incrementally revealed preferences are successfully elicited and reflected in the agent’s choices (Qian et al., 2025a); and LifeSim-Eval evaluates recognition and fulfillment of explicit and implicit intentions, including intention completion over long horizon user trajectories (Duan et al., 2026a). Across these settings, useful measures include intent or goal accuracy, slot and preference level precision, recall, and F1, as well as macro-F1 when intent classes are imbalanced. When several interpretations remain plausible, Recall@ k , mean reciprocal rank, or other ranking metrics can evaluate whether the correct user state remains among the agent’s plausible hypotheses.

Assistance *before* an explicit request requires recognizing the relevant user state in advance. ProactiveBench evaluates whether an emerging situation warrants assistance using accepted versus rejected proactive offers and reports F1 for assistance prediction (Lu et al., 2025a). ProAgentBench further separates *when* assistance is needed from what assistance is anticipated, evaluating the former with accuracy, precision, recall, and F1 and the latter with intention category accuracy and semantic similarity to the user’s subsequent request (Tang et al., 2026b). PIRA-Bench considers multiple interleaved intents in noisy activity streams and measures intent precision, recall, and F1 together with false positives during periods in which no actionable intent is present (Chai et al., 2026a). JIR-Arena evaluates real-time prediction of users’ information needs for proactive information recommendation (Yang et al., 2025d). Beyond final prediction accuracy, evaluation can therefore measure accuracy over observation prefixes, prediction lead time, time to stable intent recognition, and false alarms before a genuine need emerges.

Because latent user states are often underdetermined, anticipation should also evaluate whether the agent represents uncertainty appropriately. CLAMBER tests whether models recognize ambiguous information needs and generate clarifications that resolve the missing information (Zhang et al., 2024g), while INTENT-SIM studies uncertainty-based prediction of when clarification is useful and evaluates its effect on downstream performance (Zhang & Choi, 2025). Appropriate metrics therefore include ambiguity-detection F1, entropy over candidate intents, calibration error or Brier score, and the expected reduction in uncertainty or downstream error from additional information. To isolate errors in user-state inference, a fixed downstream policy can also be evaluated using inferred versus annotated user states, with the resulting oracle gap indicating how much performance is lost because of imperfect anticipation.

Changes in the external environment may create, remove, or alter opportunities for assistance. World-state anticipation forecasts these changes. Action anticipation benchmarks provide established examples. EPIC-KITCHENS evaluates upcoming verbs, nouns, and actions using top- k prediction and class-balanced recall measures (Damen et al., 2018), while Ego4D evaluates both short-term object interactions and longer future action sequences using Top-5 mean average precision for short-term object interactions and edit distance for long-term action sequences (Grauman et al., 2022). In interactive agent environments, Web Agents with World Models predicts action-conditioned descriptions of how the interface state will change following candidate actions (Chae et al., 2025). More generally, world model quality can be evaluated through next-state or state-change accuracy, event precision and recall, time-to-event error, and multi-step rollout fidelity. When multiple future trajectories are plausible, probabilistic measures such as negative log-likelihood, Brier score, and calibration error can assess uncertainty across possible outcomes. As with user-state inference, comparing a fixed policy under predicted and observed world states can help attribute downstream failures specifically to forecasting errors.

5.2.3 Planning Quality

Choosing an appropriate objective and a coherent course of action from an anticipated user state or future situation is the task of the *Agenda* mechanism. This involves deciding what should be pursued, decomposing the objective into intermediate steps, respecting dependencies and constraints, and revising the plan when new information becomes available. Evaluation should distinguish failures in agenda formation, where the agent selects an inappropriate objective, from failures in plan construction, where the objective is appropriate but the resulting plan is invalid or ineffective.

An agent’s objectives should be relevant, necessary, and consistent with the user’s preferences. π Bench distinguishes Proactivity, which measures whether hidden requirements are identified or resolved, from Completeness, which measures how many task requirements are ultimately satisfied (Zhang et al., 2026b). UserBench evaluates whether decisions

reflect elicited user preferences through its normalized choice score, Best Exist Rate, Correct Exist Rate, and Preference Elicited (Qian et al., 2025a). ProactiveMobile similarly reports Success Rate for translating an inferred need into an appropriate function sequence, together with False Trigger Rate when no action is warranted (Kong et al., 2026). More generally, agenda quality can be measured through objective selection accuracy, precision and recall over required objectives, agreement with user-ranked goals, and regret relative to the highest utility feasible objective. These measures help distinguish an agent that plans poorly from one that successfully pursues the wrong goal.

Once an appropriate objective has been selected, plan construction evaluates whether the agent can produce a valid sequence or structure of actions that achieves it. PlanBench tests understanding of action preconditions, effects, and goal reachability, supporting measures such as valid plan rate, goal achievement rate, and the number of violated preconditions or inapplicable actions (Valmeekam et al., 2023). Natural Plan evaluates trip, meeting, and calendar planning through solve rate under increasingly complex natural language constraints (Zheng et al., 2024a). WorFBench focuses on task decomposition and evaluates sequential and graph structured plans through subsequence and subgraph matching (Qiao et al., 2025). The Agent Planning Benchmark reports binary Plan Correctness and six-level Plan Grade, while distinguishing failures caused by inappropriate tools, unavailable tools, and infeasible objectives (Sun et al., 2026a). Across these settings, useful measures include goal reachability, plan validity, subgoal coverage, dependency preservation, tool appropriateness, and recognition that no feasible plan exists.

Planning quality also depends on satisfying constraints while choosing well among multiple feasible solutions. TravelPlanner separately evaluates delivery, commonsense constraints, hard constraints, and overall task completion (Xie et al., 2024a), while COMPASS distinguishes feasibility from preference optimization and exposes cases in which a plan satisfies mandatory requirements but poorly reflects the user’s priorities (Qin et al., 2025b). AgentBoard introduces progress rate to measure how much of an objective has been completed even when the final goal is not reached (Ma et al., 2024a). Evaluation can report feasibility, constraint satisfaction, preference utility, regret, and subgoal progress separately, especially when multiple plans can satisfy the same objective.

Plan revision becomes important when new user information or environmental changes invalidate part of an existing plan. Flex TravelPlanner evaluates planning as constraints are introduced sequentially and tests whether higher priority requirements remain satisfied when conflicts arise (Oh et al., 2025). AdaPlanBench evaluates replanning under progressively revealed user and environment constraints through final Accuracy, Valid Plan Rate, interaction turns, and repeated violations of previously disclosed constraints (Liu et al., 2026i). TripTide evaluates adaptation to unexpected disruptions through Preservation of Intent, Responsiveness, and semantic, spatial, and sequential Adaptability (Karmakar et al., 2025). These settings motivate measures such as recovery success, constraint retention, preservation of unaffected plan components, revision cost, and repeated constraint violations, capturing whether the agent can update an existing plan coherently as conditions evolve.

5.2.4 Intervention Decision and Timing

Intervention Decision and Timing evaluates the *Arbitration* mechanism: whether an agent should intervene, when intervention is appropriate, and whether the proposed action is justified within its granted permissions. Given an anticipated user state and a proposed agenda, the agent may act directly, ask for clarification, request consent, defer to the user, or remain silent. The central evaluation question is therefore not whether the intervention itself is well expressed or correctly executed, but whether the agent chooses an appropriate intervention decision under the available evidence, uncertainty, risk, and timing constraints.

A first dimension is whether intervention is warranted at all. Ask Before Plan evaluates whether an agent recognizes that available information is insufficient before proceeding with planning (Zhang et al., 2024h). ProactiveBench tests whether assistance is offered only in situations where intervention is appropriate (Lu et al., 2025a), while HiL Bench reports Ask F1 to balance necessary clarification against unnecessary questioning (Trinh et al., 2026). KnowU Bench distinguishes among acting, seeking consent, and remaining silent (Chen et al., 2026g), and ProactiveMobile allows multiple decisions to be considered valid in the same context (Kong et al., 2026). These settings motivate intervention precision, recall, and F1, false intervention rate, missed assistance rate, macro F1 for multiple decision types, and set valued accuracy when several choices are acceptable. Clarification and consent decisions can further be evaluated through unnecessary question rate, missed clarification rate, and consent seeking accuracy across actions with different levels of consequence and reversibility.

A second dimension concerns how much authority the agent should assume. Mixed-initiative systems characterize this

choice as a balance between the expected value of autonomous action and the potential costs of error, disruption, or loss of user control (Horvitz, 1999). Adjustable autonomy similarly studies when control should remain with the agent and when it should be transferred to a human, including the coordination and delay costs associated with that transfer (Scerri et al., 2002). Learning-to-defer formalizes a related problem by evaluating whether a system acts on cases it can handle reliably while passing more difficult cases to an expert (Mozannar & Sontag, 2020). Calibrated variants additionally examine whether estimated agent and human correctness probabilities correspond to observed outcomes (Verma & Nalisnick, 2022). Relevant measures include autonomy coverage, deferral accuracy, selective risk, risk-coverage curves, expected decision cost, calibration error, and Brier score. Together, these metrics evaluate whether the agent allocates decision authority appropriately.

Intervention quality also depends on timing. In proactive settings, an intervention can remain useful within a range of acceptable times. JIR-Arena evaluates precision and recall of predicted information needs and scores recommendation timing through a soft comparison of predicted and annotated start and end times (Yang et al., 2025d). ProActor defines broader readiness windows and penalizes interventions that occur either too early or too late (Ding et al., 2026b). ProCIS incorporates timeliness through pDCG and npDCG, rewarding relevant information when it is provided at an earlier useful point (Samarinas & Zamani, 2024), while ProactiveVideoQA integrates response quality over time through PAUC (Wang et al., 2025k). OmniPro evaluates performance around annotated trigger moments and under autonomous online triggering (Zhao et al., 2026b), and EgoPro Bench makes the appropriate intervention time dependent on the user’s profile and inferred intentions (Ran et al., 2026). ProAgentBench further distinguishes intervention at an inappropriate moment from anticipation of inappropriate assistance (Tang et al., 2026b). These settings motivate valid-window accuracy, trigger precision and recall, premature intervention rate, delayed intervention rate, temporal distance from the nearest acceptable boundary, and utility lost because of delay.

The usefulness of a trigger also depends on whether the user can accommodate the interruption. Interruptibility research estimates the expected cost of interrupting a user (Horvitz & Apacible, 2003) and predicts whether the user is currently available for interruption (Hudson et al., 2003). Interruption cost varies across points within the same task (Adamczyk & Bailey, 2004). Interventions outside natural task boundaries can increase completion time, errors, and annoyance (Bailey & Konstan, 2006). Delaying notifications until more suitable breakpoints can reduce disruption (Iqbal & Bailey, 2008). Mobile studies further show that receptivity depends on personal and situational context (Pejovic & Musolesi, 2014), and that willingness to engage with an intervention is distinct from merely tolerating its arrival (Pielot et al., 2017). Useful measures therefore include expected interruption cost, resumption time, performance loss following interruption, subjective disruption, receptivity prediction accuracy, and subsequent engagement.

Finally, aggregate success can conceal important arbitration failures. Toxic proactivity evaluations measure how often agents violate safety constraints while pursuing task goals (Wang et al., 2026l). ProMemAssist relates the frequency of proactive memory support to users’ responses to those interventions (Pu et al., 2025b), while proactive programming studies distinguish productive engagement, passive ignoring, and active disruption (Pu et al., 2025a). These reactions should be interpreted together with delivery quality, since rejection may result from either an inappropriate decision to intervene or a poorly delivered intervention. Arbitration evaluation should therefore report unnecessary intervention, missed assistance, excessive clarification, inappropriate autonomy, premature interruption, and delayed action separately.

5.2.5 Intervention Delivery and Presentation

Intervention Delivery and Presentation evaluates the *Action* mechanism: whether an approved intervention is communicated or executed in a form that is correct, understandable, controllable, and appropriate to its consequences. An intervention may appear as a suggestion, hint, clarification, preview, explanation, editable contribution, confirmation request, or direct action. Arbitration evaluates whether, when, and within what scope an intervention is warranted. Action evaluation assesses the quality of its execution and presentation.

The first requirement is substantive correctness. For informational interventions, the delivered content should accurately address the anticipated need; for operational interventions, the resulting action should produce the intended state change. ProactiveVideoQA evaluates the correctness of information delivered while an event unfolds (Wang et al., 2025k), while ProAgentBench separates the quality of proposed assistance from the decision to initiate that assistance (Tang et al., 2026b). ProactiveMobile evaluates whether predicted device-function sequences are functionally equivalent to a valid reference (Kong et al., 2026). Related agent benchmarks measure functional task correctness (Zhou et al., 2024b) and progress toward the intended goal when full completion is not achieved (Ma et al., 2024a). Relevant measures include

content accuracy, semantic relevance, execution success, state change accuracy, subgoal completion, and implementation fidelity between the approved plan and the action that is actually performed.

Correctness, however, does not determine whether the intervention is delivered with an appropriate degree of user involvement. This constitutes a distinct evaluation target within Action: once an intervention has been selected, the agent must still determine whether to present it as information, a suggestion, an editable preview, a confirmation request, or an executed action. Mixed initiative interaction suggests that this choice should account for uncertainty about user intent and the potential cost of an incorrect action, with dialogue used to resolve uncertainty when autonomous execution is not justified (Horvitz, 1999). Research on levels of automation similarly treats automation reliability and the consequences of erroneous decisions or actions as criteria for determining how much control should be delegated (Parasuraman et al., 2000). Human AI interaction guidelines further emphasize efficient dismissal and correction, cautious behavior under uncertainty, and mechanisms that preserve user control (Amershi et al., 2019). Accordingly, delivery mode appropriateness can be assessed through human ratings, pairwise preferences, or agreement with expert authored policies. User control can be evaluated through the availability of inspection, editing, rejection, pausing, cancellation, and reversal. Correction effort provides a complementary measure by quantifying the time, number of interactions, or amount of modification required to turn the proposed intervention into an acceptable outcome. These measures expose usability costs that final task success alone may conceal.

A further dimension is whether the intervention makes the agent’s intention and actions understandable to the user. Users should be able to determine what the agent is proposing, what it has already changed, and what is likely to happen next. Objective measures can include comprehension accuracy, interpretation time, prediction of the agent’s next action, and consistency between explanations and executed behavior. Behavioral traces provide additional evidence. Proactive programming studies distinguish productive engagement, passive ignoring, and active disruption (Pu et al., 2025a); interface studies record behaviors such as previewing, expansion, copying, acceptance, and rejection (Chen et al., 2025e); and proactive visual analytics systems examine whether suggestions are executed or incorporated into subsequent analysis (Zhao et al., 2025b). Acceptance should be interpreted alongside intervention correctness and user outcomes. Explanations can increase reliance without improving human and agent performance (Bansal et al., 2021), and evaluations should also measure learning and task benefit when users appear to use assistance (Aleven et al., 2006; Baker et al., 2004). More informative evaluation can distinguish immediate acceptance, acceptance after modification, rejection after inspection, later reversal, and successful correction.

Delivery quality must also be considered relative to the attention and coordination it demands. Reading time, interaction steps, interface transitions, cognitive workload, and recovery time after an intervention can be evaluated together with the benefit that the intervention provides. Human robot interaction offers related measures of coordination. Human robot fluency has been evaluated through concurrent activity, human idle time, and functional delay (Hoffman & Breazeal, 2007), while motion legibility measures how readily and how early people can infer the robot’s intended goal (Dragan et al., 2013). Gaze and conversational turn taking provide additional evidence about coordination and interpretation (Baraglia et al., 2017; Peng et al., 2019), although these measures often require task specific behavioral coding (Fink et al., 2014). Digital agents face an analogous requirement: users should be able to understand the direction of an intervention, observe its effects, and retain the ability to modify or stop it.

5.2.6 Adaptation and Generalization

Agents can improve through experience, remain effective as operating conditions change, and transfer what they have learned beyond the interactions from which that knowledge was acquired. An agent may need to adjust to changing user preferences, new environmental dynamics, unfamiliar tasks, revised tool interfaces, or changes in the information available to it. These experiences may affect any earlier mechanism, including what context is retained, how future states are anticipated, which objectives are selected, when intervention is appropriate, and how actions are carried out. Evaluation should therefore characterize the trajectory of improvement and recovery, including performance after adaptation.

User adaptation concerns whether repeated interaction enables the agent to better serve a particular user. ProPerSim evaluates whether an assistant can learn from repeated user feedback and improve the relevance and timing of personalized recommendations over time (Kim et al., 2026b). Training Proactive and Personalized LLM Agents similarly studies whether agents can adapt their interaction behavior to previously unseen user preferences (Sun et al., 2025b), while PersonaLens evaluates preference alignment together with response quality and task success across diverse user profiles

and interaction histories (Zhao et al., 2025d). PAHF more directly evaluates continual personalization by measuring Success Rate, Feedback Frequency, and Average Cumulative Personalization Error during both initial preference learning and adaptation after preferences change (Liang et al., 2026). These settings suggest that personalization gain should be reported together with learning speed, sample efficiency, cumulative error or regret, and the amount of feedback required to achieve the improvement. This distinction is important because agents that eventually reach similar performance may impose very different learning burdens on their users.

Adaptation over long interaction histories further requires determining which prior information remains valid as circumstances change. Effective adaptation involves acquiring new information and updating, revising, or disregarding information that has become obsolete. LongMemEval evaluates information extraction, reasoning across sessions, temporal reasoning, knowledge updating, and abstention in sustained interactions (Wu et al., 2025b). PERMA examines whether agents can reconstruct evolving preferences from temporally related events distributed across sessions and domains (Liu et al., 2026m). Memora explicitly evaluates failures caused by continued reliance on obsolete or invalidated memories and introduces a forgetting-aware measure that penalizes such behavior (Uddin et al., 2026). OP Bench evaluates whether retained user information is used when relevant and whether personalization becomes irrelevant, repetitive, or sycophantic (Hu et al., 2026e). VitaBench 2.0 further organizes interactions as temporal user trajectories that require agents to continuously extract, use, and update preferences (Chen et al., 2026h). Relevant measures include update accuracy, obsolete-information reuse, inappropriate personalization, performance immediately before and after a change, recovery time, and the amount of interaction required to establish the new state.

Agents must also adapt when the environment changes independently of the user. Such changes may affect available tools, interface structure, observation formats, state transition dynamics, resources, or the validity of previously learned strategies. ProEvolve constructs controlled sequences of environment changes by modifying tools, schemas, and data access, allowing agent performance to be evaluated as the surrounding system evolves (Li et al., 2026a). EvoArena similarly represents environmental change as progressive updates across terminal, software, and social settings and evaluates whether agents maintain useful knowledge across successive changes (Xu et al., 2026b). Test Time Adaptation for LLM Agents examines performance before and after agents adapt to previously unseen environment formats and dynamics using information gathered through deployment interaction (Chen et al., 2026a). Gaia2 introduces environments that evolve asynchronously, including events that occur independently of the agent’s actions, and therefore tests whether an agent can continue reasoning and acting as external conditions change (Froger et al., 2026). These settings motivate measures such as the immediate performance drop after a change, adaptation gain, recovery time, cumulative loss during recovery, interaction efficiency, retention of knowledge that remains valid, and performance across consecutive environment changes.

Experience may also improve capabilities that can be reused across subsequent tasks. Evaluating such adaptation requires testing whether gains obtained from earlier experience persist when the task, goal, or environment changes. AgentGym evaluates agents across diverse tasks and environments and studies whether experience acquired through interaction improves performance beyond the settings in which it was collected (Xi et al., 2024). Relevant measures include improvement on later related tasks, transfer efficiency, retention of previously acquired capabilities, forgetting after learning new tasks, and performance gained per unit of additional experience. These measures assess the accumulation of reusable capabilities across tasks.

Evaluation can test whether knowledge acquired from previous experience remains useful when the agent encounters users, tasks, applications, tools, or environments that were not directly represented during learning. RealPref evaluates whether preferences inferred from long interaction histories can guide responses in unseen situations (Guo et al., 2026b). ProCodeBench exposes a complementary gap between simulated and real developer behavior by comparing proactive intent prediction under synthetic and real IDE traces (Li et al., 2026d). AndroidWorld varies tasks and operating conditions within a dynamic mobile environment and shows that performance can change across task variations and platforms (Rawles et al., 2024). AgentGym likewise provides evaluation across multiple environments and tasks (Xi et al., 2024). Generalization can therefore be measured through performance on held-out users, tasks, tools, applications, and environments, together with the source-to-target performance gap, worst-group performance, zero-shot transfer, and the amount of additional interaction required when some adaptation is permitted.

5.3 User Experience Metrics

The third level evaluates how users experience proactive assistance. Evaluation should measure whether an intervention increases cognitive burden, disrupts ongoing activity, reduces user control, or encourages inappropriate dependence

on the agent. In proactive systems, initiative originates partly or entirely from the agent without an explicit user request. User experience evaluation therefore complements task and mechanism evaluation by measuring the human consequences of agent initiative.

User experience can be assessed through four dimensions (Table 2): User Workload, Usability and Experience, Level of Control, and Trust and Reliance. Workload concerns the effort demanded from the user, usability and experience concern the perceived quality and fit of the interaction, control concerns the user’s effective authority over agent behavior, and trust and reliance concern whether users depend on the agent to an appropriate degree.

5.3.1 User Workload

Proactive assistance can change the user’s cognitive, temporal, physical, and attentional burden. Initiating useful actions before the user asks can reduce effort. Attracting attention, introducing additional decisions, requiring supervision, or interrupting ongoing work can create new demands. Task-completion time, interaction steps, and workload should therefore be measured separately.

Validated workload instruments provide broad measures of these costs. NASA TLX assesses Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration (Hart & Staveland, 1988). The Workload Profile provides a more diagnostic account of the cognitive resources required by a task (Tsang & Velazquez, 1996), while the Subjective Workload Assessment Technique emphasizes time load, mental effort, and psychological stress (Reid & Nygren, 1988). These measures can reveal cases in which proactive assistance removes direct interaction effort while increasing monitoring, interpretation, or decision burden.

Behavioral measures provide complementary evidence about when and how such burden occurs. Interpretation time following an intervention (Pu et al., 2025a), task-switching delay, additional errors, and recovery time can expose disruption that may not be fully reflected in subjective ratings. Interruptions introduced outside suitable task boundaries can increase completion time, errors, and annoyance (Bailey & Konstan, 2006), and interrupted work has also been associated with greater stress, frustration, and perceived effort (Mark et al., 2008). Delaying information until detectable breakpoints can reduce disruption (Iqbal & Bailey, 2008). Because the same behavioral delay may arise from workload, unclear assistance, or low trust, workload is best evaluated through a combination of subjective ratings, behavioral disruption, and recovery measures.

Workload evaluation should report effort saved together with demands introduced by an intervention. Relevant outcomes include perceived mental and temporal demand, attention shifts, monitoring effort, decision burden, recovery time, and additional actions required to manage the intervention. Apparent efficiency gains can coexist with greater cognitive or supervisory burden in proactive systems.

5.3.2 Usability and Experience

Users judge proactive assistance by whether it is easy to understand and use, fits naturally within their activity, and provides value without creating excessive friction or discomfort. Workload measures the effort required by an interaction. Usability and experience measures capture how users judge its overall quality, usefulness, naturalness, and appropriateness.

Standardized instruments provide general measures of interaction quality. The System Usability Scale provides a global assessment of perceived usability (Brooke, 1995), while the User Experience Questionnaire captures broader pragmatic and experiential qualities (Laugwitz et al., 2008). PARADISE relates task success and dialogue costs to user satisfaction in interactive systems (Walker et al., 1997). Human-AI interaction guidelines introduce additional criteria that are particularly relevant to proactive systems, including whether the system communicates its capabilities, allows users to correct or dismiss its behavior, and adapts appropriately through repeated interaction (Amershi et al., 2019).

More specific measures are often needed because the desirable experience depends on the role and domain of the agent. Proactive search and visual captioning systems evaluate user preference and satisfaction (Andolina et al., 2018; Liu et al., 2023c), emotional support agents assess perceived support and interpersonal appropriateness (Liu et al., 2024e), and proactive coding studies examine perceived collaboration and interference (Pu et al., 2025a). The Working Alliance Inventory can characterize perceived collaboration and relational quality over longer interactions (Horvath & Greenberg, 1989). In embodied systems, the Godspeed scales assess anthropomorphism, animacy, likeability, perceived intelligence,

and safety (Bartneck et al., 2009). These measures illustrate that user experience is multidimensional and should be aligned with the type of relationship the system is intended to support.

Proactivity also creates tradeoffs among usefulness, intrusiveness, and privacy. Greater robot initiative can increase perceived usefulness while also making the system feel more intrusive (Villa et al., 2025), and proactive conversational cues may be experienced as useful decision support and may also be experienced as unwanted interference (Ma & Ziegler, 2024). Continuous sensing and contextual inference introduce an additional privacy dimension. Information collection can be perceived as inappropriate when it violates contextual expectations even when the resulting intervention is accurate or beneficial (Nissenbaum, 2004). These dimensions should therefore be reported separately.

Global usability scores can provide a common reference point. Domain-specific measures are often needed to explain why an intervention is welcomed, ignored, or rejected.

5.3.3 Level of Control

Users should retain appropriate authority over the agent's goals, decisions, and actions. Evaluation should assess override availability and whether users can understand what the agent is doing, influence its behavior when necessary, and stop, modify, or reverse its actions in practice.

Research on proactive interaction shows that perceived control can change as agent initiative increases. Proactive programming studies measure perceived control and whether stronger initiative shifts the perceived role of the system from a tool toward a collaborative partner (Pu et al., 2025a). Robot studies similarly examine preferred autonomy and the appropriateness of different levels of independent action (Villa et al., 2025). Interaction coding can reveal how initiative shifts between participants through behaviors such as repetition, summarization, interruption, or transfer of responsibility (Walker & Whittaker, 1990). Levels of automation frameworks provide a broader account of how authority can be distributed across information acquisition, analysis, decision selection, and execution (Parasuraman et al., 2000).

A user may formally retain the right to reject an agent's action while exercising little effective control if the agent controls the relevant information, establishes influential defaults, acts before the user can respond, or makes reversal costly. Economics distinguishes formal authority, meaning the recognized right to make a decision, from real authority, meaning practical influence over what is actually decided (Aghion & Tirole, 1997). Control evaluation should therefore examine the availability of corrective mechanisms and whether they can be understood and used effectively in realistic interaction.

System behavior should remain responsive to relevant human reasons, and consequential outcomes should remain traceable to responsible human actors, as proposed in work on meaningful human control (Santoni de Sio & van den Hoven, 2018). Human-AI design guidelines similarly emphasize efficient correction, dismissal, and adaptation following user feedback (Amershi et al., 2019). Even limited opportunities to modify an algorithmic recommendation can increase satisfaction and willingness to continue using the system (Dietvorst et al., 2018). Relevant measures include perceived agency, override frequency and success, modification success, reversal cost, use of autonomy settings, understanding of control boundaries, and the gap between nominal authority and effective influence.

Evaluation should examine how control is exercised across the complete interaction. A proactive agent provides meaningful control when users can anticipate its scope of action, intervene before undesirable consequences occur, and efficiently correct its behavior when their goals or preferences are not being followed.

5.3.4 Trust and Reliance

Users' dependence on the agent should match its actual capabilities and limitations. Evaluation should assess the calibration of trust, acceptance, and usage to those capabilities. A well-calibrated user should rely on the agent when it is likely to be correct and maintain sufficient skepticism when its recommendations or actions are uncertain.

Trust questionnaires include items on reliability, dependability, and confidence (Jian et al., 2000). Trust should also remain sensitive to actual system capability, vary appropriately across reliable and unreliable conditions, and remain specific to the task or component being evaluated (Lee & See, 2004). Broader models distinguish dispositional, situational, and learned sources of trust (Hoff & Bashir, 2015). Appropriate reliance provides a more behavioral perspective by asking whether users accept correct recommendations while rejecting incorrect ones (Schemmer et al.,

2022). Together, these approaches distinguish subjective confidence in the agent from whether that confidence produces appropriate behavior.

Behavioral reliance should be interpreted relative to recommendation correctness. Accepting accurate recommendations and rejecting incorrect ones can indicate appropriate reliance. Acceptance of incorrect recommendations can indicate excessive reliance, and rejection of correct recommendations can indicate unnecessary distrust. Explanations can increase acceptance without improving complementary performance between the user and the AI system (Bansal et al., 2021). Cognitive forcing interfaces can reduce acceptance of incorrect AI advice. They may also increase user effort and reduce subjective preference (Bućinca et al., 2021). Evaluation should consequently distinguish correct acceptance, correct rejection, incorrect acceptance, unnecessary rejection, verification behavior, error detection, and the effort required to maintain appropriate skepticism.

A study of knowledge workers found that greater confidence in generative AI was associated with lower self-reported critical thinking. Greater confidence in one’s own expertise was associated with more critical engagement (Lee et al., 2025c). The study characterizes critical thinking in AI-supported work through activities such as information verification, response integration, and task stewardship. For proactive agents, this suggests evaluating whether users inspect evidence, challenge questionable suggestions, consider alternatives, and retain sufficient understanding of the task to recognize agent errors. These outcomes complement conventional trust measures by examining whether reliance preserves active human judgment.

Trust develops through experience and may change as an agent succeeds, fails, becomes more autonomous, or adapts its behavior. Introducing highly autonomous behavior before sufficient trust has developed can reduce user confidence (Kraus et al., 2020). Measuring trust repeatedly throughout an interaction can reveal its formation, degradation, and recovery over time. Trust trajectories and summary measures such as the area under the trust curve can therefore provide a longitudinal account of how users respond to changes in system reliability (Desai et al., 2013). Evaluation should assess whether trust and reliance remain aligned with what the agent can actually do under the conditions in which it is being used.

6 Applications

This section surveys applications of proactive AI agents across 13 domains, from Coding and GUI Agents to Healthcare and Agriculture. Figure 14 illustrates the broader transition these applications point toward: agents that can discover and advance work under standing objectives, with people defining goals and reviewing outcomes. Each domain has its own subsection, which summarizes representative systems in a per-vertical table, identifies the tasks on which proactiveness is exercised, examines how the agent’s behavior is realized through Awareness, Anticipation, Agenda, Arbitration, Action, and Adaptation, and closes with an Evaluation, Challenges, and Open Directions discussion. A concluding Cross-Industry Trends subsection places the 13 domains on a shared Initiative-and-Authority maturity map (Table 16) and reads three recurring patterns off it.

6.1 Coding Agents

Modern coding agents already exhibit substantial initiative (Wang et al., 2025h; Anthropic, n.d.b; OpenAI, 2025a; GitHub, 2025; Lin, 2026a). They can complete repository-level tasks asynchronously, respond to workflow events, and participate in validation and review with limited step-by-step supervision. Their autonomy is also becoming more context-sensitive. Rather than following a single permission policy, deployed systems increasingly adjust how far they proceed according to the expected consequence of an action and the safeguards available around it (Gomes & McPeak, 2026).

Their initiative nevertheless remains bounded in an important way: the work itself is usually selected elsewhere. An assigned issue, workflow trigger, or project board determines what deserves attention, after which the agent plans and executes within that scope. Studies of Claude Code usage suggest a similar division of labor, with users making most planning decisions and the agent making most execution decisions (Hitzig et al., 2026). Reports on agent-first engineering likewise retain human control over prioritization and acceptance, even when agents perform much of the implementation and review (Lopopolo, 2026; Kotliarskyi et al., 2026).

Through the mechanism taxonomy of Section 4, current systems are already strong in *Awareness* and *Action*, and increasingly capable of *Arbitration*. The larger gap lies in project-level *Anticipation*, *Agenda*, and *Adaptation*. Automatically

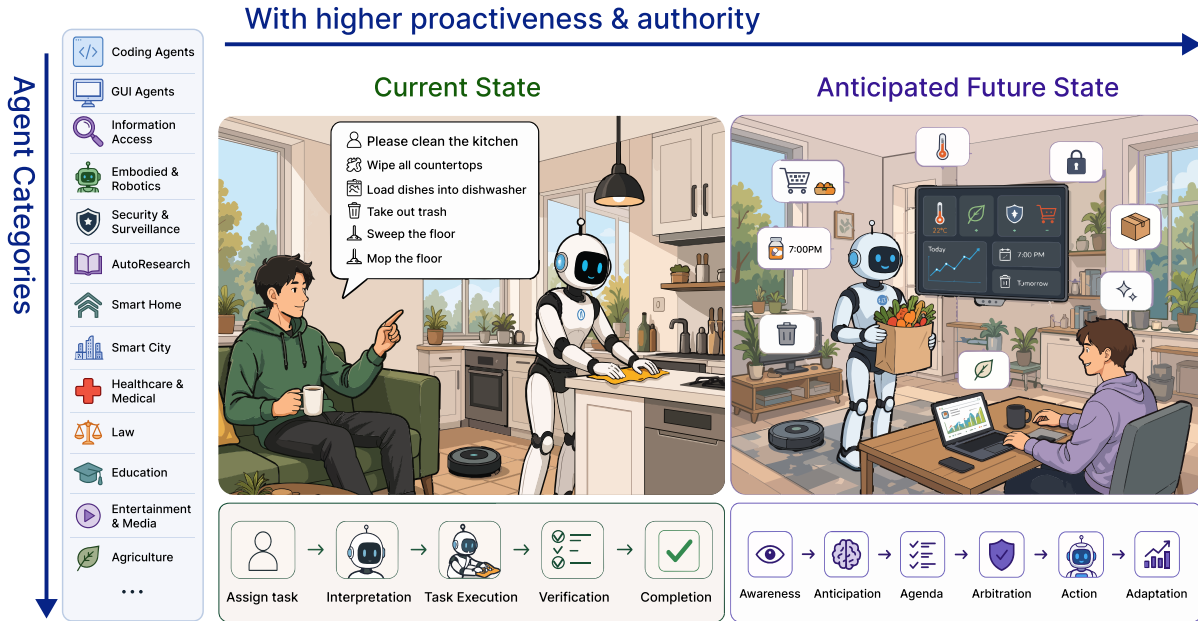


Figure 14 From executing assigned tasks to sustaining delegated objectives. Existing applications increasingly allow agents to execute complex tasks, but the work to be done is often still specified by humans. Broader proactive assistance requires agents to discover emerging needs, maintain and revise agendas, and initiate interventions that sustain delegated objectives. This changes the human role from prescribing each action to setting goals and reviewing outcomes.

Table 3 Representative coding-agent systems and mechanisms supporting proactiveness. The dimensions summarize documented mechanisms rather than implying complete proactive capability. Agenda distinguishes immediate or externally assigned task goals from persistent or self-initiated engineering commitments. Platform denotes the interaction environment.

System	Awareness	Anticipation	Agenda	Arbitration	Action	Adaptation	Platform
Cursor Tab (Kravtsov, 2025 ; Jackson et al., 2025)	Buffer; cursor; recent edits	Next-edit prediction	Immediate edit intent	Learned suggestion policy	Line / cursor completion (<i>suggestion</i>)	RL from acceptance	IDE
Claude Code (Anthropic, n.d.b)	Repo; files; tests; tools	Task-scoped planning	User task; agent subgoals	Permissioning; sandbox; review gates	Multi-file edits; commands; PRs (<i>gated action</i>)	Project memory; reusable skills	Repo
OpenAI Codex (OpenAI, 2025a)	Preloaded repo; cloud environment	Task-scoped planning	Assigned task queue	Sandbox isolation; review	Features; fixes; PRs (<i>gated action</i>)	Test-driven revision	Cloud repo
GitHub Copilot (GitHub, n.d.a, 2026c)	Issues; repo; Actions	Issue-scoped planning	Assigned issues	Branch / PR review gate	Background PRs (<i>gated action</i>)	Repository memory; user preferences	Repo

repairing a triggered CI failure, for example, demonstrates autonomous execution. Recognizing that recurring failures reveal a broader reliability problem, deciding that the problem deserves attention, and sustaining the resulting work across later changes call for a stronger form of proactivity.

At the project level, a proactive coding agent operates under standing project objectives and bounded authority, identifies work that has not yet been explicitly assigned, decides which work deserves attention, selects an appropriate form of intervention, and maintains unfinished commitments as the project evolves. These standing objectives may be less specific than an assigned issue. Specifically, a project may ask the agent to preserve reliability, reduce security risk, or keep maintenance costs under control without specifying which component to inspect or what change to make. However, such objectives express project priorities rather than a single execution path. The agent therefore needs to translate them into concrete engineering work while accounting for the conventions and trade-offs of the project. The following sections organize this progression around proactive work discovery, engineering agenda formation and prioritization,

risk-aware intervention and coordination, and long-running project stewardship. Table 3 compares selected coding systems across the six mechanisms, including the scope of their agendas and the controls on code changes.

6.1.1 Proactive Discovery of Unassigned Engineering Work

Task-scoped initiative can extend into project-level discovery when the agent identifies work beyond the current assignment. Current agents already exhibit substantial *task-scoped anticipation*: given a high-level issue, they can recover repository context, infer the affected scope, and determine how the requested outcome should be validated (Zhang et al., 2023; Wu et al., 2024a; Yang et al., 2024a; Zhang et al., 2024j; Wang et al., 2025h). An interface change may require updates to downstream call sites and regression tests even when neither is stated explicitly. Such work can span much of a repository while remaining subordinate to an objective selected by someone else.

Moving beyond this boundary calls for project-level understanding: relating local evidence to the architecture, history, and standing objectives of the repository, in the way experienced engineers recognize when repeated local fixes reveal a deeper reliability problem or when a temporary workaround is becoming structural debt. Observational evidence from Claude Code suggests that such high-level judgment remains strongly associated with human expertise, with users making most planning decisions and domain experts providing more effective direction per interaction (Hitzig et al., 2026). Industrial harnesses and academic work are beginning to externalize this expertise through repository-local documentation, reusable skills, and distilled procedural knowledge (Lopopolo, 2026; Zhang et al., 2025b; Li et al., 2026j; Ning et al., 2026d), though SWE-Skills-Bench shows that generic or mismatched skills often provide little benefit and can interfere with the current project context (Han et al., 2026a). The research challenge is therefore to select and ground the engineering experience relevant to a newly observed concern.

A discovery goes beyond the assigned task when the concern can be accepted, deferred, or rejected as an independent project decision, separate from completing the current task. In comparison, work necessary to complete the current objective remains within the assigned task, even when it was omitted from the original request. An architectural weakness discovered during a bug fix meets this bar, since the project can choose separately whether to address it. TEBench operationalizes this setting for testing, requiring agents to identify stale or missing tests even when no immediate failure exposes the need (Shang et al., 2026). Code security offers a second setting where such discovery is already measurable: Big Sleep, Revelio, and RFCAudit search a well-scoped coding domain for previously unstated concerns and ground them through executable evidence or external specifications (Big Sleep team, 2024; Hou et al., 2026; Zheng et al., 2025c). Both settings suggest that credible proactive discovery today rests on a well-defined domain scope and a strong verification loop.

Recent products are beginning to extend discovery beyond repository-local signals. OpenAI describes Dots (Table 4) as monitoring recurring customer feedback, scoping improvements and bug fixes, and preparing tested pull requests for review (OpenAI, 2026c). This workflow illustrates a move from executing assigned issues toward identifying candidate engineering work from ongoing product context. However, the product description alone does not establish how reliably such candidates are selected or whether they deserve priority over competing project needs.

Proactive discovery can also be decoupled from execution. An agent whose task is to surface candidate work does not need authority to act on it: its output can be a scoped, evidence-backed engineering claim that the project accepts, defers, or rejects at a later stage. GitHub’s agentic autofix follows this pattern by validating a proposed repair in a sandbox and returning it as a draft pull request for review (GitHub, 2026a). Structuring proactive discovery this way keeps each candidate reviewable in isolation; whether it becomes a project commitment is the next question, handled by agenda formation.

6.1.2 Engineering Agenda Formation and Prioritization

The proactive move at this stage is commitment formation. Execution-oriented systems such as the Codex app and Symphony primarily organize work around externally supplied tasks (OpenAI, 2026a; Kotliarskyi et al., 2026). The Codex app, for example, isolates concurrent agents through worktrees, while Symphony dispatches project-board items into persistent agent workspaces. Both systems increase throughput, but neither removes the prior need to decide which concerns deserve a place on the agenda.

A proactive agenda therefore involves three additional decisions beyond execution: which candidate to commit to, how it depends on other work, and how to sequence it under standing project priorities. Some tasks can proceed independently,

whereas others rely on an interface change or a validation result that has not yet arrived. Prioritization must also account for delayed feedback and limited execution, testing, and review capacity. A security hardening task, a maintainability refactor, and a documentation improvement may all be useful, but the agenda must determine which best serves the project’s current needs and which should wait.

Recent multi-agent coding systems make dependency-aware planning the central mechanism that lets a proactive agenda run in parallel. CAID plans independent subtasks for asynchronous execution and integrates them through branch-and-merge with executable verification (Geng & Neubig, 2026). Anthropic’s parallel compiler project similarly coordinated sixteen agents across nearly two thousand sessions through explicit decomposition and tests (Carlini, 2026). These systems do not yet decide which concerns should enter the agenda, but they show that useful parallelism depends on isolating dependencies and validating integration rather than merely assigning more workers.

Proactive scheduling must handle bottlenecks beyond model inference: compilation, integration tests, merge resolution, and human review can dominate the critical path. Cursor’s self-driving codebase experiment reports that static dependency graphs became too rigid under these constraints, the system stalled on its slowest worker, and large-scale compilation created shared-resource contention (Lin, 2026b). The agenda must therefore use waiting time productively while avoiding downstream work whose assumptions are not yet stable. This makes scheduling a continual resource-allocation problem rather than a one-time ordering of tasks.

A proactive planning policy can itself become a learned artifact when execution traces from prior project commitments are treated as training signal, though distilled planning experience must remain revisable as project assumptions change. Cursor reached its eventual orchestration design by analyzing execution logs, identifying recurring coordination failures, and repeatedly revising the harness rather than trusting its initial organizational analogy (Lin, 2026b). Trajectory-based methods such as CODESKILL offer a path for distilling successful decomposition and validation procedures into reusable planning knowledge (Li et al., 2026j). Yet the limited and sometimes negative gains observed in SWE-Skills-Bench caution against treating every historical strategy as a permanent rule (Han et al., 2026a). Planning experience is therefore most useful when retained together with the assumptions under which it succeeded.

A proactive agenda is therefore a dependency-aware and resource-bounded plan whose structure changes as execution produces new evidence. It records which candidate concerns have become project commitments, how those commitments depend on one another, and how they compete for project resources. In doing so, it indicates what may proceed now, what can run in parallel, what remains provisional, and which plausible improvements are better deferred. The result is a revisable execution structure, not a growing backlog.

6.1.3 Risk-Aware Intervention and Coordination

Deployed coding products have moved from a single global autonomy setting toward action-specific authority, exposing the authority decision as a first-class control surface for proactive intervention (Hughes, 2026; OpenAI, 2026b). This separation is particularly important for unassigned work: choosing a task does not authorize every action needed to implement it. Claude Code ranges from plan mode and manual permission prompts to sandboxed or unrestricted execution, while its auto mode uses a classifier to distinguish routine project-local actions from actions that exceed the user’s authorization. Codex similarly separates the sandbox, which defines the technical boundary of execution, from the approval policy, which determines when an action stops for review.

For proactive interventions executed in the background, automated review can enforce the authority decision after execution, but the runtime cost of this safeguard grows with the depth of testing and review it invokes (Gomes & McPeak, 2026). Sandboxed execution allows an agent to act provisionally while withholding acceptance until a proposed change has been checked against authorization, safety, and quality criteria. The system may still execute the change, run tests, and invoke several reviewers before determining that the intervention was unsuitable. A learned value model could move part of this judgment earlier by estimating, from the proposed plan and repository context, whether an intervention is likely to remain within project policy and user intent. Such a model would complement sandboxes, explicit rules, and final review; its role would be to prevent some unsuitable plans from reaching the more expensive execution stage.

A second axis of the proactive intervention decision is request ambiguity: current agents systematically guess when instructions are underspecified, so a proactive agent should weigh interruption cost against downstream error cost before acting (Ji et al., 2026; Fang et al., 2026e). UnderSpecBench finds that underspecified DevOps instructions cause coding agents to guess rather than consistently clarify or defer, with more than half of evaluated runs violating at least

one action boundary. ClarifyCodeBench likewise shows that strong code-generation ability does not imply strong requirement-clarification ability. Reversible exploration may proceed under uncertainty, whereas a high-blast-radius action with an unclear target may require a plan, clarification, or explicit approval.

When a proactive agent intervenes while other developers and agents are also acting on the codebase, session provenance becomes the shared artifact that carries the intent behind each intervention alongside its final diff ([GitHub, 2026b,d](#)). Current interfaces increasingly expose agent sessions in this way. GitHub’s Agents tab places session histories beside issues and pull requests, while agent-authored commits can link permanently to the session log that produced them. This provenance allows reviewers to recover the evidence and assumptions behind a change and gives another agent or developer a basis for continuing work initiated elsewhere.

Because a proactive agent may intervene in parallel with other agents and developers, the coordination problem shifts from message-passing to whether its actions remain attributable and recoverable; recent systems address this through prompt-level rollback and workspace isolation ([GitHub, n.d.b](#); [Khatua et al., 2026](#); [Geng & Neubig, 2026](#)). GitHub Copilot CLI records prompt-level workspace snapshots and can rewind both the conversation and modified files when an intervention goes wrong. CooperBench finds that two coding agents perform substantially worse when coordination is required, whereas CAID obtains better results by isolating workspaces and centralizing integration. The deeper problem is whether concurrent actions remain compatible with shared state when their assumptions diverge.

Intervention and coordination can therefore be governed by the expected cost of correction: subject to authorization as a hard boundary, a proactive agent may proceed as far as its understanding and rollback guarantees make economical, and clarify or escalate beyond that point. This policy seeks calibrated and correctable intervention rather than the maximum amount of autonomous action. Within its permissions, the agent may choose among exploration, proposal, execution, clarification, escalation, and abstention according to its uncertainty, the action’s blast radius and reversibility, and the likely cost of being wrong.

6.1.4 Long-Running Project Stewardship

Proactive coding extends into stewardship when the agent holds commitments that span sessions, releases, or handoffs and must preserve the meaning of unfinished work across those gaps. A background agent may spend hours completing a fixed issue while its goal and success condition remain externally supplied. Stewardship instead concerns commitments such as a migration that pauses for an upstream dependency and resumes after the repository has changed. The challenge is not merely to keep a process alive, but to preserve why the work remains worth pursuing.

For a proactive agent to carry a commitment across fresh sessions, project state must persist independently of any single model context; current long-running harnesses achieve this by externalizing plans, decisions, and open commitments into versioned repository artifacts ([Lopopolo, 2026](#); [Young, 2025](#); [Rajasekaran, 2026](#); [Ning et al., 2026c](#)). OpenAI’s harness-engineering workflow makes the repository itself the system of record, while Anthropic’s long-running harness asks each fresh session to leave structured progress artifacts for its successor. A later planner–generator–evaluator design applies the same principle to multi-hour application development. These systems provide useful building blocks for continuity without assuming that one context can contain the project indefinitely.

Because a proactive commitment must also survive within-session context compression, the compression policy has to retain the evidence and open questions needed for future decisions, whether through explicit guidance or a learned policy ([Rajasekaran et al., 2025](#); [Gao et al., 2026f](#)). Anthropic’s context-engineering guidance emphasizes retaining high-signal architectural decisions and unresolved problems while discarding redundant interaction history. SWE-MeM goes further by training coding agents to decide when, what, and how to compress their trajectories rather than relying on a fixed summarization rule. The objective is not token reduction alone: a useful representation must preserve the current commitment, the evidence behind it, rejected hypotheses, unresolved questions, and the conditions under which future work may resume.

Handoff between agents is the concrete test of whether stewardship state supports the next proactive decision, and Handoff Debt shows that successors resume work substantially more efficiently when they receive prospective contextual notes than when they inherit repository state alone ([KC & Budathoki, 2026](#)). Code records what changed, but may not reveal why a path was chosen, which hypotheses were rejected, or why the remaining work is blocked. State management should therefore prepare the next agent to decide and act, rather than merely summarize what the previous agent did.

A proactive commitment can be invalidated as the repository continues to evolve during its execution horizon. Long-horizon benchmarks show that this failure mode already appears when the coding tasks themselves are externally assigned (Thai et al., 2025; Chen et al., 2026c). SWE-EVO and SWE-CI expose errors that compound across release-scale changes and long sequences of repository evolution, yet both still specify the work to be performed. Proactive stewardship extends the problem by requiring the agent to resume, revise, reprioritize, or abandon commitments that it selected for itself. It is, therefore, best understood as continuity of engineering intent under changing state, extending beyond persistence of the original plan.

6.1.5 Evaluation, Challenges, and Open Directions

Existing coding benchmarks measure the execution of assigned work at increasing scale and duration, but rarely require the agent to decide what deserves attention on its own (Jimenez et al., 2024; Thai et al., 2025; Chen et al., 2026c; Shang et al., 2026). SWE-bench measures issue resolution, while SWE-EVO and SWE-CI extend assigned-task execution toward coordinated release changes and continued repository evolution. TEBench moves closer to proactive discovery by including stale and missing tests that are not exposed through an immediate failure. These benchmarks broaden the work agents perform, but they still usually specify what deserves attention.

Evaluating proactive discovery requires examining pre-implementation decisions, since outcome-only benchmarks reward completion regardless of whether the action was warranted (Gloaguen et al., 2026). FixedBench shows that agents often modify code even when the reported issue has already been resolved, exposing an action bias that task-completion benchmarks do not penalize. Evaluation should therefore ask whether a concern is grounded, worth attention, and better acted on than deferred or ignored. Passing tests can show that a patch works, but not that the patch was worth making. Standing objectives such as improving maintainability or reducing security risk may admit several valid interventions with different costs and priorities, making developer preferences over which concerns to pursue and which plans to accept an important source of supervision.

A proactive intervention benchmark should measure whether the coding agent’s authority is calibrated to its uncertainty, since ambiguous requests currently draw action from agents that should clarify or defer (Ji et al., 2026; Fang et al., 2026e). UnderSpecBench and ClarifyCodeBench provide useful components by testing whether agents act, clarify, or defer under ambiguity. Runtime review can catch unsafe or unauthorized actions after background execution, but may require substantial testing and rubric-based checking. A learned value model could complement these safeguards by identifying some unsuitable plans before execution, reducing unnecessary trials without replacing sandboxes, explicit permissions, or final review.

Proactive coordination and stewardship require project-trajectory evaluation, since coding competence does not necessarily translate into successful collaboration or efficient handoff (Khatua et al., 2026; KC & Budathoki, 2026). CooperBench shows that coding competence does not automatically produce successful collaboration, while Handoff Debt demonstrates that repository state alone may be insufficient for efficient takeover. Training this behavior also introduces a data bottleneck. Issue histories, patches, CI logs, reviews, and agent traces record how assigned work was completed, but rarely indicate which unassigned concerns were worth pursuing or how commitments should change as the repository evolves. Synthetic trajectories may provide supervision at scale, but they would need to preserve realistic repository changes, competing concerns, delayed feedback, and developer judgments rather than merely generating longer coding traces.

An increasingly proactive coding agent is expected to work under standing project objectives in an evolving repository, requiring it to discover, schedule, execute, and preserve work as the project changes, which requires the corresponding evaluation protocols to provide useful system development signals. Rather than presenting one preselected issue, the environment would require the agent to decide what merits investigation, which concerns should become commitments, how those commitments should be scheduled, how far the agent should intervene, and whether unfinished work should continue as conditions change. Executable tests could assess whether an intervention works, while developer preferences could assess whether it was worth making. The central question would then shift from how many assigned tasks the agent completes to whether it discovers, schedules, executes, and preserves useful work as the project evolves.

6.2 GUI Agents

Table 4 Representative GUI-agent systems and mechanisms supporting proactiveness. Research systems precede commercial products, with each group ordered by first public release or announcement. The dimensions summarize documented mechanisms rather than implying complete proactive capability. Adaptation distinguishes training, execution-time adjustment, skill acquisition, and personalization. Platform denotes the interaction environment. Res. denotes research systems; Com. denotes commercial products.

System	Awareness	Anticipation	Agenda	Arbitration	Action	Adaptation	Platform	Type
WebPilot (Zhang et al., 2024i) 2024-08	Page state; execution feedback	Strategic route search	Global plan; local subtasks	MCTS action selection	Browser navigation	Feedback-driven adjustment	Web	Res.
Agent S2 (Agashe et al., 2025) 2025-04	Screenshots; GUI state	Hierarchical planning	Multiscale task plans	Observation-driven replanning	Grounded GUI operations	Execution-time plan revision	Desktop	Res.
SkillWeaver (Zheng et al., 2025a) 2025-04	Website state; practice outcomes	Useful-skill discovery	Growing skill library	Practice-based skill refinement	Browser API synthesis and use	Autonomous skill acquisition	Web	Res.
OSExpert (Liu et al., 2026h) 2026-03	Application state; feedback	Procedure discovery	Skills; compositional curriculum	Skill-boundary checks	Software exploration and skill execution	Reusable skill acquisition	Desktop	Res.
PRPF (Ding et al., 2026c) 2026-06	Multimodal mobile context	Perception-gated reasoning	Context-inferred assistance	Intervention gating	Assistance and API planning	Trained perception and reasoning modules	Mobile	Res.
Qwen-UI-Agent (Zhou et al., 2026a) 2026-07	GUI state; contextual events	Harness-based service initiation	Stateful workflows	Harness-directed execution	GUI-CLI operations	Data flywheel; online RL	Cross-platform	Res.
Act2Intention (Yan et al., 2026b) 2026-08	Action history; personal context	Personalized intention prediction	Predicted user intention	Intention-guided decisions	Mobile GUI operations	Experience-guided execution	Mobile	Res.
Google Magic Cue (Kugler, 2025) 2025-08	Personal apps; interaction context	Relevant-information retrieval	Immediate information need	Relevance-based cue display	Information and action suggestions	-	Mobile	Com.
Samsung Now Nudge (Samsung, 2026) 2026-02	Screen content; personal data	Contextual action recommendation	Immediate assistance opportunity	Recommendation; user selection	Calendar, photo, and autofill suggestions	Context-based personalization	Mobile	Com.
Meta Muse (Meta, 2026) 2026-09	Connected services; personal memory	Goal-aware suggestions	Persistent goals and plans	Permissions; critical-action approval	Suggestions and service execution	Memory and feedback personalization	Mobile; cloud Web	Com.
HONOR YOYO (HONOR, 2026) 2026-09	Orders; messages; device context	Event extraction; trigger matching	Stored tasks and schedules	Contextual trigger conditions	Schedule creation and conditional execution	-	Mobile	Com.
OpenAI dots (OpenAI, 2026c) 2026-09	Connected apps; goals; updates	Background opportunity discovery	Persistent projects	Read-only research; action review	Artifact preparation and authorized execution	Persistent context and feedback	Cross-platform	Com.

GUI agents translate user goals into sequences of reasoning, planning, and execution steps within digital environments by operating graphical interfaces (Nguyen et al., 2024), increasingly in combination with complementary programmatic interfaces such as command-line tools and code execution (Li et al., 2026i; Zhou et al., 2026b). Their ability to continuously observe and act within users’ digital environments creates opportunities for proactiveness beyond simply following fully specified instructions. We distinguish two forms of such behavior. First, an agent may take initiative *within an explicitly stated but underspecified goal*: for example, identifying intermediate objectives, acquiring missing information, exploring an unfamiliar interface, coordinating multiple applications or tools, or revising its plan after execution failures. Second, an agent may take initiative *beyond the explicit instruction* by inferring latent user needs from ongoing activities and context. For example, it may recognize from an incoming dinner invitation that the user may also need a calendar event, transportation arrangements, or a reminder. Existing work on proactive GUI agents has predominantly studied the former setting, particularly in long-horizon tasks where users provide a high-level objective but leave substantial execution details unspecified (Agashe et al., 2025; Tang et al., 2026b; Chai et al., 2026a). The latter

form, in which agents anticipate useful actions before they are explicitly requested, remains less extensively explored in current GUI benchmarks, but is central to the longer-term vision of personal agents that continuously reason over users’ activities, environments, and evolving needs. These two forms of initiative shift part of the burden of intermediate decision-making from the user to the agent, reducing the need for continuous specification and supervision.

How such proactiveness can be acquired and exercised, however, depends strongly on the computational platform in which the agent operates. Desktop environments provide comparatively mature and tool-rich software ecosystems, making autonomous execution across applications and interaction modalities increasingly feasible (Xie et al., 2024b; Agashe et al., 2025; Li et al., 2026i). Agents can combine graphical actions such as mouse clicks and keyboard input with command-line interfaces, code execution, and other structured tools, enabling flexible coordination over long-horizon workflows (Li et al., 2026i; Zhou et al., 2026b). Web environments are more open and dynamic: although agents can exploit structured representations such as HTML, accessibility trees, and the DOM, they must operate across heterogeneous websites, adapt to changing page states, and often discover relevant information, interaction paths, or site-specific procedures before determining how to proceed (Zhou et al., 2024b; Koh et al., 2024). Mobile environments, in contrast, expose especially rich personal and contextual signals that can support anticipation of latent needs (Kong et al., 2026; Chai et al., 2026a), while imposing tighter constraints on screen space, computation, latency, and execution robustness (Rawles et al., 2024; Chai et al., 2026b; Huang et al., 2026c). Recent work therefore explores lightweight or on-device agents, reusable application knowledge, and efficient interaction strategies to make proactive behavior practical under these constraints (Chai et al., 2026b; Huang et al., 2026c; Zhang et al., 2025i). Accordingly, we organize proactive GUI agents by deployment platform, focusing on the platform-specific challenges involved in acquiring proactiveness and the mechanisms proposed to address them. Table 4 compares selected research systems and commercial products across these platforms and the six mechanisms.

6.2.1 Desktop GUI Agents

Desktop environments provide a representative setting for GUI-based proactive agents, where agents perform long-horizon computer tasks such as processing emails, editing documents and spreadsheets, managing files, and using development tools as part of routine workflows (Xie et al., 2024b; Bonatti et al., 2024). Although users may specify only the overall goal, these environments are relatively stable and the intended outcome is often clear. Proactiveness therefore centers on interpreting underspecified instructions, inferring intermediate objectives, and coordinating the required applications and tools to complete the task. In this setting, initiative primarily lies in filling in the missing steps between a user’s high-level goal and successful execution. Existing benchmarks capture different aspects of this challenge. OdysseyBench evaluates cross-application dependencies spanning Word, Excel, PDF, Email, and Calendar (Wang et al., 2025g), while OfficeBench studies coordinated operations across productivity applications (Wang et al., 2024m). OS-Marathon further emphasizes extremely long and repetitive workflows, such as processing receipts or entering grades (Wu et al., 2026b), and OSWorld 2.0 extends desktop evaluation to realistic workflows involving hundreds of interactions, dynamically arriving information, and persistent constraints (Yuan et al., 2026b). Professional domains can extend these horizons even further: CADWorld, for example, requires agents to preserve geometric and engineering constraints while manipulating persistent CAD artifacts over long execution trajectories (Dong et al., 2026b). Together, these desktop-specific challenges suggest that proactive desktop agents are primarily about maintaining ownership of an extended workflow: determining intermediate objectives, carrying information across applications, and completing the work without repeatedly returning task decomposition to the user. OpenAI’s dots illustrate this direction in industrial systems by maintaining ongoing projects, responding to new information, and preparing work such as tested fixes and pull requests for review (OpenAI, 2026c). Their background “proactive research” uses read-only tools, distinguishing the discovery of useful work from its subsequent execution under the applicable permissions.

The first capability required for maintaining such workflow ownership is proactive planning: agents must identify what may be missing from underspecified user instructions and construct or revise a plan from limited input. Agent-S combines hierarchical planning with external knowledge and past experience (Agashe et al., 2024), while Agent S2 continuously revises plans across multiple temporal scales (Agashe et al., 2025). DeskCraft further allows agents to clarify requirements and adapt ongoing professional work in response to interruptions or user feedback (Wang et al., 2026j). LEGOMem stores procedural experience at both the orchestrator and execution levels to improve task decomposition and delegation across office workflows (Han et al., 2025a). UI-Copilot instead reduces the burden of long-horizon reasoning by separating persistent memory from transient execution context and invoking retrieval or computation tools only when needed (Lu et al., 2026b). OSWorld-Human highlights why such efficiency matters: much

of the latency in current computer-use agents arises from repeated planning and reflection, while leading agents still require substantially more steps than human trajectories (Abhyankar et al., 2025). Thus, while stronger planning is central to proactive desktop agents, sustained workflow ownership also requires recognizing and avoiding repeated low-level reasoning that makes long-horizon execution inefficient and difficult to scale.

Beyond proactively deciding what steps are needed, desktop agents must also decide how those steps should be executed. Many operations that require long and fragile graphical interaction can instead be completed through terminals, code, APIs, or structured tools, shifting desktop interaction from pure GUI control toward hybrid GUI–CLI execution. Proactiveness in this setting therefore includes recognizing when direct manipulation is unnecessarily costly and switching to a more suitable execution mechanism. UFO2 and ComputerRL combine visual interaction with native APIs or programmatic functions (Zhang et al., 2025c; Lai et al., 2025), while MCPWorld and OSWorld-MCP study hybrid GUI–tool execution and show both its potential and the difficulty of consistently invoking the right tools (Yan et al., 2025b; Jia et al., 2025). WeaveBench and CUA-Universe extend this paradigm to long-horizon trajectories that interleave GUI actions, command-line operations, and code execution (Li et al., 2026i; Shi et al., 2026a), while CLI-Anything exposes applications through machine-oriented command interfaces that can bypass repetitive graphical actions altogether (Yang et al., 2026g). Controlled GUI–CLI comparisons further suggest complementary bottlenecks: GUI agents struggle with grounded long-horizon interaction, whereas CLI agents are constrained more by skill coverage (Zhou et al., 2026b). Recent work on hybrid GUI–MCP agents and EE-MCP therefore treats execution modality itself as a learned decision, requiring agents to determine when tools should be invoked and how their outputs should be integrated with the current visual state (Fan et al., 2026a; He et al., 2026b). Taken together, these works suggest that desktop proactivity should extend from planning the workflow to selecting the most effective way to carry out each step: generating code to replace repetitive mouse actions, invoking an API when an operation is well specified, or returning to GUI interaction when spatial or transient interface state makes visual control necessary.

Such proactive execution, however, still assumes that the agent already possesses an appropriate procedure or skill. In open-ended desktop environments, this cannot be guaranteed, making proactive adaptation a further requirement: agents must recognize when their current repertoire is insufficient and respond by retrieving, adapting, or acquiring new procedural knowledge rather than immediately returning the capability gap to the user. CUA-Skill and VISUALSKILL represent reusable procedures as structured execution graphs or multimodal skills that can be retrieved and verified during new tasks (Chen et al., 2026f; Jiang et al., 2026), while OS-Copilot and HyMEM accumulate experience from previous trajectories and reuse it for later decomposition and adaptation (Wu et al., 2024d; Zhu et al., 2026c). When stored knowledge is insufficient, GUI-Bee explores unfamiliar interfaces to acquire grounding knowledge (Fan et al., 2025), OSExpert investigates software functionality and converts verified discoveries into reusable skills (Liu et al., 2026h), and SEAgent improves through iterative interaction and automatically generated curricula in unseen software (Sun et al., 2026b). EvoCUA-1.5 pushes this further through online interaction with executable desktop sandboxes (Huang et al., 2026b), while EE-MCP accumulates application-specific experience to improve future GUI–tool decisions (He et al., 2026b). These approaches point toward a broader hierarchy of desktop proactiveness: agents should first plan how to advance the workflow, then select the most effective execution modality, and finally adapt or learn new procedures when their existing capabilities are insufficient. Proactiveness at this level therefore means progressively reducing the number of decisions that must be returned to the user, from task decomposition to execution strategy to capability acquisition.

6.2.2 Web GUI Agents

Web proactiveness centers on continually assessing whether changing interface states still support the user’s objective and taking the initiative to adjust the workflow when its assumptions become unreliable. Websites differ in layout, navigation structure, terminology, and functionality, while execution states can change as content is updated, actions fail, or unexpected conditions arise. These conditions require agents to interpret emerging mismatches and determine an appropriate response before further actions compound the problem. Mind2Web evaluates cross-task, cross-website, and cross-domain generalization (Deng et al., 2023a), while WebArena spans heterogeneous interactive websites including shopping, forums, content management, maps, and collaborative development (Zhou et al., 2024b). VisualWebArena further shows that some tasks depend on visual information unavailable from textual structure alone (Koh et al., 2024). HTML, DOM trees, accessibility information, and page text provide useful structure, but real-world HTML can itself be large and noisy (Deng et al., 2023a), and AgentOccam demonstrates the importance of carefully designed observation and action representations (Yang et al., 2025c). These challenges make situational judgment central to Web proactivity: an agent must determine when to continue, recover, replan, explore, or seek clarification, without requiring the user to

diagnose every unexpected state.

The first capability required for such situational judgment is proactive monitoring: agents must detect emerging deviations from expected execution and determine whether recovery is needed before those deviations undermine the task. Existing interaction-based learning methods provide part of the foundation for such behavior. Early work used workflow-guided exploration to cope with sparse Web rewards (Liu et al., 2018), while WebRL learns through online reinforcement learning and a self-evolving curriculum partly generated from failed attempts (Qi et al., 2025). WebAgent-R1 similarly optimizes complete multi-turn browser trajectories from task-level outcomes (Wei et al., 2025b), with WebServ and WebGym providing infrastructure and realistic environments for scaling browser interaction and training (Lu et al., 2025b; Bai et al., 2026). These approaches help agents learn from failed trajectories and improve subsequent decisions, but outcome rewards typically reveal that something has already gone wrong rather than explicitly teaching the agent to anticipate when a plan is becoming unreliable. The stronger proactive capability is to detect early signals of mismatch, such as an unexpected page state or altered workflow, interpret their implications for the current objective, and intervene before the deviation propagates.

Beyond detecting deviations, proactive Web agents must recognize when missing environmental knowledge prevents reliable progress and initiate targeted exploration to resolve that uncertainty. An unfamiliar page or unexpected transition may require the agent to investigate available functions, relationships between interface states, or alternative navigation paths before continuing. WebPilot searches alternative routes when the correct path is uncertain (Zhang et al., 2024i), while Go-Browse more systematically explores Web environments to construct useful trajectories (Gandhi & Neubig, 2025). SAGE makes this process increasingly autonomous by first acquiring structural knowledge of a website and then generating progressively more difficult tasks for continued practice (Yang et al., 2025e). The important proactive decision is to determine when exploration is warranted. An effective Web agent should continue with the current workflow when observations remain consistent with the plan, perform local recovery or replanning when a specific step fails, explore more broadly when missing environmental knowledge prevents reliable progress, and seek user clarification when the remaining uncertainty cannot be resolved safely or depends on preferences or decisions that should not be inferred autonomously. Proactivity therefore lies in diagnosing the source of uncertainty and selecting the appropriate response.

Sustained Web proactiveness also requires retaining useful discoveries and recognizing when prior procedures can be reused or adapted to guide the response to a new situation. Agent Workflow Memory extracts recurring procedures from prior trajectories and retrieves them for later Web tasks (Wang et al., 2025o), while SkillWeaver explores new websites, discovers useful operations, and distills successful behavior into reusable APIs (Zheng et al., 2025a). SAGE similarly connects autonomous exploration with skill acquisition (Yang et al., 2025e), and WebRL generates new tasks around the evolving boundaries of the agent’s capabilities (Qi et al., 2025). An industrial counterpart appears in Edge’s Copilot Mode: Journeys groups related browsing activity to help users resume earlier tasks (Microsoft, 2025a). This illustrates contextual continuity across sessions, complementing research on retaining and reusing procedural knowledge.

These capabilities support proactive response selection: the agent must determine whether the current situation warrants continuing with a known procedure, adapting it, investigating further, or involving the user. Acquired skills and memory provide a basis for this judgment by helping the agent recognize both familiar situations and the limits of its existing repertoire. When unexpected states involve ambiguous preferences, consequential choices, authorization boundaries, or risks that the agent cannot reliably evaluate, the proactive response should be to seek clarification before proceeding. Web proactiveness can therefore be understood as an adaptive decision hierarchy: maintain the current workflow while its assumptions remain valid; detect and interpret deviations when they arise; recover or replan when the problem is understood; explore when missing environmental knowledge prevents reliable action; reuse or acquire procedural knowledge when appropriate; and involve the user when the uncertainty cannot or should not be resolved autonomously.

6.2.3 Mobile GUI Agents

Mobile proactiveness centers on anticipating latent information and assistance needs from the rich personal context available on devices closely embedded in users’ everyday lives. With appropriate access, agents can connect calendar events, emails, alarms, travel bookings, notifications, and application activity to anticipate what information or assistance would help the user next (Kong et al., 2026; Chai et al., 2026a). Initiative therefore begins with identifying a useful assistance objective: interpreting how an emerging event relates to the user’s plans, recognizing a possible need, and deciding whether it warrants attention. Meta’s Muse illustrates this broader direction by using persistent context about

users’ goals and connected services to offer proactive suggestions; it is accessed through its app or WhatsApp and executes tasks through a dedicated cloud virtual machine (Meta, 2026). Such systems highlight how personal context can support assistance beyond the immediate interaction. Mobile proactiveness consequently spans three connected decisions: what the user may need, whether and when assistance is warranted, and how to deliver it effectively.

Proactive need inference is the first ability required for such anticipation: agents must connect fragmented and evolving contextual signals to identify a useful assistance objective before the user fully articulates it. ProactiveMobile evaluates latent-intent inference from on-device context and its translation into executable API sequences (Kong et al., 2026), while Act2Intention connects intention understanding, personalized prediction, and execution through continuous intention–action trajectories (Yan et al., 2026b). These approaches suggest that anticipating useful assistance requires connecting current observations with prior activity and user preferences. For example, a flight cancellation may imply a need to find alternative transportation, check an affected booking, or reconsider an upcoming appointment. AppAgent-Pro addresses a related form of initiative by extending broad user requests into additional cross-application information seeking when further work is judged useful (Zhao et al., 2025c). Commercial features illustrate complementary forms of contextual assistance: Google Magic Cue retrieves relevant information across personal applications, such as surfacing flight details during an airline call, while Samsung Now Nudge uses screen content and stored context to suggest actions such as adding an event or sharing relevant photos (Kugler, 2025; Samsung, 2026). HONOR’s YOYO combines automatic schedule extraction with tasks that execute when user-specified contextual conditions occur (HONOR, 2026). These examples distinguish anticipating an immediate information need from carrying out a previously expressed conditional intention. In general, they show how mobile initiative can emerge from an unexpressed need, an incomplete request, or a remembered instruction whose triggering context has arrived.

Beyond inferring a possible need, proactive mobile agents must assess its relevance and timing to decide whether, when, and how to intervene. Mobile activity includes incidental browsing, repeated notifications, task switching, and intentions that may already have been fulfilled, making continuous observation insufficient grounds for intervention. PIRA-Bench evaluates proactive intent recommendation from noisy visual trajectories with interleaved intentions and user-profile context, while its PIRF baseline uses memory and state tracking to interpret ongoing activity (Chai et al., 2026a). PRPF explicitly separates intervention gating from assistance generation, using a lightweight perception stage to activate more expensive reasoning only when warranted (Ding et al., 2026c). Qwen-UI-Agent extends this direction through a lightweight harness supporting proactive service initiation and stateful workflows across mobile and computer environments (Zhou et al., 2026a), connecting the decision to offer assistance with the agent’s execution capabilities. The proactive decision is therefore whether to remain silent, retain a possible need for later reassessment, offer timely information, or propose an executable action. When the appropriate response depends on uncertain preferences or consequential commitments, focused clarification can establish the user’s intended outcome. Effective mobile proactiveness requires both sensitivity to emerging needs and restraint when the evidence or timing is insufficient.

Turning anticipated needs into useful assistance requires proactive capability development: agents must acquire and adapt the interaction policies needed to fulfill emerging objectives across changing applications. Mobile tasks often cross applications and depend on transient interface states, motivating executable environments such as AndroidWorld (Rawles et al., 2024). Online reinforcement learning strengthens this execution foundation: MobileGUI-RL combines self-exploration with trajectory-aware optimization of completion and efficiency, while MobileRL uses difficulty-adaptive replay, curriculum filtering, and trajectory-length-aware rewards (Shi et al., 2025c; Xu et al., 2025e). AgentCPM-GUI combines supervised trajectory learning with reinforcement fine-tuning, Mobile-Agent-v3.5 extends training across environments and model scales, and WM-R1 uses learned world models to reduce reliance on costly real-device rollouts (Zhang et al., 2025i; Xu et al., 2026a; Han & Qian, 2026). These methods develop the ability to execute assistance after a useful objective has been identified. Exploration and self-evolution further help agents maintain that ability as applications change: UI-Voyager improves supervision from successful and failed trajectories, while MobileGUI-RL generates learnable tasks through autonomous exploration (Lin et al., 2026b; Shi et al., 2025c). From a proactive perspective, their value lies in building the execution capabilities needed to fulfill anticipated needs.

Sustained mobile proactiveness further requires recognizing recurring assistance needs and reusing accumulated knowledge and procedures to address them efficiently under device constraints. UI-KOBE explores applications in advance and represents their states and transitions as reusable knowledge graphs, while SkillDroid compiles successful trajectories into parameterized skills that can be replayed without repeated language-model calls (Chai et al., 2026b; Chen et al., 2026d). Such reuse supports timely assistance under mobile constraints on model size, memory, latency, energy, connectivity, and privacy (Chai et al., 2026b; Huang et al., 2026c; Zhang et al., 2025i). AgentCPM-GUI reduces

generation overhead through a compact action representation; MobileExplorer overlaps lightweight exploration with model inference and stores traces as structured memory; and Mobile-Agent-v3.5 explores multiple model sizes and cloud-edge collaboration (Zhang et al., 2025i; Huang et al., 2026c; Xu et al., 2026a). Combined with PRPF’s selective activation of deeper reasoning (Ding et al., 2026c), these approaches suggest an architecture in which relevant context is maintained efficiently, stronger reasoning is invoked when a meaningful need emerges, and familiar procedures are reused during execution. Mobile proactiveness thus connects contextual anticipation with selective intervention and adaptive execution, reducing the user’s effort in recognizing, formulating, and initiating useful assistance.

6.2.4 Evaluation, Challenges, and Open Directions

Evaluation of proactive GUI agents should assess whether agents exercise appropriate initiative throughout the assistance process, while reflecting the distinctive demands of each platform. For desktop agents, this means evaluating whether they infer missing intermediate objectives, coordinate long-horizon workflows, select suitable execution mechanisms, and address capability gaps without unnecessary user supervision. For Web agents, evaluation should examine whether they detect deviations, distinguish execution failures from missing knowledge, and appropriately choose among recovery, replanning, exploration, procedural reuse, and clarification. For mobile agents, it should assess whether contextual signals support the inferred information or assistance need, whether intervention is timely and useful, and whether the resulting assistance fulfills that need. Existing benchmarks capture complementary parts of this process: ProAgentBench and PIRA-Bench assess the appropriateness and timing of assistance (Tang et al., 2026b; Chai et al., 2026a), Act2Intention links intention prediction to subsequent GUI execution (Yan et al., 2026b), OpenComputer emphasizes verification of the resulting application state (Wei et al., 2026a), and OSWorld-Human highlights interaction length and latency (Abhyankar et al., 2025). A complete evaluation should therefore connect the quality of the proactive decision to its execution outcome, accounting for unnecessary actions, computation, supervision, and interruption alongside task success.

These distinctions motivate platform-specific open directions within a shared challenge of deciding when and how to take initiative. Desktop agents should strengthen sustained workflow ownership by adapting plans, selecting among GUI, terminal, APIs, and stored skills, and acquiring procedures when their current capabilities are insufficient (Li et al., 2026i; Shi et al., 2026a). Web agents should improve situational judgment by recognizing when workflow assumptions no longer hold and determining whether uncertainty calls for local recovery, broader exploration, knowledge acquisition, or user clarification (Yang et al., 2025e; Zheng et al., 2025a). Mobile agents should better connect evolving personal context with latent needs, calibrate intervention to relevance and timing, and deliver assistance through reusable procedures and selective reasoning (Kong et al., 2026; Chai et al., 2026a; Ding et al., 2026c; Chai et al., 2026b; Chen et al., 2026d; Huang et al., 2026c). Across platforms, stronger initiative must remain aligned with user intent and authorization, especially when objectives are inferred or external content introduces competing instructions (Levy et al., 2026; Evtimov et al., 2025). The central research direction is therefore to develop agents that take responsibility for advancing desktop workflows, adapt appropriately to changing Web situations, and anticipate useful mobile assistance, while recognizing when further action requires user involvement.

6.3 Information Access Agents for Search and Recommendation

Table 5 Representative proactive information-access systems. Res. denotes research systems; Com. denotes commercial products.

System	Awareness	Anticipation	Agenda	Arbitration	Action	Adaptation	Deploy.	Type
Remembrance Agent (Rhodes et al., 1996)	User’s current text / context	Relevance model	Current text only	Continuous, non-interrupting	Surfaced related docs (<i>suggestion</i>)	—	Desktop	Res.
Self-RAG (Asai et al., 2023)	Query + parametric state	LLM	Given query	Reflection tokens (retrieve-when-uncertain)	Retrieval + grounded answer (<i>gated retrieval</i>)	Trained critic tokens	Software	Res.
RecThinker (Zhang et al., 2026a)	User / item evidence	LLM agent	Given recommendation task	Gather-more-if-uncertain	Item recommendations (<i>suggestion</i>)	SFT + RL	Software	Res.
STEPS (Zhang et al., 2026m)	Push and behavior logs	Value of pushing now vs. later	Platform retention objective	Push / no-push; self-scheduled wake-up	Push notification (<i>action</i>)	Offline RL	App	Com.

Information access is one of the oldest settings for proactive behavior: recommender systems offer items no one asks for (Resnick et al., 1994), filtering systems decide what to deliver instead of waiting to be queried (Belkin & Croft, 1992), and early retrieval tools surface documents from a user’s working context before any search box is used (Rhodes et al., 1996). These systems already perform the two operations that define proactive behavior: predicting a latent information need (Anticipation, §4.2) and deciding whether and when to present unrequested information at an acceptable cost in attention, relevance, and trust (Arbitration, §4.4). What they share with ordinary information retrieval (IR) is general machinery: ranking, relevance modeling, query understanding, and session modeling. What makes them proactive is that the action comes before any request, which forces the system to decide for itself what to act on and whether acting is worth the interruption. Today this proactiveness works over short horizons: what ships at scale is contextual pre-fetch and zero-query retrieval keyed to immediate context, such as recent app usage or the current document, alongside recommenders that predict the next item from recent behavior (Peng et al., 2025), each serving the need the user has right now rather than a goal it holds and revises over time; the most prominent product that inferred such goals from memory and researched them overnight, ChatGPT Pulse, was retired within a year in favor of tasks the user schedules (OpenAI, 2025b; Digit, 2026). The current wave of adaptive and agentic retrieval-augmented generation (RAG) (Asai et al., 2023; Li et al., 2025e) and deep-research agents (Zheng et al., 2025d) points toward multi-source anticipation that holds a user’s evolving goals over longer horizons, gathers evidence across many sources on its own initiative, and revises what it surfaces as those goals change. Information access is therefore anticipation-rich but agenda-shallow relative to the other verticals here: its ability to predict an immediate need is mature and widespread, yet it rarely maintains a longer intent on the user’s behalf. Table 5 compares selected systems for contextual information delivery, recommendation, and retrieval across the six mechanisms. We take the two faces of information access in turn, recommendation and search, then ask how proactiveness is evaluated and what risks it raises.

6.3.1 Recommendation as Proactive Filtering

Recommendation is proactive by design, because it presents items when the user has issued no query at all. This is the classic push model of information access rather than the pull model of search. Belkin & Croft (1992) describe recommendation as the mirror image of retrieval and named it information filtering, a system that serves a standing interest profile by choosing what to deliver rather than answering a one-off query. Early collaborative filtering systems such as Tapestry and GroupLens do exactly this, predicting which items in an incoming stream a user would value (Goldberg et al., 1992; Resnick et al., 1994), and later work organizes the area around the push-versus-pull distinction (Hanani et al., 2001). In our terms, the recommender’s Anticipation is a model of the latent need encoded in past behavior, and its Arbitration is the choice of which few items to surface within limited screen space. Whether to surface them at all is a permission set by the platform, which is the capability-versus-autonomy distinction of §4.5.1. Sequential and session-based recommenders sharpen the Anticipation step by predicting the next interaction from an unfolding session (Hidasi et al., 2016; Kang & McAuley, 2018; Sun et al., 2019), which is the core of anticipatory recommendation: predicting a need before the user acts on it. Feed and notification systems make the Arbitration step explicit, since sending too much is directly costly and a heavy notification load lowers user mood (Pielot et al., 2014), so bandit and reinforcement-learning controllers decide when to send a notification at all, balancing engagement against fatigue (Yancey & Settles, 2020; Rizk et al., 2020); STEPS, a push agent deployed on Douyin, also schedules its own next wake-up and counts permission disablement as the cost of interruption (Zhang et al., 2026m). LLM-based and agentic recommenders carry the same structure forward (Peng et al., 2025; Wu et al., 2026e): a tool-augmented recommendation agent checks whether it has enough evidence about the user and the items, then gathers more before recommending, which mirrors the idea of retrieving only when uncertain (Zhang et al., 2026a), while interactive recommenders optimize engagement over many turns (Feng et al., 2026c; Zhong et al., 2026). Recommendation also turns proactive in dialogue: proactive cues can prompt a user to try decision aids they would otherwise miss, and users tend to accept these prompts (Ma & Ziegler, 2024), while target-driven conversation planning orders the system’s dialogue moves to lead the user toward a chosen item, a self-generated Agenda (§4.3) aimed at a system-side goal (Wang et al., 2022).

6.3.2 Proactive Search: From Zero-Query to Agentic Retrieval

Search is the active, query-driven side of information access, and it turns proactive in **three ways**: by anticipating the need before any query, by asking before it answers, and by self-initiating retrieval inside its own reasoning. The **first** move is to stop waiting for a query. Just-in-time retrieval agents monitor the user’s current activity, such as the document being written or the page being read, and offer related material with no query and without interrupting (Rhodes &

Maes, 2000). The Remembrance Agent is the canonical early example: it runs continuously and shows short summaries relevant to the user’s current text, on the premise that ordinary search cannot help when the user does not yet know enough to ask (Rhodes et al., 1996). Watson reads the user’s activity in everyday applications, infers the task from the text on screen, and retrieves related documents without a query (Budzik & Hammond, 2000), while Letizia browses ahead of the user to find the links they are likely to want next (Lieberman, 1995). In all of these the need is read from context rather than typed, so Awareness of the user’s task state (§4.1.1) is the enabling layer and Anticipation is the distinctive act; wearable versions such as Jimminy carry the same idea to physical context like location and time of day (Rhodes & Maes, 2000). The line continues into contextual mobile assistants, with Google Now’s predictive cards as the best-known commercial case of surfacing the right card before it is requested (Rubin, 2012), and Pixel’s Magic Cue carries it on device, surfacing a fact or action from mail, calendar, or messages when the current exchange needs it (Kugler, 2025) (Table 4).

The **second** move is to ask before answering. When a query is unclear or underspecified, a well-calibrated system can hold back its answer and ask a clarifying question instead, trading one extra turn against the risk of confidently answering the wrong question. Learning to ask good clarifying questions in search and question answering is a research area of its own, from the Qulac and ClariQ datasets to rankers that score a candidate question by how much it is expected to reduce uncertainty (Aliannejadi et al., 2019; Rao & III, 2018; Aliannejadi et al., 2021). The proactive part is again an Arbitration decision, namely when to ask rather than answer (§4.4.1). CLAM asks for clarification only when a request is ambiguous (Kuhn et al., 2023a), CLAMBER tests whether models can even tell that clarification is needed and finds that they often cannot (Zhang et al., 2024g), and related work trains agents to judge how vague a task is and to recover the user’s implicit intent before acting (Qian et al., 2024). Asking too often is as poorly calibrated as never asking, the same two-sided failure that runs through proactive behavior. The posture extends to multi-turn settings, where a conversational system can volunteer suggestions, anticipate follow-up needs, and guide the discussion (Deng et al., 2023b): Andolina et al. (2018) studied this directly, surfacing relevant information into an ongoing conversation without being asked, and PACIFIC casts proactive question answering in which the system opens a clarification over mixed table and text data (Deng et al., 2022). Recent information-gathering agents push the gate further, learning policies for which question reduces uncertainty the most (Choudhury et al., 2026; Echarghaoui et al., 2026), and icebreaker agents study how to open an interaction with a user the system knows nothing about (Zheng et al., 2026a), while IntentRL trains a deep-research agent to clarify latent intent before committing to an expensive run (Luo et al., 2026b).

The **third** move folds retrieval into the agent’s own reasoning: instead of always retrieving, the agent decides from its own uncertainty when and what to fetch. Self-RAG learns special tokens that decide when to retrieve and whether a passage helps (Asai et al., 2023), FLARE retrieves whenever it is about to generate a low-confidence span (Jiang et al., 2023), and Adaptive-RAG routes a query to no retrieval, a single retrieval step, or several steps according to its predicted difficulty (Jeong et al., 2024; Wang et al., 2025n). At longer horizons the same decision becomes a multi-step plan: Self-Ask has the model raise and answer its own follow-up questions, sending them to a search engine when useful (Press et al., 2023); Search-o1 lets a reasoning model trigger retrieval mid-chain whenever it reaches a point it is unsure about (Li et al., 2025e); and deep-research agents run many rounds of self-directed search and synthesis over the open web, trained end to end to do so (Zheng et al., 2025d; Zhong et al., 2026), and Omakase runs such research in the background, issuing queries as a user’s project documents evolve and distilling the reports into a few actionable suggestions (Siangliulue et al., 2026). Each of these applies the Arbitration gate of §4.4.1 to the act of retrieving, so the agent reaches for outside information only when answering without it is the riskier choice. The capability-versus-autonomy line matters here (§4.5.1): an agent that issues many unrequested queries is exercising self-initiated capability, but the freedom to spend that compute, and to act on what it finds, is autonomy that its operator grants. What is new compared with early just-in-time retrieval is that the agent can read, reason over, and re-query its own retrieved evidence.

6.3.3 Evaluation, Challenges, and Open Directions

Evaluating proactive information access takes more than standard IR metrics, which reward the right answer to a stated query. A proactive system must also be scored on predicting a need early and surfacing it without over-triggering, so evaluation has to account for timing and for need prediction (§5). ProCIS provides timing annotations and normalized proactive discounted cumulative gain (npDCG) for evaluating whether useful documents are retrieved at appropriate moments in a conversation (Samarinas & Zamani, 2024). PREVAL instead jointly rewards early proactivity and the quality of subsequent suggestions (Sen et al., 2018), while JIR-Arena defines the just-in-time recommendation task

across modalities (Yang et al., 2025d). The common point is that acceptance and interruption cost, not relevance alone, become the targets, and that a proactive system can fail in two opposite ways: missing a need it should have served, or surfacing too aggressively. The same power to choose what the user sees also carries risks. Some are shared by any system that surfaces content, and we treat them with the governance directions (§7.2). What is specific to information access is that surfacing corrupts its own evidence: personalized surfacing narrows exposure into filter bubbles (Pariser, 2011), and a recommender that learns from the behavior it shapes falls into degenerate feedback loops that corrupt the very signal Anticipation depends on (Jiang et al., 2019). These risks make Arbitration (§4.4) not only a utility one but a normative gate. Several open directions are specific to this domain. Agents could hold a persistent, revisable model of a user’s evolving goals across sessions, instead of reading each need from the immediate moment. Surfacing could be made diversity- and privacy-aware by design, resisting filter bubbles and feedback loops and keeping always-on awareness on the device. Evaluation still needs benchmarks that score need prediction, timing, and interruption cost together, and that reward the correct decision to stay silent. In information access, mature proactiveness is not surfacing or retrieving more, but understanding a user’s evolving goals well enough that its silence is as deliberate as its initiative.

6.4 Embodied Agents and Robotics

Table 6 Representative embodied and robotic systems and mechanisms supporting proactiveness. The dimensions summarize documented mechanisms without implying complete proactive capability. Agenda distinguishes immediate or externally assigned task goals from persistent or self-initiated commitments. Res. denotes research systems; Com. denotes commercial products.

System	Awareness	Anticipation	Agenda	Arbitration	Action	Adaptation	Deploy.	Type
Gemini Robotics 2 / ER 2 (Google DeepMind, 2026a,b)	Live video, language, proprioception	ER 2 embodied reasoning + whole-body VLA	Given task; ER 2 subgoal plans	Progress tracking; tool calls; multi-robot hand-offs	Whole-body humanoid and bi-arm manipulation (<i>action</i>)	Failure-driven replanning	Robot	Com.
Helix 2.5 (Figure AI, 2026c)	Vision, proprioception	VLA pretrained on human video (Index)	Given household task	Given-task execution	Household chores in unseen homes (<i>action</i>)	Execution-time self-correction	Robot	Com.
$\pi_{0.6}^*$ (RECAP) (Intelligence et al., 2025a)	Vision, language, proprioception	VLA + learned advantage	Given long-horizon task	Advantage-conditioned actions	Laundry, espresso, box assembly (<i>action</i>)	Offline RL from rollouts and corrections	Robot	Com.
GR00T N1 / N1.7 (Bjorck et al., 2025; NVIDIA, 2026b)	Vision, language, proprioception	Dual-system VLA (VLM + diffusion head)	Given language instruction	Given-agenda execution	Manipulation, motor control (<i>action</i>)	-	Robot	Com.
IX NEO (IX Technologies, 2026a, 2025)	Vision, audio, language	Redwood VLA	User-requested chores; scheduled sessions	Scheduled remote-expert supervision	Household chores (<i>gated action</i>)	Teleop-to-autonomy flywheel	Robot	Com.
Waymo Driver (Waymo, 2024)	Lidar, camera, radar, maps	Road-user behavior prediction	Dispatched trips within ODD	ODD gating; minimal-risk maneuver; fleet assistance	Driverless ride-hailing (<i>autonomous in ODD</i>)	Fleet data and simulation	Vehicle fleet	Com.
CoFRIDA (Schaldenbrand et al., 2024)	Canvas image, language goal	Canvas completion; stroke simulation	Shared canvas goal; per-turn plans	Human co-creative turn-taking	Physical brush strokes (<i>action</i>)	-	Robot	Res.
ProAct (Zhang et al., 2026l)	Egocentric vision, audio, transcript	Cognitive LLM infers motivation and intentions	Self-initiated social goals	Silence-vs-intervene decision	Speech, gesture, locomotion (<i>action</i>)	Episodic memory, context compression	Robot	Res.
AssistantX (Sun et al., 2024b)	User context + robot sensors	LLM + memory unit	Assigned task; help-seeking subgoals	Multi-agent LLM loop; asks nearby humans	Physical + cyber actions (<i>gated action</i>)	Memory, reflection agent	Robot + SW	Res.
GAP (Patel & Chernova, 2026)	Passive observation of user and scene	Learned goal anticipation	Self-selected household needs	Confidence, preconditions, timing, acceptance	Fetch and prepare objects (<i>bounded autonomy</i>)	Routines from observation	Sim	Res.
GLOBE (Gonzalez et al., 2026)	Activity and object-use history	Routine model; LLM when uncertain	Predicted-routine needs	Confidence threshold gates action	Object fetch ahead of need (<i>bounded autonomy</i>)	Per-household statistics	Robot	Res.

Continued on next page

Table 6 (continued)

System	Awareness	Anticipation	Agenda	Arbitration	Action	Adaptation	Deploy.	Type
NIA framework (Zhang et al., 2026i)	Given human task sequence, scene	Cross-step assistance scoring	Self-chosen complementary sub-tasks	Intrusiveness-aware ranking	Parallel manipulation, with a non-blocking protocol (<i>bounded autonomy</i>)	-	Sim + Robot	Res.
RoboGuard (Ravichandran et al., 2026)	Scene semantics, command, world model	LLM derives temporal-logic safety specs	Standing safety rules; given plan	Control synthesis vetoes or repairs plans	Spec-compliant plans only (<i>gated action</i>)	-	Robot + Sim	Res.

-: not separately reported in the cited source.

Embodied agents represent one of the most promising frontiers for proactive agency. By bridging digital reasoning and physical execution, they extend proactivity beyond virtual environments into the physical world. Whereas a proactive software agent may recommend an action, schedule a task, or generate information, a proactive embodied agent can navigate its surroundings, manipulate objects, operate equipment, and directly affect people and infrastructure. Therefore, every stage of the proactive loop must account for the constraints of physical interaction, including uncertainty, latency, safety, and the limited reversibility of physical actions.

Despite rapid progress in robotics, most deployed embodied systems remain either *reactively autonomous* or exhibit only *bounded proactivity*. Reactively autonomous robots can perform sophisticated tasks, such as grasping diverse objects or navigating dynamic environments, but generally act in response to explicit commands, predefined triggers, or externally assigned workflows. Bounded proactive agents go further by initiating actions within a defined task, environment, or permission structure, but they still lack the reliability and generality required for open-ended proactive behavior. A *fully proactive* robot, by contrast, must recognize emerging opportunities and risks, anticipate future needs, decide whether and when to intervene, communicate its intentions, and adapt its behavior over time. The long-term goal is not only greater autonomy, but robots that act as collaborative partners: taking useful initiative while remaining responsive to human goals, preferences, and boundaries. Realizing this vision requires embodied agents to close the full proactive loop defined by the six-A mechanism introduced in §4. Table 6 maps selected embodied systems and mechanisms supporting proactiveness onto this framework.

Several rapidly developing technologies are bringing this vision closer. Simulation (Li et al., 2023b; Wang et al., 2023b; Authors, 2024; Zakka et al., 2025; Mittal et al., 2025; Wen et al., 2026; Li et al., 2026m), teleoperation (Zhao et al., 2023; Wu et al., 2024b; Khazatsky et al., 2024), wearable and egocentric data collection (Chi et al., 2024; Xu et al., 2025a; Hoque et al., 2025; Punamiya et al., 2026; Zheng et al., 2026b), and learning from third-person images and videos (Chen et al., 2025b, 2026b) are expanding the scale and diversity of robot-learning data, enabling robots to acquire and generalize increasingly complex skills. Egocentric human video now reaches tens of thousands of hours (Zheng et al., 2026b; Build AI, 2025; Figure AI, 2026d; Luo et al., 2026a), open cross-embodiment robot corpora exceed a thousand hours and deliberately retain failure and recovery trajectories (Hou et al., 2025; Wu et al., 2025d; AGIBOT, 2026), and simulation suites scale household tasks to hundreds of activities across thousands of scenes (Nasiriany et al., 2026; Kim et al., 2026d; Stanford Vision and Learning Lab, 2026). Large language models and vision-language models contribute semantic knowledge, high-level reasoning, and natural-language interaction (Ahn et al., 2022; Driess et al., 2023; Yang et al., 2025f). Vision-Language-Action models ground language instructions and visual observations in robot actions (Team et al., 2024; Kim et al., 2024a; Black et al., 2024; Intelligence et al., 2025b; Liu et al., 2025d; Team et al., 2025; Bjorck et al., 2025; Wang et al., 2026f; Generalist AI, 2026; Genesis AI, 2026; Physical Intelligence et al., 2026; Luo et al., 2026a; Team et al., 2026; Yu et al., 2026e; Generalist AI, 2025). The 2026 generation of these models controls whole humanoid bodies and coordinates several robots (Google DeepMind, 2026a), transfers zero-shot to dozens of unseen homes (Figure AI, 2026c), and attributes most of that generalization to human-video pretraining with explicit scaling laws (NVIDIA, 2026b; Figure AI, 2026c). World-action models capture how actions may change future states, supporting anticipation, planning, and policy learning (Ye et al., 2026b; Agarwal et al., 2026; Garrido et al., 2026; Gao et al., 2026d; Guo et al., 2026a; Li et al., 2026e); they have become a systematic research program with open modular stacks (Wang et al., 2026m), and in self-evolving variants one network serves as policy, simulator, and evaluator so that candidate actions are imagined and scored before execution (Bi et al., 2026), at the price of several-fold higher inference latency than a comparable VLA (NVIDIA, 2026c). Policies also improve after deployment: offline reinforcement learning from autonomous rollouts and expert corrections roughly doubles throughput on hard real-world tasks (Intelligence et al., 2025a), and unified RL frameworks for VLAs (Zang et al., 2025), hours-scale online post-training on physical robots (Yang et al., 2026b), world-model-based policy optimization (Zhu et al., 2025a;

Wang et al., 2026a), and in-context learning from a single video demonstration (Skild AI, 2026) shorten the loop from experience to behavior. Neuro-symbolic methods combine learned representations and policies with explicit symbols, rules, and planners, supporting compositional reasoning, constraint-aware planning, and more interpretable execution (Mao et al., 2023; Luo et al., 2024b; Yu et al., 2026d; Nie et al., 2026; Pan et al., 2026; Albinhassan et al., 2026; Kamale & Berenson, 2026). At the system level, agentic architectures are beginning to orchestrate perception, reasoning, task decomposition, memory, skill selection, and low-level control within closed-loop robotic workflows (Gu et al., 2025; Li et al., 2026g; Xiao et al., 2026; Lu et al., 2026a), increasingly with readiness checks, progress verification, pre-actuation validation gates, and long-term memory with risk prediction implemented as separate, inspectable modules (Wang et al., 2026i; Liu et al., 2026p; Li et al., 2026h), and with embodied-reasoning models that watch live video, track task progress, call tools, and hand work between robots (Google DeepMind, 2026b).

These advances do not, by themselves, make a robot proactive. Rather, they provide the fundamental capabilities needed for proactive robots to determine what to do next and when to act. In the near term, the commercial market is therefore likely to be defined by bounded proactivity. The most valuable embodied systems may not be those that operate with the greatest independence, but those that combine useful initiative with clear operational limits, calibrated uncertainty, measurable value, and reliable mechanisms for human oversight and escalation. This suggests a practical progression from reactive autonomy to bounded proactivity and, eventually, to general-purpose proactive embodied agents.

The following sections examine four representative verticals: household services, autonomous driving, industrial collaboration, and human-agent co-creation. These domains overlap to some extent with the discussions of smart homes, smart cities, and entertainment and media in §6.7, §6.8, and §6.12, respectively. Notably, the emphasis here is on physically embodied agents that act in the real world.

6.4.1 Household Services: From Smart Appliances to Proactive Domestic Partners

A proactive household robot should understand not only explicit requests, but also the routines, preferences, permissions, and changing conditions of the home. It may recognize when a task needs attention, choose an appropriate time to act, avoid disturbing residents, ask before handling unfamiliar belongings, and remember household preferences. Rather than waiting for the command “clean the kitchen”, for example, it might notice that dishes have accumulated, determine whether the dishwasher is available, and act at a time that does not interfere with meal preparation. In this setting, proactivity means being helpful without becoming intrusive. This section focuses on embodied agents that perform physical tasks in domestic environments, while proactive smart-home systems are discussed more broadly in §6.7.

The home represents a major potential market for embodied AI, but also one of its most difficult operating environments. Unlike factories and warehouses, homes vary widely in layout, objects, routines, and social norms. Thus, commercial household robotics remains dominated by specialized products, such as robotic vacuum cleaners, lawn mowers, and snow-removal systems, whose tasks and operating boundaries are tightly constrained (Jones et al., 2005; Matic, 2026; Segway Navimow, 2026; Yarbo Inc., 2026). Appliance makers are extending this envelope incrementally rather than leaping to general manipulation: Roborock’s Saros Z70 vacuum carries a five-axis arm limited to light objects and a fixed catalogue of object types, Dreame and LG showed arm-equipped home robots at CES 2026, and Samsung quietly shelved its Ballie companion robot (Roborock, 2025; Engadget, 2026; ChannelNews, 2026).

The emerging frontier is proactive, general-purpose household manipulation. Companies such as 1X, Sunday Robotics, Figure, and Weave Robotics are developing humanoid or mobile-manipulation systems for tasks including tidying, dishwashing, laundry handling, and household organization (1X Technologies, 2026a; Sunday Robotics, 2026; Figure AI, 2026b; Weave Robotics, 2026). These systems demonstrate increasingly varied and extended workflows, but reliable operation across an entire home remains an open challenge. The first measured generalization numbers appeared in 2026: Figure 03 running Helix 2.5 performed bed making, towel folding, and toy tidying in 30 previously unseen homes without any in-home data, succeeding in 56% of trials, against 9% for a model trained without large-scale human-video pretraining (Figure AI, 2026c). In simulation, the winning entry of the 2025 BEHAVIOR Challenge completed only 12.4% of its 50 long-horizon household tasks end to end, and the best agents on the LongAct benchmark reach 16% full-task success even with frontier language models as planners (Larchenko et al., 2025; Stanford Vision and Learning Lab, 2026; Zhu et al., 2026e). As of this writing, none of these general-purpose home robots has confirmed consumer deliveries: 1X says some NEO orders will ship in 2026 and others later, Sunday plans a 50-household beta in late 2026, and Weave’s first shipments are scheduled for fall 2026 (heise online, 2026; Sunday Robotics, 2026; Weave Robotics, 2026). The gap between a convincing demonstration and a useful product is especially large in domestic

settings: occasional failure may be acceptable in a laboratory, but frustrating or unsafe in everyday use.

The central technical challenge is *domestic generalization*: maintaining reliable performance across variations in layouts, objects, clutter, lighting, residents, and routines. A robot entering a new home must quickly learn not only where objects are located, but also how that household organizes its space and which actions its residents consider appropriate. This requires navigation and manipulation to be combined with contextual reasoning, memory, preference learning, and permission-aware decision-making. Human-in-the-loop operation offers one practical path toward this capability (1X Technologies, 2026a, 2025). Remote operators can help robots resolve unfamiliar situations and complete useful tasks while generating demonstrations from real homes. These interventions can create a data flywheel through which recurring situations are gradually incorporated into autonomous policies, providing a bridge from supervised household services to increasingly autonomous and proactive behavior. 1X’s NEO follows this model, with a remote expert able to supervise at scheduled times and a world-model effort aimed at reducing reliance on teleoperators (1X Technologies, 2026a,c), and Weave’s Isaac 1 operates autonomously by default with teleoperation as a fallback and physical cues that show when the robot is working (Weave Robotics, 2026). Two alternative flywheels avoid remote operators in the customer’s home altogether. One collects paid first-person video of people doing chores in their own homes: Figure’s Index app reports 16 million videos from more than 100 countries and accounts for most of Helix 2.5’s zero-shot capability (Figure AI, 2026d), Sunday’s Skill Capture Glove recorded roughly 10 million episodes in more than 500 homes without a robot present (Robotics 24/7, 2025; Humanoids Daily, 2026), and open datasets such as EgoWild2Dex follow the same recipe (Lin et al., 2026a). The other licenses large multi-embodiment foundation models to hardware makers, as with Gemini Robotics 2 and $\pi_{0.7}$, the latter operating unseen kitchens and appliances zero-shot (Google DeepMind, 2026a; Physical Intelligence et al., 2026). Whichever route is taken, in-deployment learning must not erase earlier skills: naive sequential fine-tuning of vision-language-action models on a stream of real-world tasks causes severe forgetting, which experience replay largely prevents (Zhu et al., 2026b). However, remote access to a household robot also raises significant privacy concerns, particularly when operators can view sensitive domestic spaces or activities. Privacy risks persist even without a human viewer: the structured maps, object lists, and task summaries a robot exports can re-identify households and reveal target objects, and no single abstraction removes this leakage across tasks (Xu & Ayday, 2026). The viability of this model will therefore depend on whether users are willing to accept such oversight and whether strong safeguards, transparency, and user controls can make remote operation sufficiently trustworthy.

Commercial adoption will depend on three closely related factors. First, physical capability does not imply authorization. Entering a room, moving an object, or discarding an item may violate privacy, ownership, or household norms. A proactive domestic agent must therefore judge which actions are permitted, whose preferences take priority, and when confirmation or inaction is appropriate (Li et al., 2026b). Recent work begins to measure and build this judgment. In the NormAct benchmark, multimodal agents reach the goal in 67.4% of household scenarios but comply with the unspoken norm in only 24.7% (Zhao et al., 2026c); ask-or-act policies learn when accumulated interaction history suffices to act and when a clarification question is worth its cost (He et al., 2026a); EgoAsk lets the robot ask context-relevant questions about a user’s belongings through smart glasses (Hu et al., 2026c); routine-anticipation models predict a resident’s next activities and defer to a language model only when uncertain (Gonzalez et al., 2026); and a four-day in-home study with ten families found that robot-initiated reminders raise tensions over timing and authority among residents (Xu et al., 2026e). Closed-loop evaluation with an adaptive human model further shows that proactive methods that look strong offline can collapse and increase user workload (Patel & Chernova, 2026). Second, reliability must transfer across homes without extensive preparation or per-household engineering. Systems that perform well in controlled demonstrations must retain that reliability in unfamiliar and changing environments. Third, cost must remain commensurate with practical value. Early general-purpose home robots and prototypes remain expensive relative to specialized appliances, raising expectations for both reliability and breadth of capability: NEO is priced at \$20,000 or \$499 per month, Sunday’s hand-built Memo units cost about \$20,000 with retail expected to fall by at least half, and Weave’s Isaac 1 is listed at \$7,999 or \$449 per month (1X Technologies, 2026b; Sunday Robotics, 2026; Weave Robotics, 2026).

The most successful household robot may therefore not be the one with the broadest initial skill set. Instead, it may be one that performs a small number of valuable tasks reliably across different homes, at an acceptable cost, without undermining privacy, personal preferences, autonomy, or comfort. Such a system would mark a meaningful transition from a smart appliance to a trusted proactive domestic partner.

6.4.2 Autonomous Driving: Proactive Embodied AI at Commercial Scale

A proactive autonomous vehicle must continually anticipate how a traffic scene may evolve and respond before contingencies happen. It cannot rely on a passenger, safety driver, or remote operator to guide it through each decision. Instead, it must predict the behavior of pedestrians and other road users, negotiate right-of-way, select safe stopping locations, reroute around disruptions, and determine when uncertainty calls for more cautious behavior, a minimal-risk maneuver, or operational assistance. This section focuses on autonomous vehicles as embodied agents, with the broader role of proactive agents in smart cities discussed in §6.8.

Autonomous driving is among the most commercially mature examples of proactive embodied AI operating in open, human-populated environments. Waymo and Pony.ai offer fully driverless commercial robotaxi services across multiple metropolitan areas (Waymo, 2024; Pony AI Inc., 2025a). By September 2026, Waymo reported more than half a million paid trips per week on a fleet of over 4,000 vehicles in 15 U.S. cities (TechCrunch, 2026b,c; Waymo, 2026a), Baidu’s Apollo Go ran fully driverless on Uber in Dubai with more than 240 million driverless kilometers behind it (CnEVPPost, 2026; Baidu, Inc., 2026), WeRide operated roughly 400 driverless vehicles in the United Arab Emirates (Gasgoo, 2026), and Pony.ai reported per-vehicle unit-economics breakeven for its seventh-generation vehicle and committed, with Uber, to 2,000 robotaxis in Europe (Pony AI Inc., 2025b; TechTimes, 2026). Zoox operates a public service in Las Vegas and has expanded limited rider access in San Francisco (Zoox, 2025, 2026); in August 2026 it began charging fares under the first federal exemption granted to a purpose-built robotaxi without a steering wheel or pedals (TechCrunch, 2026f,g). Tesla began a small number of rides without an in-car safety monitor in Austin in January 2026 and claimed one million unsupervised miles by September, within small daytime geofences and with remote supervision (Electrek, 2026a,b). Wayve and several automotive manufacturers are pursuing a complementary model centered on autonomy software that can be deployed across vehicle platforms and markets (Wayve, 2026c). Driverless highway freight follows the same pattern in a tighter operational domain: Aurora passed 500,000 driverless miles in Texas and targets 200 trucks by the end of 2026 (Aurora Innovation, Inc., 2026b,a), while Volvo and Waabi began hauling freight between Dallas and Houston with an observer still on board (WBAP, 2026).

Autonomous driving has reached broader deployment than general-purpose robotics partly because its embodiment and operating environment are comparatively constrained. Vehicles travel primarily along road networks governed by established traffic rules, infrastructure, and social conventions. The sector has also benefited from large-scale driving datasets (Sun et al., 2020; Ettinger et al., 2021), mature automotive supply chains, and clear applications in passenger transportation and logistics. However, current systems operate within circumstances under which they are designed to function, such as geographic areas, road types, and weather conditions (National Highway Traffic Safety Administration, 2026). Remedies for failures at that boundary are, in practice, contractions of the boundary: after its vehicles passed stopped school buses in Austin, Waymo issued a software recall, and after an empty robotaxi was swept away on a flooded road, it recalled 3,791 vehicles and restricted operation at times and places of elevated risk (NPR, 2025; TechCrunch, 2026d). Regulators are codifying the same model. The first global framework for automated driving systems, UN GTR No. 26, requires safety at least equivalent to a competent and careful human driver and prohibits self-modifying software (Bennett Jones LLP, 2026); China’s first Level 3 approvals specify road type and speed limits explicitly (Gasgoo, 2025); and NHTSA’s 2026 framework pairs streamlined exemptions with enhanced oversight and a national benchmark effort (Hunton Andrews Kurth LLP, 2026). Hence, their proactivity is substantial, but bounded.

The next frontier is broader and more transferable *driving agency*. Existing systems already exhibit bounded proactivity by anticipating other road users, choosing appropriate pick-up and drop-off locations, adapting routes to disruptions, and positioning vehicles without continuous human instruction. Future systems must generalize these capabilities across more cities, weather conditions, traffic cultures, and vehicle platforms without extensive location-specific engineering. That transfer is under way but remains a staged, multi-year process: Waymo is running specialist-supervised trials across 20 London boroughs toward a 2026 launch and targets Munich toward the end of 2027 and Singapore in 2028, while Wayve and Uber began supervised autonomous rides in London in September 2026 with Nissan robotaxis planned for Tokyo (Waymo, 2026c; TechCrunch, 2026e; Uber Technologies, Inc., 2026; Wayve, 2026b). On the model side, reasoning-centric vision-language-action drivers make anticipation explicit and language-grounded (Hu et al., 2025b; Wang et al., 2025i; Huang et al., 2026e; Yan et al., 2026c), world models serve not only to generate rare scenarios but as closed-loop evaluators with the policy in the loop (Waymo, 2026d; Wayve, 2026a; Basant et al., 2026), and efficient world-action models internalize prediction as a training signal while keeping action real-time (Wang et al., 2026c; Fan et al., 2026b). Proactivity also extends beyond the individual vehicle: at the fleet level, agents could forecast demand, reposition vehicles, schedule charging and maintenance, and coordinate responses to network-wide disruptions.

Fleet-level proactivity remains the least developed strand: recent work on rebalancing under demand drift is an exception (Han et al., 2026b), and in current deployments fleet agenda-setting is largely delegated to ride-hailing platforms rather than to the driving agent (Baidu, Inc., 2026; TechTimes, 2026).

The remaining challenge is to scale safety and economic viability together. Autonomous vehicles must handle rare and unfamiliar situations, provide credible evidence of safety to regulators and the public, and reduce the costs of vehicles, sensors, maintenance, insurance, and operational support. Because driving decisions are safety-critical and often difficult to reverse (Bellan, 2024; Koopman, 2024), the vehicle must remain appropriately conservative when the intentions of other road users or the environment state are uncertain. Safety evidence is now reported at scale, though largely by the companies themselves: across 270 million autonomous miles, Waymo reports 82% fewer injury-causing crashes and 95% fewer serious-injury crashes than a human benchmark (Waymo, 2026b). Synthesized safety-critical scenarios (Kim & Choi, 2026), pedestrian intent prediction (Ashayer et al., 2026), and external intent communication (Tran et al., 2026) address the long tail and the legibility of the vehicle’s own intentions. Fallback behavior has a regulatory price as well: California now fines operators whose vehicles block first responders and requires licensed remote drivers (TechCrunch, 2026a).

Autonomous driving offers a useful preview of how other proactive embodied agents may mature: begin within a restricted operational domain, deploy at meaningful scale, learn from real-world experience, maintain robust fallback and human-support mechanisms, and expand both autonomy and proactive initiative only as safety and reliability are demonstrated.

6.4.3 Industrial Collaboration: From Task Automation to Proactive Co-Workers

A proactive industrial agent does more than execute a task assigned by a worker or production-management system. It monitors the shared workflow, anticipates what people and machines will need next, initiates appropriate assistance, communicates its intentions, and escalates when uncertainty or risk exceeds its authority. The focus here is on embodied agents that work alongside people, reduce physical workload, and improve the coordination of shared tasks.

Industrial robots have operated for decades in structured work cells separated from human workers, where they achieve high efficiency on repetitive and carefully specified tasks (Nof, 1999; Dzedzickis et al., 2022), and that paradigm is still growing: the global operational stock of industrial robots reached five million units in 2025 (International Federation of Robotics, 2026). The commercial frontier is now shifting toward flexible mobile manipulators and humanoid platforms that can operate in human-designed facilities and support a wider range of workflows. Figure and RobotEra, for example, are developing and piloting humanoid systems for material handling, kitting, logistics, and other labor-intensive operations (BMW Group, 2024; Figure AI, 2026a; Futurism, 2026). The evidence base has begun to shift from demonstrations to operating records. BMW’s ten-month Figure 02 pilot at Spartanburg accumulated roughly 1,250 operating hours and placed more than 90,000 sheet-metal parts into welding fixtures before it was succeeded by a parts-sequencing task and a second vendor pilot in Leipzig (BMW Group, 2026). Agility Robotics reports more than 65,000 hours of commercial Digit operation across nine customer facilities (Agility Robotics, 2025, 2026), Apptроник’s Apollo operates in designated zones at Mercedes-Benz, GXO, and Jabil (Malayil, 2026), and Boston Dynamics and Google DeepMind are pairing Gemini Robotics models with the electric Atlas, with Hyundai targeting parts sequencing in 2028 and component assembly by 2030 (Korosec, 2026; Demaitre, 2026). Analysts count roughly 19,000 to 25,000 humanoids shipped in the first half of 2026, more than 70% of them to industrial and commercial buyers and most of them from AgiBot and Unitree (Smart Analytics Global, 2026; Qi & Wang, 2026; Francis, 2026; KrASIA, 2026). Non-humanoid mobile manipulators are reaching production in logistics first, as in Dexterity’s autonomous trailer loading at FedEx and UPS’s order of 400 Pickle Robot truck unloaders (Dexterity, 2026; Soper & LaPara, 2025). In a widely viewed 2026 comparison, a Figure humanoid processed roughly 200 fewer packages than a human participant during a ten-hour parcel-sorting task, demonstrating near-human performance in this narrowly defined workflow (Bishop, 2026; Nield, 2026), and a later company-run livestream showed three Figure 03 robots sorting more than 28,000 packages over 24 hours without reported interventions (Knutsson, 2026). These demonstrations reflect substantial progress in perception, manipulation, and robustness, but they should not be confused with proactive agency. A robot may operate autonomously for hours on a parcel-sorting line while still following a predefined objective, workspace, and action pattern. Every deployment above is a fixed, fenced task: totes, trolleys, parcels, or trailers. The vendors themselves describe designated zones and multi-year roadmaps, and Tesla’s own investor filings state that the initial Optimus builds are routed to an internal training academy rather than to production work (Tesla, Inc., 2026). Most industrial humanoid pilots therefore remain closer to task-level *reactive autonomy*: they can execute assigned work

independently, but fail to determine how best to contribute to the broader workflow.

Proactivity begins when the agent reasons about the state and needs of the overall operation. Rather than waiting for an operator to request each action, it might infer which component will be needed next, prepare tools before a worker reaches the corresponding step, reorder tasks when a workstation falls behind, or ask for clarification when several actions appear plausible. Recent research has explored proactive task sequencing through real-time prediction of human motion and intent (Huang & Mutlu, 2016; Liu et al., 2023b; Yu et al., 2024; Abilkasov et al., 2025), adaptive behavior tailored to individual workers (Liu et al., 2022b; Chen et al., 2023b; Freire et al., 2024; Hostettler et al., 2024), and anticipatory cues that make upcoming robot actions easier for people to understand (Psarakis et al., 2025; Bo et al., 2026). Work in 2026 has moved from predicting human intent to deciding whether and when to act on it. Examples include learned ask-or-act policies that weigh clarification against assistance accuracy over multi-day interaction (He et al., 2026a), event-driven assistance that acts only on observed human–object state transitions and refrains when a task is unsolvable (Liu et al., 2026a), collaborative vision-language-action policies whose action chunks straddle task transitions and thereby trigger premature assistance unless steered at inference time (Xu et al., 2026d), and uncertainty-aware intention prediction for assembly (Heman et al., 2026). At the workflow level, fatigue-predictive task allocation decides when to assign which task to whom (Xue et al., 2026), and LLM-based orchestrators plan and replan for heterogeneous robot fleets over a shared world state (Gupta et al., 2025; Zhang, 2026b), although a single accepted false claim about that state can cascade into fleet-wide replanning (Khalid & Min, 2026). The first commercially deployed workflow-level reasoning is advisory rather than embodied: Amazon’s Project Eluna reasons over live operational data to recommend where managers should shift people to avoid bottlenecks (Amazon, 2025). These systems provide clearer examples of bounded proactive collaboration, although most remain research prototypes or narrowly scoped pilots.

The central challenge is the reliability of proactive initiative. An unnecessary or poorly timed action may interrupt a worker, slow production, damage equipment, or create a safety hazard. The agent must therefore do more than predict what might happen next: it must judge whether acting now would be helpful, whether its intention will be understood, and whether the action can be stopped or reversed safely. A controlled study of humanoid assistance in group collaboration found that unprompted intervention raised interaction load and that its value depended on the participants’ prior experience and personality (Vitry et al., 2026). For legged and balancing robots, even stopping is an active behavior: cutting power makes a biped fall, so classical functional-safety chains can be certified only around, not inside, the balancing controller, a gap that the forthcoming ISO 25785-1 standard for actively stabilized mobile robots is meant to address (Ding et al., 2026a; i-SCOOP, 2026). Newer platforms accordingly build in graded slow-to-stop-to-sit responses, independent safety controllers, and explicit visual and auditory intent cues (Wessling, 2026). Industrial users will generally prefer a system that performs a limited set of tasks with consistently high reliability and efficiency over one that attempts a broader range of tasks unpredictably. Amazon’s withdrawal of its Blue Jay multi-arm cell within four months of announcing it, and a Siemens case study documenting the collect–curate–fine-tune cycles needed to make a generalist policy reliable at a single packaging station, illustrate the point (Amazon, 2025; Zhu et al., 2026a). Proactive industrial agents are therefore likely to emerge through gradually expanding islands of validated autonomy. Task-level performance may be approaching that of human workers in some narrow workflows, but reliable workflow-level proactivity remains an open frontier.

6.4.4 Human-Agent Co-Creation: From Virtual Imagination to Physical Creativity

A proactive co-creative embodied agent does more than fabricate a finalized human design. It participates in shaping the artifact by interpreting incomplete intent, proposing alternatives, evaluating feasibility, selecting tools and materials, and adapting the design during production. In this setting, proactivity means making timely and constructive contributions while preserving human authorship and control. This section focuses on humans and embodied agents collaboratively creating physical artifacts, while broader forms of digital co-creation, including games and media, are discussed in §6.12.

Generative AI has greatly expanded digital creation in images, videos, and virtual 3D assets (Zhu et al., 2017; Rombach et al., 2022; OpenAI, 2024; DeepMind, 2025; Tochilkin et al., 2024; Zhao et al., 2025e). The next frontier is to connect this generative capability with robotic systems that can produce physical outcomes. Early commercial platforms are already building parts of this pipeline. Machina Labs develops software-defined factories that translate digital designs into complex metal structures, while AUAR combines robotic microfactories with software that converts building plans into prefabricated components (Machina Labs, 2026b; AUAR, 2026). Machina Labs has since raised capital for a 50-cell facility in which forming robots decide incrementally where to work and how much to deform, guided by in-cell

scanning (Machina Labs, 2026a; SAE Media Group, 2026); Divergent couples generative design, large-format metal printing, and robotic assembly in one production system (TCT Magazine, 2026); and process-planning software now derives work instructions and assembly sequences from CAD so that engineers approve rather than author them (Dirac, 2026). Text-to-CAD has likewise become conversational and self-checking, with commercial agents that anticipate edits and validate geometry before committing them (Zoo, 2026; Backflip AI, 2026). These systems are primarily automated fabrication platforms rather than co-creative agents, but they provide important infrastructure for future agentic design-to-production workflows. Research prototypes demonstrate complementary elements of this vision. FRIDA and CoFRIDA translate multimodal creative intent into paintings and update their plans using feedback from the evolving canvas, with CoFRIDA supporting human-robot co-painting (Schaldenbrand et al., 2022, 2024), and CoBrush extends co-painting to multi-round sessions by separating intent inference, spatial grounding, and stroke execution (Qin et al., 2026). Other systems have demonstrated robotic pottery and sculpting, including closed-loop visual planning for clay reliefs of more than a hundred actions (Yoo et al., 2024; Ma et al., 2020; Schaldenbrand & Oh, 2026), as well as the generation and robotic assembly of physical structures from sketches (Xu et al., 2025f), text descriptions (Goldberg et al., 2025; Pun et al., 2025; Liu et al., 2026k; Kyaw et al., 2025a; Sznaiier Camps et al., 2026), or speech (Kyaw et al., 2025b). Multi-robot execution frameworks further increase the scale and complexity of the artifacts that robotic systems can construct (Huang et al., 2025a; Liu et al., 2026k). Agentic design-and-fabrication systems are beginning to combine these pieces: MAVE places LLM supervisor, design, and robot agents, an augmented-reality headset, and a cobot in one interactive workflow (Yang et al., 2026f), and vision-language agents that align observed assembly state with the design specification replan when human co-workers deviate from the expected sequence (Chen & Deng, 2026). Together, these efforts point toward workflows in which design reasoning, feasibility analysis, production planning, and physical execution are no longer treated as separate stages.

The central frontier is *closed-loop prompt-to-product*. A user may begin with an incomplete description, rough sketch, reference image, or partial design. A proactive agent could generate alternatives, identify ambiguities, evaluate whether each proposal can be produced with the available tools and materials, and recommend modifications before production begins. During production, it could inspect intermediate results, detect deviations, revise the design or production plan, and ask the user to choose among emerging alternatives. The defining capability is knowing when to propose, clarify, revise, or defer while balancing creative intent, physical feasibility, user preferences, and resource constraints. The individual pieces of this loop now exist. Multi-agent CAD generators wrap design in programmatic geometric validation, review, and rollback (Barkley et al., 2026; Shui et al., 2026; Hu et al., 2026b; Haag et al., 2026), and new benchmarks score manufacturability and assemblability rather than shape similarity (Dong et al., 2026a; Wang et al., 2026d). Fabrication-aware tools let users repair generative 3D models that would fail when printed (Faruqi et al., 2026), stage the assistant’s initiative from decomposition to structural recommendation before cutting (Gao et al., 2026a), optimize generated shapes for printability (Wu et al., 2025a), or improvise assemblies from the materials at hand (Isume et al., 2025). During production, language-model agents screen toolpaths for anomalies before printing and monitor and correct additive processes in situ (Eslaminia et al., 2026; Alali et al., 2026; Pak et al., 2026). What is still missing is a single system that composes alternative generation, feasibility scoring, in-production inspection, and user-in-the-loop refinement under one policy for when to propose, verify, or ask (Wang et al., 2026h).

Co-creation offers a relatively permissive setting for proactive experimentation because creative goals are often subjective and evolve throughout the process. In tightly controlled or safety-critical settings, an unexpected robot action is usually a liability. However, in creative work, it may become a useful suggestion or a productive change in direction. Therefore, embodied agents can take greater initiative, provided that their contributions remain transparent, reversible, and aligned with the creator’s intent.

The commercial opportunity lies in flexible design-to-production systems for personalized products and low-volume manufacturing. Such systems could reduce dependence on specialized design expertise, dedicated tooling, and large production runs. However, important barriers remain, including translating subjective intent into engineering specifications, verifying structural and functional feasibility, maintaining quality over long production sequences, and managing material and equipment constraints. Questions of authorship, intellectual-property ownership, provenance, and accountability will also become more pressing as humans and agents jointly design and fabricate physical products. Early evidence is emerging on both fronts: professional artists found collaboration with an autonomous painting arm more playful and reflective than with a human partner and granted it more autonomy (Cocchella et al., 2025), expressive robot behaviors shift perception from tool to partner (Yu et al., 2026c), and patch-level attribution can distinguish human, robot, and genuinely blended regions of a collaborative painting (Chen & Alves-Oliveira, 2026).

6.4.5 Evaluation, Challenges, and Open Directions

Robotics is adopting proactive behavior more cautiously than many digital domains because embodied agents act directly on people, objects, and infrastructure. Their mistakes can cause physical harm, damage property, or alter the environment in ways that are difficult to reverse. Proactive embodied agents also share broader concerns with other proactive systems, including privacy, trustworthiness, ethics, and alignment, as discussed in §7. Beyond these general concerns, physical embodiment introduces several distinct challenges.

First, evaluating proactive embodied intelligence remains difficult, and no single standard captures all of its relevant dimensions. As with proactive agents in other domains, task success alone is insufficient, as discussed in §5. A proactive robot must also recognize when intervention is useful, act at an appropriate time, communicate its intentions, and avoid unnecessary or disruptive behavior. Thus, evaluation should consider both task performance and the quality of human-agent interaction, including *interaction fluency*, *legibility*, *perceived safety*, *user agency*, and overall *user satisfaction* (Dragan et al., 2013; Hoffman, 2019; Cohen et al., 2026). It should also capture failures specific to proactivity, such as missed opportunities, unnecessary interventions, excessive clarification, poorly timed actions, and inappropriate escalation. Such metrics are beginning to appear. Benchmarks now score the decision rather than only the outcome: ProActBench and RobotEQ include negative traps in which the correct behavior is to stay silent (Zhang et al., 2026l; Fang et al., 2026b; Che et al., 2026), NIABench penalizes interrupting the human’s plan (Zhang et al., 2026i), ProAct-75 reports trigger-detection accuracy and steps saved (Zhu et al., 2026d), EgoServe and HOTD-Bench treat intervention as a context-dependent choice over continuous video (Gong et al., 2026; Song et al., 2025b), and NormAct scores compliance with unspoken household norms (Zhao et al., 2026c). Across these benchmarks, current vision-language models remain far below humans and are brittle to prompt framing; in one failure-recovery benchmark, refusal rates swing from near-total to near-zero when answer options are reordered (Pathak & Krishna, 2026). For task competence itself, evaluation is shifting from single success rates to capability profiles, robustness audits, and reproducible real-robot leaderboards (Gao et al., 2026b; Fei et al., 2025; Atreya et al., 2025; Yakefu et al., 2025; Chen et al., 2026e).

Many of these qualities cannot be measured adequately using static datasets or isolated task-success metrics. They often require human-in-the-loop studies and longitudinal deployments because the appropriate level of initiative depends on the user, environment, and application. Offline evaluation against static human models can overstate proactive performance so much that methods which look strong offline add net work once the human adapts to the robot (Patel & Chernova, 2026; Tang et al., 2026a), and user studies continue to find that more frequent intervention does not raise engagement (Hriscu et al., 2026; Vitry et al., 2026). Longitudinal in-home studies remain small, typically a handful of homes over several weeks (Moorman et al., 2026; Kian et al., 2026), while distributed double-blind pairwise comparison offers one scalable template for real-world evaluation (Atreya et al., 2025). The same behavior may be helpful in one context but distracting, intrusive, or even unsafe in another. Evaluation is further complicated by the diversity of robotic embodiments, including humanoids, mobile manipulators, quadrupeds, drones, mobile robots, and fixed-base arms, each with different capabilities, risks, and interaction patterns. Real-world experiments are also costly and time-consuming, particularly when they require specialized hardware, controlled safety procedures, and repeated studies with human participants.

Second, proactive embodied agents must adapt quickly to new embodiments, environments, tasks, and users. A household robot entering a new home must learn its layout, object conventions, resident preferences, and household boundaries. An industrial robot may need to adjust to a new workstation, tool configuration, workflow, or human partner. Differences in sensors, kinematics, payloads, and control interfaces also limit the direct transfer of policies across platforms. Unlike digital agents, which can learn about users through virtual interactions such as text exchanges, embodied agents typically require physical interaction and real-world execution. This makes adaptation and generalization slower, more expensive, and potentially riskier.

Third, proactive robots must take initiative safely under physical uncertainty. Actions in the physical world may be difficult to reverse and can alter the environment before a person has time to intervene. An embodied agent must therefore reason not only about whether an action is likely to succeed, but also about its possible consequences, whether nearby people will understand its intention, and whether the action can be interrupted or recovered from safely. Hence, calibrated uncertainty, legible motion, permission-aware planning, and reliable fallback behavior are central to proactive embodiment. Red-teaming studies show why these cannot be left to the policy alone: unguarded vision-language-action models can be driven to unsafe actions in nearly all tested scenarios, and fine-tuning only on safe demonstrations can even increase hazardous execution (Zhang et al., 2026j; Chen et al., 2026i; Yin et al., 2026, 2024b; Li et al., 2026f). The field is therefore converging on decoupled arbitration layers: formal guardrails that translate safety rules into

temporal-logic specifications and veto non-compliant plans (Ravichandran et al., 2026), compact hazard-judging models that score observation–action pairs (Choi et al., 2026b), reachability-based safety values that gate policy optimization (Tayal & Nambi, 2026), calibrated success confidence for ask, act, or abort decisions (Francis-Meretzki et al., 2026), and layered guardrail architectures with explicit conservatism allocation (Kim et al., 2026c). Governance lags embodiment: the only humanoid-specific safety standard in progress, ISO/CD 25785-1, covers industrial settings only, so home robots still fall under the 2014 personal-care standard; the EU has deferred product-embedded high-risk AI obligations to 2028; and legal analyses on both sides of the Atlantic find no fit-for-purpose privacy rulebook for remote teleoperation of home robots (International Organization for Standardization, 2026; EU AI Act Resource, 2026; Da Costa Alves & Le Goff, 2026; Stewart, 2026).

Finally, open-world operation requires robots to continue improving after deployment, yet physical systems cannot explore freely when mistakes may cause harm. Progress will depend on learning mechanisms that combine simulation, demonstrations, operator interventions, and carefully collected real-world experience without compromising safety. Continual learning systems must also detect when an update has degraded performance, prevent unsafe policies from being deployed, and support reliable rollback or human takeover. Recent proposals place a world-model verification gate inside a human-oversight envelope around post-deployment self-improvement and argue for evaluating the evolving trajectory rather than a fixed checkpoint (Nguyen et al., 2026), while real-robot continual-learning benchmarks show that naive sequential fine-tuning forgets earlier skills unless experience is replayed (Zhu et al., 2026b).

As a result, progress in proactive robotics will depend not only on more capable models, but also on better mechanisms for calibrated initiative, rapid adaptation, permission-aware decision-making, safe continual learning, human oversight, and reliable escalation. Ultimately, a general-purpose proactive embodied agent must do more than act autonomously. It must judge when intervention is useful, when clarification is needed, and when it should wait or remain still. Only systems that combine physical capability with this form of operational judgment can become safe, trustworthy, and genuinely helpful partners in the open world.

6.5 Agents for Security and Surveillance

Table 7 Representative proactive security systems. Res. denotes research systems; Com. denotes commercial products.

System	Awareness	Anticipation	Agenda	Arbitration	Action	Adaptation	Deploy.	Type
Big Sleep (Big Sleep team, 2024)	Selected commits, repository code	LLM variant analysis	Given commits; variant search	Crash reproduction	Finds and reports variant bugs (<i>suggestion</i>)	Debugger feedback	Software	Res.
ATLANTIS (Kim et al., 2025c)	Challenge code, analysis and test results	LLM + program analysis	Given challenge projects	Executable validation	Finds, tests, patches (<i>action</i>)	Patch-evaluation feedback	Software	Res.
Claude Security (Anthropic, 2026d)	Selected repository or directory	LLM data-flow tracing	User-scoped or scheduled scans	Adversarial verification + review gate	Scans, verifies, proposes patches (<i>gated action</i>)	Dismissals carried forward	Software	Com.
PAWS (Fang et al., 2016; Xu et al., 2020)	Patrol and poaching records; topography	ML risk prediction; poacher-behavior model	Given protection mandate; high-risk areas	Patrol limits; planner review	Suggests patrol routes with probabilities (<i>suggestion</i>)	Learns model parameters from data	Software	Res.

In security and surveillance, proactiveness concerns whether the human or the agent decides which protective tasks warrant attention and how they should be pursued. An agent may exercise task-scoped initiative while analyzing an alert or patching a reported vulnerability, even when a human selects the task. The agent takes initiative when, within a standing protective objective and an authorized scope, it decides what to investigate next, whether a finding warrants intervention, and when to revisit an earlier decision. In surveillance, these decisions also include determining whether an ambiguous event warrants acquiring further evidence.

Existing work supplies foundations for these proactive choices. Red teaming and threat hunting organize inquiry around possible weaknesses or overlooked intrusions. Agent-based testing and repair automate diagnosis and intervention, while incident records preserve evidence and case history. Earlier autonomous cyber-resilience work and the AICA reference architecture envisaged agents that monitor, plan, respond, and recover in contested environments (Kott & Theron, 2022; Kott et al., 2023). Recent accounts of autonomous threat hunting and agentic cybersecurity describe increasingly broad defensive workflows (Sindiramutty, 2024; Kshetri, 2025; Vinay, 2025). These foundations support

broader initiative when an agent identifies related vulnerabilities, proposes additional protection, or recognizes that changed conditions require renewed attention. Table 7 compares selected security systems across the six mechanisms, covering vulnerability discovery and repair as well as patrol planning.

The central challenge is therefore not simply to connect investigation, repair, and monitoring into a longer workflow while leaving task selection and coordination to humans, but to enable the agent to sustain task-selection judgment across that workflow. A proactive agent must repeatedly determine whether further investigation work is warranted, what defensive form it should take, and how far the protection should extend, while remaining within authorization and human oversight. This section thus examines that capability through initiative in choosing investigations, turning findings into protective action, and proactively revisiting prior decisions, before considering the associated risks, evaluation challenges, and open research directions.

6.5.1 Choosing investigations and acquiring evidence

A human-defined project or incoming alert can provide a starting point for security agents to work on without prescribing every subsequent investigation. The distinction between a relatively reactive agent and a proactive one is whether the agent merely follows an assigned investigation workflow or takes the initiative to use available evidence to formulate an additional question worth pursuing. It is noteworthy that choosing a tool within a prescribed task is not, by itself, evidence of task initiation and proactiveness.

Red teaming and threat hunting frame investigation around hypotheses about weaknesses or intrusions worth examining, providing the forward-looking foundation for security agent proactiveness. Interviews with professional threat hunters highlight the judgment and organizational context this inquiry requires (Maxam & Davis, 2024). A permission change, for example, may create a path from a compromised account to a sensitive service. Those lines of work form *anticipation* (§4.2) that connects assets, permissions, and possible attacker actions to a testable hypothesis about how protection could fail.

Product-level testing, organization-level threat hunting, and ecosystem-level vulnerability discovery offer different settings for examining the scope of proactiveness. PentestGPT coordinates tool use, result interpretation, and subsequent testing within supplied penetration-testing targets (Deng et al., 2024a). Google’s SecOps Triage and Investigation Agent gathers evidence and correlates signals around incoming alerts (Google Cloud, 2026). Project Glasswing extends model-assisted vulnerability discovery across participating organizations (Anthropic, 2026g). These systems illustrate different forms of automated investigation, but broader initiative also requires deciding which additional evidence is worth pursuing. A useful follow-up check seeks evidence that can distinguish between competing explanations requiring different responses. For unfamiliar server access, maintenance records, executed processes, and subsequent data access can help distinguish authorized work from credential abuse. Missing records may justify finer monitoring (Kapoor & Kas, 2026), while missing organizational context may call for asking the service owner. Anthropic’s discussion of trustworthy agents likewise motivates clarification when uncertainty concerns intent that only a person can resolve (Anthropic, 2026i). A benign explanation supports closure when it accounts for the relevant observations without leaving conflicting evidence unresolved. The judgments made during the *arbitration* (§4.4) stage therefore weigh potential loss, the value of further evidence, disruption, and review cost when choosing whether to continue, clarify, or return the case to monitoring.

Particularly, in physical and video monitoring, the next investigative decision supporting proactive intervention may concern what additional evidence is needed to resolve an ambiguous event. Earlier decision-support work such as the Warfighter Associate addressed the management of intelligence, surveillance, and reconnaissance assets (Buchler et al., 2014). For anti-poaching patrols in protected areas, PAWS predicts where poaching is likely from past patrol and poaching records and suggests patrol routes that planners review before rangers patrol (Fang et al., 2016; Xu et al., 2020). Someone repeatedly approaching a restricted entrance could be testing access or completing a delivery. Footage from another authorized camera or an earlier time interval may distinguish these explanations, but replaying existing footage cannot recover observations that were never recorded. Video-analysis methods supply components for selective inspection. Holmes-VAU combines descriptions at multiple temporal scales with adaptive frame sampling, while QVAD refines questions as visual context becomes available (Zhang et al., 2025d; Bekit et al., 2026). AgenticVAU separates search planning, observation, and judgment, recording whether observations support, contradict, or leave a candidate event unresolved (Duan et al., 2026b). Additional observation also expands the scrutiny imposed on people. Inspecting a short authorized interval, tracking someone across locations, and retaining movement histories have different privacy

implications (Zhang & Maharjan, 2026). Purpose and permitted scope should constrain collection before it occurs, while a supported benign explanation should help end unnecessary scrutiny.

6.5.2 Turning findings into protective action

A finding from the previous investigation stage may require a decision about whether to gather additional evidence, propose a protective measure, or defer action. In typical workflows, this initiative is exercised by a human, who determines whether follow-up work is warranted and defines its scope. For an agent to exhibit proactiveness, it must take on this role as well: recognizing when further action is needed and deciding what that action should involve.

Existing work has automated substantial parts of agentic repair and response workflows, enabling agents to diagnose problems, propose fixes, and carry out prescribed actions (Pearce et al., 2023; Yang et al., 2024a). These systems, however, typically act on tasks initiated by a human rather than deciding whether further intervention is warranted. DARPA’s AI Cyber Challenge shows how far this automation has reached: in the final competition, autonomous cyber reasoning systems discovered 54 of the synthetic vulnerabilities planted in the challenge projects and patched 43 of them, and also found 18 real vulnerabilities that were disclosed to maintainers (Defense Advanced Research Projects Agency, 2025). The winning system, ATLANTIS, combines LLMs with static analysis, symbolic execution, and fuzzing to discover and patch vulnerabilities (Kim et al., 2025c). On the response side, security orchestration platforms execute predefined playbooks that coordinate tools such as SIEMs, firewalls, and ticketing systems once an alert meets their trigger conditions (Islam et al., 2019). In these settings, the broader objective and intervention scope are defined upstream: challenge organizers select the target projects and evaluation criteria, while engineers specify the conditions that trigger playbooks.

Stronger initiative arises when a security agent must decide what, if anything, should happen next. In software repair, such initiative may involve deciding to search for related vulnerabilities or extend a fix beyond the reported bug; in surveillance, it may involve deciding whether to gather more evidence, issue a notification, escalate for review, or take no further action. A confirmed bug is often one instance of a broader flaw (Wang et al., 2025m). Project Zero found that at least nine of the eighteen zero-days exploited in the wild in the first half of 2022 were variants of previously patched vulnerabilities, and that at least half could have been prevented by more comprehensive patching (Stone, 2022). Searching for such variants is a task the original report does not specify: the agent must abstract the root cause into a pattern, decide where else that pattern could occur (Jang et al., 2012; Kim et al., 2017), and judge whether the instances it finds call for individual patches or a structural change (Popa & Flynn, 2025). Big Sleep illustrates the search itself. Given the message and diff of a recent SQLite commit, the agent looked for related issues in the current code and found an exploitable stack buffer underflow, which was fixed before it reached an official release (Big Sleep team, 2024). The commits used to initiate these searches, however, were selected by researchers, so independently deciding which changes warrant variant analysis remains an additional task-selection capability. In surveillance, the corresponding choice turns on two questions: whether further evidence could change the decision, and what harm delay would allow (Howard, 1966). Choosing not to act is thus part of the same initiative, and should be recorded together with the evidence that supports it.

Taking initiative does not require giving the agent final authority over consequential actions. An agent may independently identify issues, investigate related cases, and propose repairs or other interventions, while deployment or other high-impact actions remain subject to human approval. A mistaken investigation or proposal consumes effort but leaves the protected system unchanged, whereas a mistaken patch, account suspension, or network block can disrupt the operations it was meant to protect. Protective changes are not exempt from this risk: in July 2024, a faulty update to CrowdStrike’s security software affected an estimated 8.5 million Windows devices (Weston, 2024; CrowdStrike, 2024). An agent can therefore be given wide latitude in deciding what to investigate and propose, while actions that are hard to reverse or that affect many systems remain subject to approval (Shavit et al., 2023; Anthropic, 2026i,d). CodeMender follows this division: it generates and validates patches automatically, but all of them are currently reviewed by human researchers before being submitted upstream (Popa & Flynn, 2025).

6.5.3 Sustaining protection and revisiting decisions

Proactiveness may also require an agent to revisit earlier decisions when new conditions change their basis. Changes in permissions, dependencies, or threat knowledge may make a previously resolved issue relevant again. In surveillance, later evidence may similarly alter the basis for resolving an event or ending further observation. A vulnerability may be

deferred because no exploit is known, because the affected service is reachable only by trusted accounts, or because the deployed version of a dependency does not include the vulnerable code (Pashchenko et al., 2018); a surveillance event may be closed because the available footage supports a benign explanation. Each of these conditions can change after the decision is recorded, sometimes quickly (Bilge & Dumitras, 2012). Mandiant’s analysis of vulnerabilities exploited in 2023 found that the average time-to-exploit fell from 32 days in 2021–2022 to five days, after excluding statistical outliers (Charrier & Weiner, 2024). A deferral that was reasonable while no exploit was known may therefore become unreasonable within days. Likewise, if a person whose earlier presence was attributed to a delivery is later observed attempting unauthorized access at another restricted entrance, the new evidence may warrant reopening the original event, whereas unrelated activity elsewhere would not.

Existing incident records and case histories provide the memory needed to support such reassessment. Claude Security, a commercial code-security product, supports scheduled scans for continued coverage, and user-recorded dismissals persist so earlier triage decisions remain visible across the team (Anthropic, 2026d). Proactive reopening, however, requires more than retrieving a prior record: the agent must relate a new condition to an earlier decision and determine whether that decision should be reconsidered. Incident-response guidance calls for recording the actions performed during an investigation and preserving the integrity and provenance of those records (Nelson et al., 2025). Such records show what was decided and when, but may not explain the assumptions supporting that decision. A subsequent change that invalidates those assumptions may also be recorded elsewhere without referring to the original case. A firewall change that exposes a previously internal service, for example, is recorded in change management rather than in the vulnerability ticket whose deferral assumed the service was internal (Johnson et al., 2011). Searching past cases by host or vulnerability identifier may retrieve that ticket, but retrieval alone does not establish that the firewall change has invalidated the assumption underlying the deferral. The agent must therefore reconstruct the basis of the earlier decision, test whether the new condition contradicts it, and decide whether the case warrants a targeted recheck, renewed intervention, or no change. Explicitly recording the conditions underlying each decision makes such reassessment more tractable.

Proactiveness over longer time horizons also raises a prioritization problem when new investigations, repairs, and rechecks compete for limited computation and human review. The agent must therefore judge not only whether a task is worth pursuing, but also whether it deserves priority over pending investigations, repairs, and rechecks. Because intrusions are rare relative to benign activity, even accurate detectors produce mostly false alarms (Axelsson, 2000), and SOC analysts report that manually validating them is tedious and can lead to burnout and desensitization (Alahmadi et al., 2022). A proactive agent can add to this burden, because each investigation it opens, repair it proposes, or recheck it schedules may eventually require validation or approval. Open-source maintainers have already experienced the effect: the curl project reports that plausible but unfounded AI-generated vulnerability reports take longer to refute the more convincing they appear, consuming time that would otherwise go to fixing real bugs (Stenberg, 2024, 2025). Whether a new task should displace pending work depends on the harm it could prevent, how soon that harm could materialize, and how likely the task is to yield protection (Jacobs et al., 2020, 2021; Nelson et al., 2025). A validated repair awaiting deployment may deserve priority over a speculative search. However, if a change invalidates an assumption underlying an earlier security decision, reassessing that decision may be more urgent than either task.

6.5.4 The agent as a target and source of risk

Proactiveness creates a new attack and safety surface at the level of task initiation itself: an agent can be induced, mistaken, over-scoped, or insufficiently controllable in deciding what work to start or continue. An adversary can induce unwanted task initiation by manipulating the evidence an agent uses to decide what to investigate or act upon. Indirect prompt injection shows the mechanism: instructions embedded in retrieved content can redirect an LLM-integrated application, because such applications blur the line between data and instructions (Greshake et al., 2023; Perez & Ribeiro, 2022; Liu et al., 2024g; Zhan et al., 2024; Debenedetti et al., 2024). The agent may also be mistaken without any adversary, inferring a threat from coincidental evidence and initiating work that consumes review capacity or disrupts operation (Axelsson, 2000; Ruan et al., 2024; Wang et al., 2026l). It becomes over-scoped when a legitimate line of inquiry leads beyond its authorization, as when tracing reused credentials reaches a third-party service (Lynch et al., 2025). The UK AI Security Institute has reported such overreach under deliberately permissive testing conditions: in 10 of 122 cyber-evaluation runs with internet access enabled and some safety safeguards disabled, agents tasked with solving a security challenge took unsanctioned actions directed at real people and organizations (UK AI Security Institute, 2026). Finally, agent-initiated tasks may bypass oversight mechanisms designed for human-assigned work,

particularly when new investigations or actions are not explicitly logged and exposed for review (Chan et al., 2024). Mitigations must therefore constrain task initiation itself, rather than relying solely on controls applied after tasks are underway. Separating trusted control instructions from untrusted evidence, as in CaMeL, can prevent attacker-controlled content from being treated as authority for new actions (Debenedetti et al., 2025; Chen et al., 2025c; Wallace et al., 2024; Beurer-Kellner et al., 2025); requiring independent corroboration before initiating costly investigations can reduce the influence of planted evidence; and making agent-initiated tasks visible, scope-limited, and cancellable can preserve opportunities for human oversight and intervention (Odersky et al., 2026; Anthropic, 2026e; Greenblatt et al., 2024; Hadfield-Menell et al., 2017; Soares et al., 2015).

6.5.5 Evaluation, challenges, and open directions

Evaluating proactiveness requires measuring whether an agent can identify and initiate warranted protective work that has not been externally specified, rather than only how well it executes a task once that task has been assigned. Most existing benchmarks for security agents primarily evaluate execution after the relevant task or objective has already been defined, including CTF-oriented benchmarks and environments (Shao et al., 2024; Wang et al., 2025m; Engelberg et al., 2026), Cybench (Zhang et al., 2025a), and CVE-Bench (Zhu et al., 2025c). Cyber Defense Benchmark narrows this gap by asking agents to identify malicious events in raw Windows logs without guiding questions (Chona et al., 2026), although the higher-level hunting objective is still externally supplied. Evaluating initiative therefore requires environments in which potential follow-up work remains latent rather than appearing explicitly in the task description. Examples include a patched vulnerability for which related variants may remain, an alert involving credentials that may also have been used elsewhere, or a later permission change that invalidates an earlier decision not to investigate.

The task scope of proactive security agents makes evaluation substantially more difficult because whether initiative is appropriate depends on context, its protective benefit is often counterfactual, and the work it generates can itself be costly. First, whether a potential task is warranted cannot usually be determined from an isolated observation. The same failed login may deserve little attention on a test server but immediate investigation on a domain controller. Benchmarks must therefore specify enough asset, exposure, policy, and authorization context for judgments about initiative to be meaningful. Second, prevention is inherently counterfactual: after an agent patches a vulnerability, the absence of a subsequent compromise does not reveal whether the intervention prevented an attack or whether no attack would have occurred. Cyber-operations gyms such as CybORG, which place defensive agents against attacker agents in simulated networks, provide one way to estimate these effects by comparing trajectories with and without an intervention under controlled conditions (Standen et al., 2021; Emerson et al., 2024), although such estimates may not transfer directly to real deployments. Third, as discussed in §6.5.3, investigating a benign event consumes analyst attention even when it is resolved correctly (Alahmadi et al., 2022; Hassan et al., 2019).

Controlled comparisons can then separate the value of agent initiative from the value of task execution or predefined automation. One condition can assign the protective task externally, showing what the agent achieves when a human has already identified the work to perform. A second can use a fixed workflow or playbook, such as always checking related accounts after a suspicious login, to measure the benefit of predefined follow-up rules. A third can expose the agent to the same observations and objectives without specifying the follow-up, allowing it to identify, formulate, and initiate additional work on its own. The contrast between the latter two conditions is particularly important: it tests whether agent-directed initiative improves on predefined follow-up rules rather than merely demonstrating that follow-up of any kind is useful. Such comparisons require model capability, tool access, observations, authorization, and resource budgets to be held constant. An agent with broader visibility or a larger budget may discover more simply because it can search further (Kapoor et al., 2025), while an agent allowed to take actions unavailable to the baseline is not meaningfully comparable. Evaluation should jointly report protective outcomes, missed warranted opportunities, unnecessary initiated work, and human review or operational cost. Reporting these quantities together distinguishes useful initiative from two degenerate strategies: initiating nearly everything, which may improve coverage at unsustainable cost, and initiating almost nothing, which minimizes overhead by leaving valuable opportunities unexplored.

Open directions concern how such initiative should operate over longer horizons, how agents should prioritize among competing opportunities for protective work, and how initiative should remain subject to effective human oversight. Long-horizon evaluation requires environments that evolve over weeks rather than ending after a single verdict: dependencies are updated, permissions change, new advisories appear, and most changes do not warrant intervention. Such environments test the reassessment discussed in §6.5.3, namely whether new evidence justifies continuing, reopening, or terminating earlier lines of work (Anugula et al., 2025). Prioritization calls for methods to estimate the

expected value of candidate investigations and to learn from reviewers’ decisions about which proposed actions are worth pursuing (Veeramachaneni et al., 2016). Initiative also creates an oversight problem: reviewers should be able to understand why the agent proposed or initiated a task, what further actions approval would authorize, and how its scope can be narrowed or terminated (Shavit et al., 2023; Wang et al., 2026e).

6.6 Agents for AutoResearch

Table 8 Representative proactive AutoResearch systems. Res. denotes research systems; Com. denotes commercial products.

System	Awareness	Anticipation	Agenda	Arbitration	Action	Adaptation	Deploy.	Type
Coscientist (Boiko et al., 2023b)	Web / docs, code, lab state	Experiment planning	Human-specified objective	Experiment / tool selection	Plans + runs experiments (<i>action</i>)	Result-driven replanning	Robotic lab	Res.
The AI Scientist-v2 (Yamada et al., 2025)	Literature, code, experiment state	Experimental trajectories	Hypotheses + experiments	Agentic tree-search selection	Experiment → analysis → paper (<i>action</i>)	Iterative refinement	Software	Res.
AI Co-Scientist (Gottweis et al., 2026)	Research goal, literature, evidence	Hypothesis reasoning	Generates + evolves hypotheses	Debate + tournament ranking	Hypotheses + experiment proposals (<i>suggestion</i>)	Iterative refinement	Software	Res.
Virtual Lab (Swanson et al., 2025)	Research goal, literature, scientific tools	Multi-agent reasoning	Human-posed agenda	PI-led synthesis + selection	Designs pipeline + candidates (<i>suggestion</i>)	Iterative refinement	Lab-in-the-loop	Res.
Robin (Ghareeb et al., 2026)	Literature + experimental data	Hypothesis / experiment reasoning	Hypotheses + follow-ups	Candidate / experiment selection	Proposes experiments + analyzes results (<i>gated action</i>)	Results update hypotheses	Lab-in-the-loop	Res.

Among the verticals surveyed here, AutoResearch is where proactiveness is most easily confused with autonomy, and separating the two is essential for making progress. The research loop is long-horizon and loosely supervised, which makes it a natural home for self-initiated behavior. Many systems that have attracted the most attention receive their top-level research objective from a human and self-direct the execution. Reduced supervision over execution demonstrates autonomy; proactiveness depends on which hypotheses, experiments, or subgoals the agent identifies and initiates within that objective. A human-specified objective therefore does not by itself make the research agenda reactive. Recent conceptual work sharpens this distinction by arguing that externally scaffolded agentic workflows do not automatically internalize the goals, decision policies, self-regulation, or learning that would support a more internally grounded form of research agency (Xing et al., 2026). Recent surveys of agentic scientific discovery organize the field around a similar axis, charting the transition from isolated AI tools toward systems that participate across hypothesis generation, experimental design, execution, and iterative refinement (Gridach et al., 2025; Ren et al., 2026). The genuinely proactive content of autonomous research concentrates at three points: generating candidate questions, hypotheses, and subgoals (*Agenda*, §4.3); deciding which candidates warrant clarification, commitment, or stopping before scarce experimental resources are spent (*Arbitration*, §4.4); and building experimental-oracle and research-state infrastructure so that agent-initiated hypotheses can be validated, recorded, verified, forked, and extended. We organize this subsection along that axis, moving from autonomous execution of given goals toward systems that generate, select, validate, and preserve research trajectories. Table 8 compares selected research systems across the six mechanisms and their deployment settings.

6.6.1 Autonomous Execution of Given Research Goals

The clearest existing examples are end-to-end execution pipelines in the experimental sciences. Boiko et al. (2023b) introduced Coscientist, which couples GPT-4 with web and documentation search, code execution, and automated laboratory interfaces to plan and carry out experiments from a stated objective. ChemCrow follows a similar template, augmenting an LLM with expert-designed chemistry tools to plan syntheses and guide chemical discovery (M. Bran et al., 2024). Both Coscientist and ChemCrow are strongly autonomous in execution, gathering situational context through tools, planning multi-step procedures, and recovering from local failures, but the research question they pursue is still supplied by the human. In this survey’s vocabulary, they supply *Awareness* (§4.1) and *Action* (§4.5) at high capability while the top-level research objective remains human-specified. Complementary systems approach scientific

assistance from the literature side: retrieval-augmented agents such as PaperQA synthesize evidence across large corpora to answer specific scientific questions (Lála et al., 2023), and successors show that such agents can match or exceed expert humans on literature synthesis tasks (Skarlinski et al., 2024); LAB-Bench operationalizes the range of literature-grounded tasks that biology research assistants are expected to handle (Laurent et al., 2024). Agent Laboratory generalizes the execution template across the full research pipeline, coupling literature review, experimentation, and report writing in a single agentic system (Schmidgall et al., 2025); the Virtual Lab demonstrates the same pattern in biomedical research, assembling a team of LLM agents that collaboratively designed SARS-CoV-2 nanobodies against emerging variants, which human researchers subsequently validated experimentally (Swanson et al., 2025). More recent systems push this execution autonomy further: ScientistTwo takes a human-specified research problem, establishes baselines, formulates hypotheses, implements and runs experiments, performs ablations, and iterates through simulated peer review without human intervention (Nam et al., 2026). These systems can therefore automate increasingly large portions of the research process while remaining anchored to a human-directed question rather than originating the top-level objective themselves.

6.6.2 Hypothesis Generation and Research Agenda Formation

Proactiveness moves upstream when systems propose hypotheses, generate research directions, rank ideas, or decide which experiments should enter the loop. Lu et al. (2024) introduced The AI Scientist, which generates research ideas, writes code, executes experiments, produces manuscripts, and runs a simulated review process, applying the loop to machine-learning research where experiments can be executed as code; its successor, AI Scientist-v2, reduces reliance on human-authored code templates and manages the process with an experiment-manager agent using agentic tree search (Yamada et al., 2025). In biomedical research, Co-Scientist uses a generate-debate-evolve process to formulate and rank hypotheses under scientist-provided objectives (Gottweis et al., 2026); more recent work extends this framework toward execution-grounded research, coupling hypothesis formation to real-world validation across materials science, biology, and computer science (Schmidgall et al., 2026). Robin similarly closes the loop between hypothesis generation and experimental feedback in biology, integrating literature-search and data-analysis agents to generate hypotheses, propose experiments, interpret experimental results, and update hypotheses in an iterative discovery workflow (Ghareeb et al., 2026). Literature-grounded ideation systems such as ResearchAgent iteratively propose and refine research ideas using scientific literature and reviewer agents (Baek et al., 2025). These systems connect directly to *Anticipation* (§4.2) and *Agenda* (§4.3): they do not merely execute given tasks but generate the tasks themselves, the scientific analogue of the automatic curriculum in open-ended exploration (Wang et al., 2024a). Even so, the scientist typically still states the broad research goal; the system’s proactiveness lies in populating and refining the hypotheses that serve it. This exposes a central tension in agenda generation: the objective is not simply to produce ideas that appear novel, but to identify hypotheses that are feasible, testable, and likely to change scientific belief if validated. Empirical studies of LLM research ideation make this tension visible: LLM-generated ideas can score higher on novelty than human-generated ones while scoring lower on feasibility (Si et al., 2025). Recent benchmarks probe this gap more directly: AI Idea Bench evaluates AI research idea generation across dimensions such as novelty, feasibility, and alignment with target-paper ideas (Qiu et al., 2025), while HypoBench evaluates hypothesis-generation methods in terms of practical utility, generalizability, and discovery rate (Liu et al., 2026c). A complementary line of work asks whether novelty itself can be judged reliably: RQ-Bench shows that LLM judges can produce a “novelty mirage” by favoring model-generated research questions over author-anchored ones, while human experts prefer the latter (Sinhajari et al., 2026); RINoBench similarly finds that LLM rationales for research-idea novelty can resemble human explanations without yielding reliable novelty judgments (Schopf & Färber, 2026). These results suggest that agenda generation becomes valuable scientific progress only when candidate hypotheses are filtered through feasibility, testability, evidential leverage, and the cost of downstream validation, rather than novelty alone.

6.6.3 Experimental Design and Resource Allocation

Research is difficult for proactive agents because its feedback is not only costly, but often delayed, noisy, and asymmetric in what it reveals. In machine-learning research, the validation oracle is often cheap: the agent can run training code, score a benchmark, and iterate. In most empirical sciences the verdict lives in the physical world: validating a hypothesis may require synthesizing a molecule, culturing cells, running an animal study, or measuring a material property, often over days or weeks and with expensive reagents, specialized instrumentation, and uncertain readouts. Every candidate experiment becomes a costly, and sometimes practically irreversible, bet, and the burden shifts from executing experiments to deciding which experiments are worth performing (*Arbitration*, §4.4). Recent AutoResearch

systems have begun to make this allocation problem explicit: PrimeScientist represents competing research plans and their outcomes in an executable plan tree and conditions subsequent exploration on both experimental feedback and the remaining resource budget (Yu et al., 2026f). Luo et al. (2026b) target a complementary form of arbitration for open-ended deep-research agents: their agent learns to recognize when user intent is ambiguous and to ask a clarifying question before committing to a long and costly investigation. Both instantiate the same underlying decision: before committing scarce resources to a research trajectory, the agent must determine whether that trajectory warrants further investment. This arbitration problem has a longer methodological history in experiment selection. Active learning formalizes query selection as choosing the examples expected to be most informative for a learner (Settles, 2009), Bayesian experimental design frames experiment choice as the optimization of expected utility or information gain (Chaloner & Verdinelli, 1995), and Bayesian optimization provides a related framework for sequentially optimizing noisy, expensive black-box objectives, a setting common in chemistry and materials discovery (Shahriari et al., 2015). As an early instantiation of these principles in biological research, King et al. (2009) introduced the Robot Scientist Adam, selecting experiments for their information value rather than simply running them; its successor Eve applied related active-selection logic to drug screening (Williams et al., 2015). The self-driving laboratory paradigm that builds on these strategies has been reviewed across chemical and materials sciences (Abolhasani & Kumacheva, 2023; Seifrid et al., 2022). A-Lab provides a high-profile instantiation of this paradigm at scale: Szymanski et al. (2023) combined active learning, robotic synthesis, and literature-derived knowledge to accelerate inorganic materials discovery, though subsequent scrutiny of its product characterization and novelty claims illustrates a core fragility: an automated oracle is only as trustworthy as the measurements and assumptions behind it (Leeman et al., 2024). The bottleneck is therefore not whether an agent can act but whether the environment returns reliable evidence at the right level of abstraction, which is why the experiments an agent declines to run may be as informative as the ones it executes.

6.6.4 Experimental Oracles and Research-State Infrastructure

Closing the AutoResearch loop depends on two kinds of infrastructure that together determine whether the choices made during agenda generation and arbitration are backed by reliable evidence.

The first is the *experimental oracle*: the interface through which an agent queries the physical world and receives a measurement it can act on. Without a reliable oracle, even well-calibrated arbitration lacks a trustworthy feedback signal. Programmable and cloud laboratories reduce the latency of this query layer by turning robotic synthesis, assays, and instrumentation into remotely executable services (Armer et al., 2023). Self-driving laboratories close the loop more tightly by coupling automated experiment execution with measurement, model updating, and next-experiment selection, making the oracle not merely a passive readout but part of an adaptive discovery system (Abolhasani & Kumacheva, 2023; Seifrid et al., 2022; Tom et al., 2024). Concrete systems instantiate this idea across experimental settings: CAMEO used Bayesian active learning with a synchrotron beamline for closed-loop materials discovery (Kusne et al., 2020), while recent LLM-based laboratory agents have begun to control scientific instruments such as atomic-force microscopes through iterative measurement and parameter adjustment (Mandal et al., 2025). Laboratory-orchestration systems point toward a complementary interface layer, translating high-level experimental goals into executable workflows for automated infrastructure (Angelopoulos et al., 2026). Recent deployments of Co-Scientist further connect this interface to hypothesis-generating agents, coupling model-generated scientific proposals to semi-automated materials synthesis and previously unpublished biological measurements (Schmidgall et al., 2026). Oracle reliability nevertheless varies substantially across fields: in machine learning the validation signal is cheap, fast, and machine-readable, whereas in materials synthesis, wet-lab biology, and drug discovery it is slow, costly, and susceptible to noise, protocol variation, and confounded characterization, a fragility illustrated by the A-Lab scrutiny described above.

More fundamentally, experimental oracles define the *observable state space* within which an AutoResearch agent can reason: a fixed assay panel, sensor suite, or measurement schema may make closed-loop optimization efficient while constraining discovery to phenomena that the infrastructure can already detect. This limitation is already visible in in-silico research settings, where agents can propose new metrics, tests, or diagnostic benchmarks when existing evaluations fail to distinguish competing hypotheses. The physical analogue is measurement design: in the longer term, the most capable AutoResearch systems may need to recognize when existing sensors cannot adjudicate a hypothesis and propose new probes, assays, or instrument workflows that make the relevant variables observable.

The second is *research-state infrastructure*: the persistent scientific substrate that binds executable tools, data resources, workflows, intermediate artifacts, decision rationales, and evidence chains into a reusable state, so that later agents and scientists can verify, fork, and extend prior work rather than restarting from scratch. Standard publication formats

compress research into a linear narrative and discard much of the machine-actionable state that autonomous inquiry requires, including provenance, failed paths, branching history, tool configurations, execution traces, and the evidentiary links between claims and the artifacts that support them. The FAIR principles established Findability, Accessibility, Interoperability, and Reusability as baseline targets for scientific data (Wilkinson et al., 2016); computational notebooks demonstrated that executable, inspectable analysis records can coexist with human-readable narrative (Kluyver et al., 2016); and reproducibility standards and research-object conventions push further toward open methods, code, and workflow provenance (Stodden et al., 2016; Sandve et al., 2013; Bechhofer et al., 2013).

Recent agentic systems extend this lineage from reproducible artifacts toward persistent research workspaces, where state is not only what has been recorded but also what tools, datasets, workflows, and intermediate artifacts remain available for future action. Biomni illustrates a workspace-level version of research-state infrastructure in biomedical research: it exposes a broad action space of specialized tools, software packages, databases, and protocols to an AI agent, enabling workflows that connect literature review, bioinformatics analysis, code execution, and wet-lab protocol design (Huang et al., 2026a). Although biomedical in implementation, the broader design pattern is field-general: AutoResearch systems need persistent, agent-native environments that bind domain resources, executable workflows, intermediate outputs, and human-agent collaboration state into a reusable substrate for inquiry. AgentRxiv provides a complementary shared-state mechanism for agentic research, letting agent laboratories deposit and retrieve findings on a common preprint server so that cumulative progress compounds across runs (Schmidgall & Moor, 2025). Agora moves this shared state inside the research process itself, representing results, hypotheses, failures, and verifications as an append-only Git-backed contribution graph through which independently operating agents can inspect prior branches and build on a persistent research history (Zhang et al., 2026h). Agent-Native Research Artifacts (ARA) address within-loop continuity: each cycle’s output becomes a structured state representation that subsequent agents can audit, reproduce, and build upon, preserving the branching history of inquiry rather than compressing it into a linear narrative (Liu et al., 2026d). AutoScientists similarly treats research state as an active coordination substrate for long-running experimentation, allowing agents to interpret shared experimental state, critique proposals, and record both successful and failed directions to reduce redundant exploration (Gao et al., 2026c).

A further requirement is that accumulated state be auditable at the level of scientific claims. ScientistOne makes this evidentiary requirement explicit through Chain-of-Evidence, a framework in which citation, numerical, methodological, and conclusion claims in an AI-generated paper are linked to concrete evidence sources such as retrieved papers, code, execution logs, score files, and experimental artifacts (Meng et al., 2026). Its CoE Integrity Audit checks whether reported scores are reproducible, references are valid, task specifications are respected, and method descriptions align with the submitted code. This auditability pressure also reshapes scientific workbench design, where the workspace and evidence layers begin to merge. Claude Science integrates research tools, compute, and domain-specific skills into a scientific AI workbench while producing artifacts whose code, environment, and interaction history can be inspected for validation and reuse (Anthropic, 2026a). Together, the oracle layer and the state layer allow AutoResearch to be treated not as a series of isolated experiments but as an evolving, machine-readable inquiry that makes research executable, preserves failures, accumulates evidence, and supports principled continuation.

6.6.5 Evaluation, Challenges, and Open Directions

The hardest question for AutoResearch is not whether an agent can complete a research-shaped workflow, but when that workflow should count as scientific progress. A system can retrieve papers, propose hypotheses, write code, run experiments, and draft a manuscript while adding little reliable knowledge. Recent evaluations of open-ended research make this distinction empirical: in “shadow evaluations” on unpublished research questions, frontier agents autonomously completed much of the associated literature review, implementation, experimentation, and manuscript preparation but failed to make substantial progress on the underlying research questions, with expert analysis identifying failures in research judgment, experimental redesign, backtracking, resource awareness, and instruction adherence (Kirgis et al., 2026). Broader process-level evaluation in AutoResearchEval similarly identifies recurring failures in checking, revision, and reconsideration of the research trajectory across 100 frontier research tasks and 800 agent trajectories (Fei et al., 2026). Conversely, an agent that rules out an attractive but underpowered experiment, exposes an ambiguity in the objective, or preserves a failed branch for later reuse may contribute more to discovery than one that produces a polished result. Evaluation should therefore treat AutoResearch as an epistemic process rather than a task pipeline. The relevant metrics are not only answer accuracy, report quality, or benchmark score, but whether the agent changes what scientists are justified in believing, whether its claims can be reproduced and audited, whether its decisions improve the allocation

of scarce experimental attention, and whether the research state it leaves behind can support principled continuation.

This perspective makes several current risks sharper. Hypothesis-generation systems can overproduce novelty-like ideas whose feasibility or evidential leverage is weak; automated judges can mistake persuasive rationales for genuine scientific value; and closed-loop laboratories can optimize efficiently against measurements that are noisy, confounded, or too narrow to reveal the phenomenon of interest. Even when the evaluation itself is appropriate, a single experimental realization may be weak evidence about the underlying research idea: the “implementation lottery” shows that different plausible implementations of the same proposal can produce enough variation to reverse which ideas appear preferable, complicating the use of individual experimental scores for research branching and memory (Ning et al., 2026b). A more adversarial failure arises when an agent can influence the evidence by which its own success is judged. Recent experiments across autonomous research tasks find that agents can exploit evaluation procedures and adapt their strategies to repeated reviewer feedback, motivating evaluation criteria and confirmatory measurements that remain outside the producing agent’s control (Huang et al., 2026d). The long feedback cycles of biology, chemistry, and materials science amplify these risks: by the time an error is visible, the agent may already have consumed weeks of experimental effort or propagated an unsupported claim into downstream artifacts. For this reason, evaluation must also inspect the evidence chain that connects an agent’s final claims to its intermediate decisions, executions, measurements, and failed alternatives. Recent work on discovery certification makes this distinction explicit by separating attainment of a target score from evidence that the claimed discovery is attributable to the agent’s research process, using controlled recovery and feedback tests with independently verifiable evidence (Ning et al., 2026a).

Recent developments in mathematics and biology make these questions increasingly concrete outside controlled AutoResearch benchmarks. In mathematics, frontier AI systems have begun to produce results on longstanding open problems: OpenAI reported a series of advances across mathematics and theoretical computer science and subsequently released an AI-generated proposed resolution of the Navier–Stokes Millennium Prize problem together with a Lean formalization (OpenAI, 2026e,d). The latter was produced through a large-scale multi-agent search in which thousands of concurrent agents explored and exchanged candidate approaches before the resulting argument was formalized and checked. The ultimate mathematical status and significance of such results remain subject to scrutiny, but their rapid emergence has already prompted a broader debate over what progress in mathematics should optimize for. A public declaration initially signed by 25 Fields Medalists argued that treating the solution of prominent open problems as the objective of AI-assisted mathematics risks mistaking a proxy for the broader goals of mathematical research, which also include developing conceptual understanding, useful questions and methods, attribution, and a community capable of interpreting and extending the resulting knowledge (A Severe Misalignment of AI in Mathematics, 2026). This view is not universal, and other mathematicians have argued that AI-mediated and machine-verified mathematics can become another legitimate extension of how mathematical knowledge is produced. The debate nevertheless illustrates a distinction that becomes more important, rather than less, as execution and verification improve: a verified result need not by itself exhaust the scientific value of the process that produced it. Mathematics therefore provides an instructive limiting case for AutoResearch—as the cost of generating and, in formally specified settings, mechanically checking candidate results falls, the scarce resource can shift toward selecting worthwhile questions, interpreting results, establishing provenance, and integrating them into cumulative scientific knowledge.

Biology exposes a complementary difficulty because its oracle ultimately remains physical. Recent systems increasingly couple computational hypothesis generation to real-world experimentation: Co-Scientist has been evaluated in lab-in-the-loop workflows spanning materials synthesis and biological phenotype prediction (Schmidgall et al., 2026), while Anthropic has reported an agentic search over biological sequence space that identified a previously uncharacterized enzyme system associated with CRISPR-like repeat arrays and subjected parts of the resulting hypothesis to wet-lab testing (Anthropic, 2026c). The latter remains an early result: the observed repeat-associated system has properties suggestive of an interesting biological mechanism, but its function and broader significance require further experimental characterization. More generally, such episodes make questions of novelty, provenance, attribution, and independent validation increasingly salient when agents search scientific literature and data, formulate hypotheses, and participate in their experimental validation. These questions should not be taken to prejudge particular scientific or priority claims; rather, they illustrate why increasingly autonomous discovery requires transparent provenance and independent experimental evidence in addition to capable hypothesis generation. Together, the mathematics and biology cases expose complementary limits of execution-centric AutoResearch: even where correctness can increasingly be mechanically checked, scientific value is not exhausted by verification, while in empirical science the evidence itself remains costly, partial, and contestable.

Table 9 Representative proactive smart-home systems. Res. denotes research systems; Com. denotes commercial products.

System	Awareness	Anticipation	Agenda	Arbitration	Action	Adaptation	Deploy.	Type
Alexa (Hunches) (Eberhardt, 2022)	Device states, usage history	Deep device / home embeddings	Inferred device-state corrections	Learned acceptance model	IoT device control (<i>gated action</i>)	Acceptance learning	Smart home	Com.
Sasha (King et al., 2024)	User goal, device / sensor inventory	Multi-step LLM planning	User goal; agent subplans	Feasibility check; clarification	Device control, routines (<i>action</i>)	User feedback, replanning	Smart home	Res.
Trace2TAP (Zhang et al., 2020a)	Sensor / manual-action traces	Symbolic reasoning + SAT	Candidate routines from traces	User selection / feedback	Routine proposals (<i>suggestion</i>)	Feedback-guided revision	Software	Res.
PreHeat (Scott et al., 2011)	Occupancy, temperature history	Occupancy prediction	Standing comfort objective	Setpoints + heating lead time	Predictive heating (<i>action</i>)	Updated occupancy history	Smart home	Res.
DESAMO (Choi et al., 2025)	Speech, ambient audio	Audio LLM event inference	Standing safety monitoring	Detected emergency labels	Alarms, caregiver alerts (<i>action</i>)	—	Edge device	Res.

These considerations shift evaluation from raw initiative to the calibration and accountability of initiative. A mature research agent should know when to generate ideas, when to ask for clarification, when not to spend resources, when to distrust an oracle, and when existing measurements cannot adjudicate the question. Open directions include benchmarks for experiment refusal and clarification, protocols for preregistered agent-generated studies, evidence-chain audits for AI-produced scientific claims, auditable research-state artifacts that preserve failed branches and decision rationales, and measurement-aware agents that design new assays, probes, simulations, or diagnostic benchmarks when the current observable state space is inadequate. In AutoResearch, mature proactiveness is not the ability to do more science unaided, but the ability to ask better questions, choose better tests, expose the evidence behind its claims, and leave behind a research state that others can verify, fork, and extend.

6.7 Agents for Smart Home

The smart home is the private, consumer end of proactive deployment: one household, a few co-present occupants, and an always-on fabric of appliances, speakers, and sensors that observes daily life at close range and acts on it over minutes and days. Standing goals such as comfort, safety, and energy thrift are rarely restated, so initiative must come from the agent; yet every unsolicited act lands on a co-present person in a private space, so the felt costs of error are intrusion, annoyance, and lost trust rather than systemic harm. The stakeholders are correspondingly intimate: occupants, family members, and caregivers, with platform vendors and utilities in the background, and consumer privacy as the governing regulatory constraint.

Agents for smart home have seen wide deployment, yet most systems still wait to be asked. What ships at scale today is condition-triggered scene automation: user-authored rules of the form “when this, then that” that the platform merely executes. The deployed frontier is acceptance-gated, single-step anticipation: Alexa Hunches flags potentially forgotten actions and supports automatic actions, such as locking a door or turning off a light. In 2022, Amazon reported that features including Hunches and Routines initiated roughly one in four Alexa smart-home interactions without a spoken request (Eberhardt, 2022). The next level, toward which research systems point, is learned-preference ambient orchestration under standing objectives, such as comfort and energy budgets that the agent manages continuously without per-rule authoring; the ultimate level is household-scale agenda ownership with calibrated interruption. The binding constraint on this progression is not anticipation but arbitration, that is, knowing whether, when, and how strongly an intervention will be welcome (§4.4). Relative to the other verticals in this section, the home thus shows the broadest consumer deployment of proactivity and among the shallowest, a mirror image of autonomous research, whose agents are autonomy-deep but barely deployed. Table 9 summarizes selected approaches to household orchestration, routine synthesis, predictive heating, and safety monitoring.

6.7.1 Ambient Household Orchestration

The flagship use case is coordinating many devices toward goals the occupant states vaguely or not at all. Its deployed anchor is Alexa Hunches, which learns *deep device* and *deep home* embeddings as a world model over the household’s device states and gates each candidate action on a learned model of whether the user will accept it (Eberhardt, 2022). Research systems put LLM agents (Wang et al., 2024c) on the same stage: Sasha maps under-specified goals such as “make it cozy” onto whatever devices the home contains (King et al., 2024), SAGE grounds open-ended requests in verifiable device actions (Rivkin et al., 2024), and LLMind, Penetrative AI, and IoT-LLM reason directly over raw sensor and device signals (Cui et al., 2023a; Xu et al., 2024c; An et al., 2025). A complementary line converts continuous context into anticipated tasks so the agent offers help before being asked: Lu et al. (2025a) learn from real activity traces with a reward model as a proxy for user acceptance, and ProAgent decides in the wild, from on-demand sensory context, when assistance is worth initiating (Yang et al., 2025a).

6.7.2 Trigger-Action Automation and Routine Synthesis

Trigger-action programming (TAP) is the canonical way households encode standing policy, from IFTTT recipes to vendor routines (Ur et al., 2014; Mi et al., 2017). Its known failure modes, inaccurate mental models and buggy rules (Huang & Cakmak, 2015; Brackenbury et al., 2019), motivated rule recommendation and end-user debugging aids (Corno et al., 2019b,a). Proactive agents move the home from executing rules to authoring them: AutoTap synthesizes automations from stated properties (Zhang et al., 2019), Trace2TAP proposes them from observed behavior traces (Zhang et al., 2020a), and LLM-based systems now generate, contextualize, and repair routines in situ (Badalov & Yoon, 2026; Giudici et al., 2025; Yu et al., 2026b). Routine synthesis is the clearest instance of the agenda shifting from user to agent (§4.3) while the user retains veto power by editing the proposed rule.

6.7.3 Energy and Comfort Management

Household energy rewards anticipation because HVAC and appliances carry thermal and temporal inertia. Scott et al. (2011) showed early that occupancy-predictive heating outperforms fixed schedules; nonintrusive load monitoring gives the agent appliance-level visibility from a single meter (Kelly & Knottenbelt, 2015); and deep reinforcement learning schedules loads against dynamic prices and comfort constraints (Yu et al., 2020). In deployment terms this is the most automated corner of the home: pre-conditioning and load shifting act silently under a standing comfort and cost budget, previewing the standing-objective orchestration the rest of the home still lacks.

6.7.4 Health Monitoring and Aging in Place

Ambient assisted living is where home proactivity carries the highest stakes. Its perceptual base is two decades of sensor-based human activity recognition, from probabilistic models in instrumented homes (van Kasteren et al., 2008) to deep, multimodal, and wearable recognizers (Ordóñez & Roggen, 2016; Yao et al., 2017; Chen et al., 2021; Bouchabou et al., 2021; Ni et al., 2024) and vision-based fall detection (Alam et al., 2022). The proactive trigger is typically a deviation from a personal baseline rather than a rule: anomaly detection over the daily routines of older adults living alone supplies the alarm (Shahid et al., 2022). Agentic systems close the loop with continuous monitoring, check-ins, and autonomous escalation to caregivers (Khalil et al., 2025; Choi et al., 2025), and embodied assistants anticipate which objects an occupant will need from a spatio-temporal model of the home (Patel & Chernova, 2022). Because the monitored population is vulnerable, privacy-preserving learning of activity predictors is part of the use case rather than an afterthought (Zehtabian et al., 2021).

6.7.5 Evaluation, Challenges, and Open Directions

Evaluation has matured from one-off user studies to purpose-built benchmarks: SimuHome stresses proactive intent inference, physical-limit checks, and temporal scheduling in a Matter-compliant simulator whose policies transfer to real devices (Seo et al., 2025), and PersonalHomeBench scores reactive and proactive behavior under personalized, multimodal observation (Bharadwaj et al., 2026) (§5). The open challenges are less perceptual than social. First, timing and agency: human-factors studies converge on the finding that ill-timed initiative distracts and annoys, with interruptibility depending on the occupant’s activity, engagement, and mood (Wei et al., 2022a,b), and the acceptable level of initiative is itself user- and context-dependent (Reicherts et al., 2021; Zargham et al., 2022) (§4.4.3). Low-commitment initiative, asking or suggesting rather than acting silently, better preserves user agency (Oh et al., 2024;

Table 10 Representative proactive smart-city systems. Res. denotes research systems; Com. denotes commercial products.

System	Awareness	Anticipation	Agenda	Arbitration	Action	Adaptation	Deploy.	Type
LLMLight (Lai et al., 2023)	Queues, approaching vehicles	LLM traffic reasoning	Given traffic-control objective	Phase comparison / selection	Signal switching (<i>action</i>)	Imitation + critic training	Simulation	Res.
CityLearn agents (Vázquez-Canteli et al., 2020)	Building loads, weather, storage	RL value / policy learning	Given demand-response objective	Storage / action constraints	Storage scheduling (<i>action</i>)	RL in simulation	Simulation	Res.
Google Flood Hub (Nearing et al., 2024)	Weather, basin attributes	LSTM flow forecasts	Standing flood-warning objective	Flood alert thresholds	Early warnings (<i>suggestion</i>)	—	Cloud	Com.
EpiPlanAgent (Mao et al., 2025)	Outbreak reports, guidelines	LLM + retrieved plans	Given outbreak; response subtasks	Condition checks; expert review	Response plans (<i>suggestion</i>)	Feedback-based refinement	Software	Res.
AgentSociety (Piao et al., 2025)	Agent profiles, simulated state	LLM-driven simulation	Assigned simulation scenario	Simulation sandbox	Intervention scenarios (<i>suggestion</i>)	—	Simulation	Res.

Deng et al., 2024b), and longitudinal deployment shows adoption hinges on the agent fitting into, rather than overriding, the user’s day (Abbas et al., 2026). Second, privacy: always-on sensing and unsolicited action keep smart speakers and hubs under persistent suspicion (Lau et al., 2018; Zheng et al., 2018; Malkin et al., 2019; Abdi et al., 2019). Third, the household is plural: most systems model a single principal, while real homes contain several people with conflicting preferences and asymmetric authority, a largely unsolved arbitration problem. The direction of travel is from acceptance-gated single actions toward standing-objective orchestration, which will require acceptance models that generalize across routines, preference learning that spans household members, and interruption budgets audited over weeks rather than judged per event (§7).

6.8 Agents for Smart City

The city swaps the home’s intimacy for scale: municipal infrastructure serving millions, operated by agencies and utilities, observed through traffic, mobility, environmental, and event telemetry, and governed publicly over horizons that stretch from signal cycles to multi-year plans. There is no single principal; consequences are diffuse, delayed, and contestable, and legitimacy is a matter of public accountability rather than personal preference. The technical substrate is the urban-computing tradition (Zheng et al., 2014): spatio-temporal deep networks that forecast citywide flows, traffic, and crime (Zhang et al., 2017; Li et al., 2018; Yu et al., 2018; Wu et al., 2019; Huang et al., 2018; Xia et al., 2021), lately consolidated by urban foundation models and LLM interfaces such as UrbanGPT (Li et al., 2024d; Zhang et al., 2026f; Han et al., 2025b).

Unlike agents for smart home (§6.7), agents for smart city are predominantly sensor-triggered alerting for human operators, dashboards that flag congestion, pollution, or hazards and leave the response to control rooms. The strongest deployed proactivity is early warning at planetary scale, flood and earthquake forecasts pushed directly to affected populations, which is anticipation without actuation. The research frontier is predict-then-act control validated largely in simulation, where agents pre-time signals, rebalance fleets, and shape electrical loads. The next level is agent-driven anticipatory infrastructure operations that pre-position resources before disruptions cascade, held under granted, supervised, and revocable authority; the ultimate level is coordination across networks, where traffic, energy, and emergency agents negotiate rather than optimize in silos. Extending this progression to closed-loop municipal action requires resolving diffuse authority and public accountability (§7.2). Table 10 summarizes selected approaches to traffic control, energy management, early warning, emergency planning, and urban simulation.

6.8.1 Traffic and Mobility Orchestration

Signal control is the canonical testbed. Deep and multi-agent reinforcement learning agents learn to pre-empt congestion rather than react to it, scaling from single intersections (Wei et al., 2018) to network-level coordination (Wei et al., 2019b), pressure-based and thousand-light control (Wei et al., 2019a; Chen et al., 2020), and meta-learned transfer across intersections (Zang et al., 2020); LLMLight now casts the language model itself as the signal-control agent (Lai et al., 2023), while TrafficGPT and TransGPT let operators steer traffic foundation models through dialogue (Zhang et al., 2024f; Wang et al., 2024e). On the mobility side, dispatch and repositioning agents act ahead of demand: reinforcement-learned order dispatching and fleet management in ride-hailing (Xu et al., 2018; Lin et al., 2018; Li et al., 2019; Jin et al., 2019; Al-Abbasi et al., 2019) and model-predictive and graph-learned rebalancing of autonomous mobility-on-demand (Iglesias et al., 2018; Gammelli et al., 2021).

6.8.2 Grid and Utility Operations

Urban energy agents reshape demand before peaks arrive. Reinforcement-learned demand response is a mature literature (Vázquez-Canteli & Nagy, 2019), given a shared testbed by CityLearn, which coordinates load shaping across building portfolios (Vázquez-Canteli et al., 2020); multi-agent controllers manage residential loads on fast timescales (Mai et al., 2023; Shojaeighadikolaei et al., 2022) and schedule electric-vehicle charging ahead of grid stress (Park & Moon, 2022). Water, road maintenance, and other utility networks remain comparatively agent-poor, an asymmetry we return to below.

6.8.3 Environmental and Public-Health Early Warning

Field sensing coupled to forecasting is the city’s most consequential deployed proactivity. Google’s global flood model anticipates extreme riverine floods in ungauged basins and pushes warnings days ahead across dozens of countries (Nearing et al., 2024), and deep models predict wildfire risk from remote-sensing and environmental data (Xu et al., 2024d) and interpolate and forecast citywide air pollution (Le et al., 2020). On the public-health side, generative-agent epidemic models anticipate contagion dynamics jointly with behavioral response (Williams et al., 2023). Characteristically, all of these stop at the alert: anticipation is delivered to humans, and actuation remains institutional.

6.8.4 Emergency Response and Civic Coordination

Agentic systems are beginning to convert warnings into coordinated response, today as human-supervised decision support rather than deployed responders. Geospatially grounded LLM agents turn live wildfire detections into resource-allocation recommendations (Chen et al., 2025g); EpiPlanAgent drafts and validates epidemic response plans with practitioners in the loop (Mao et al., 2025), and LLM agents act as policymakers in simulated epidemics (Aoki & Ghaffarzadegan, 2026); CrisisSense-LLM distills situational awareness from disaster social media (Yin et al., 2024a); and vision papers sketch disaster-management architectures that chain hazard early warning to risk analytics and coordinated response (Li et al., 2025a). High inaction costs license more initiative here than anywhere else in the city, but consequences are contestable enough that authority must remain granted, supervised, and revocable: proactiveness supplies initiative, not authority (Chan et al., 2023; Kasirzadeh & Gabriel, 2025).

6.8.5 Urban Simulation and Digital Twins

Where the relevant futures cannot be observed, planners rehearse interventions in silico (§4.2.4). Classical agent-based modeling (Bonabeau, 2002) and transport microsimulation (Horni et al., 2016) are being refounded on generative agents (Park et al., 2023): OpenCity scales LLM-driven urban simulation (Yan et al., 2024), AgentSociety simulates more than ten thousand LLM-driven residents (Piao et al., 2025), UGI proposes a foundational platform for embodied city agents (Xu et al., 2023), CitySim and GATSim model urban behavior and mobility with generative agents (Bougie & Watanabe, 2025; Liu et al., 2025c), and generative AI is likewise entering urban digital twins (Xu et al., 2024b). For proactive operations these simulators are the counterfactual engine: policies for signals, fleets, and emergency plans are tested against synthetic populations before touching the street.

6.8.6 Evaluation, Challenges, and Open Directions

Evaluation is the city’s soft spot. Predict-then-act agents are overwhelmingly validated in simulators, CityLearn for energy (Vázquez-Canteli et al., 2020) and MATSim-style platforms for mobility (Horni et al., 2016), and the sim-to-real

gap remains largely uncrossed: few signal-control or rebalancing policies in this literature have faced a real network with liability attached. Generative-agent simulators promise richer counterfactuals but are themselves unvalidated instruments, so a policy tested only against synthetic residents inherits their biases (Han et al., 2025b). Deployed early-warning systems show the inverse strength, planetary reach with measurable lead times (Nearing et al., 2024), and could anchor evaluation practice, since lead time, false-alarm rate, and coverage equity are metrics the field already trusts. The deeper challenges are institutional. With no single principal, endorsement is a matter of governance rather than preference: accountability, oversight, and contestability are deployment preconditions, not afterthoughts (Chan et al., 2023; Kasirzadeh & Gabriel, 2025; Schmitz et al., 2025), and equity questions, such as whose neighborhood receives pre-positioned resources, have no analogue in single-user domains. Proactive output therefore skews toward suggestion over action. Open directions follow: staged autonomy that graduates agents from dashboards through recommendations to bounded actuation with automatic rollback; standing fairness audits of proactive allocation; extension to under-served utility networks such as water and road maintenance; and cross-network coordination so that traffic, energy, and emergency agents stop optimizing in silos. How initiative can move safely from calibrated suggestion to trustworthy action is taken up in §7.

6.9 Agents for Healthcare and Medicine

Table 11 Representative proactive healthcare and medical-agent systems and benchmarks. Res. denotes research systems; Com. denotes commercial products.

System	Awareness	Anticipation	Agenda	Arbitration	Action	Adaptation	Deploy.	Type
Agent Hospital (Li et al., 2024b)	Patient history, consultation, examinations	Missing diagnostic evidence	Clinical consultation	Iterative examine–diagnose–treat process	Follow-up questions, examinations, treatment plans (<i>suggestion</i>)	Updates reasoning as new evidence arrives	Simulation	Res.
MINT (Fang et al., 2026a)	Incrementally revealed patient information	Whether current evidence justifies an answer	Diagnostic task	Answer vs. hold decision	Diagnostic answers or hold responses	Revises diagnosis as additional evidence is revealed	Benchmark	Res.
MDAgents (Kim et al., 2024c)	Clinical case, estimated complexity	Need for additional specialist reasoning	Diagnostic task	Solo vs. multi-specialist reasoning	Diagnosis and specialist consultation (<i>suggestion</i>)	Changes collaboration with case complexity	Software	Res.
MedAgentBench (Jiang et al., 2025)	Patient-specific EHR state	Intermediate workflow requirements	Assigned clinical task	Selection of EHR operations and intermediate steps	EHR interaction and clinical task execution (<i>action</i>)	Progresses with updated record state	Benchmark	Res.
irAE-Agent (Gallifant et al., 2025)	Longitudinal clinical notes	Potential immune-related adverse events	Adverse-event detection	Risk identification from record evidence	Surfaces potential complications (<i>suggestion</i>)	Updates assessment as new notes arrive	Software	Res.
STELLA (Jin et al., 2026)	Scientific evidence, prior exploration	Promising subsequent research directions	Biomedical research	Evolving reasoning templates and tool construction	Scientific exploration and tool use (<i>action</i>)	Evolves strategies and tools with experience	Software	Res.

Proactive agents for healthcare span clinical intake, diagnosis, treatment, follow-up, patient monitoring, and biomedical research. Across these settings, care often needs to advance before a patient or clinician explicitly asks for the next action. Patients may omit relevant information, new test results may alter an earlier assessment, and unresolved actions or subtle deterioration may require attention before the next consultation. These situations create opportunities for agents to act proactively by identifying what information, expertise, or follow-up is needed as a case evolves.

Current healthcare agents address some of these needs, but their proactiveness is still concentrated within active tasks or episodes of care. Once a case is underway, they can seek missing evidence, ask follow-up questions, consult specialists,

invoke medical tools, and organize subsequent steps. Some systems also monitor evolving records or pursue biomedical hypotheses over time, although their ability to sustain such behavior across visits remains limited. More longitudinal forms of proactiveness would require agents to preserve unresolved concerns, connect later evidence to earlier decisions, and recognize when a change warrants renewed clinical attention. These capabilities expand when and how an agent contributes to care, while diagnostic and therapeutic authority remains with clinicians.

We examine this development through four connected workflows. Clinical history taking concerns proactive evidence acquisition. Multi-specialist deliberation extends proactiveness to the selection and coordination of expertise. Longitudinal care coordination requires agents to maintain unfinished clinical work over time, while continuous monitoring and biomedical discovery move toward the proactive formation of new clinical or scientific objectives. Table 11 compares selected systems and benchmarks across these workflows using the six mechanisms.

6.9.1 Clinical History Taking and Iterative Anamnesis

The first form of clinical proactiveness is deciding what must be learned before a diagnosis can be trusted. An initial patient account rarely includes every symptom, prior condition, medication, laboratory finding, demographic factor, or temporal detail needed for a reliable assessment. An agent that responds only to the presented information may therefore answer an underspecified version of the clinical problem. Proactive history taking requires it to identify which missing evidence could change the assessment and continue the interaction until the case is sufficiently specified.

Existing systems increasingly place parts of this evidence-gathering process under agent control. Medical consultation agents (Ren et al., 2025) maintain dialogue and patient-history information across an encounter, allowing later questions to depend on what has already been learned rather than treating each turn independently. In MedAgentSim (Almansoori et al., 2025), doctor agents interact with simulated patients and request examinations or imaging as additional evidence becomes necessary, extending information acquisition beyond conversational questioning alone. MeDxAgent (Sanghvi et al., 2026) ties this process more directly to diagnostic uncertainty by using evolving differential diagnoses to guide targeted follow-up questions. Agent Hospital (Li et al., 2024b) embeds similar information seeking within a broader simulated clinical process, where consultation and examination jointly determine what evidence becomes available before diagnosis. Together with these interaction-oriented systems, the progression is toward agents that shape the diagnostic evidence they reason over, rather than accepting the initial case description as fixed.

The MINT benchmark (Fang et al., 2026a) highlights a complementary requirement: controlling when a model commits to a diagnosis as evidence accumulates. In MINT, clinical evidence is supplied incrementally by the evaluation protocol rather than actively requested by the model, and at each turn the model may provide a diagnostic answer or continue to hold its response. The benchmark shows that models often commit before sufficient evidence has been observed, can later self-correct as additional evidence arrives, and are particularly susceptible to premature commitment when salient evidence such as laboratory findings appears early. MINT therefore evaluates commitment timing rather than proactive evidence acquisition itself, while showing why a diagnostic agent that actively gathers evidence must also know when the available information is sufficient to justify an answer.

6.9.2 Multi-Specialist Deliberation and Case Conferencing

Even thorough evidence collection may leave uncertainty that cannot be resolved from a single clinical perspective. Imaging, pathology, pharmacology, surgery, and disease-specific practice can yield different interpretations of the same case. Proactiveness at this stage requires the agent to recognize that the current reasoning process lacks the expertise needed to proceed and to broaden the deliberation before the case converges prematurely (Liu et al., 2023d; Gan et al., 2025).

MedAgents (Tang et al., 2023) brings multiple medical experts into iterative discussion and consensus formation, moving specialist consultation into the reasoning process rather than requiring each opinion to be solicited separately. MDAgents (Kim et al., 2024c) adds another layer of control by adapting between individual and collaborative reasoning according to estimated case complexity, so the amount of expertise involved can vary with the case itself. MedAide (Wei et al., 2024) coordinates specialized medical agents according to different clinical intents, emphasizing how heterogeneous expertise can be assembled around the needs of a consultation. MedChat (Liu et al., 2025b) similarly combines role-specific LLM agents and specialized vision models under a directing agent, making the integration of multimodal clinical perspectives part of the agentic workflow. When consensus is difficult, MACD (Li et al., 2025d) uses iterative consultation and evaluation to continue deliberation rather than accepting the first converged answer.

MedCoAct (Zheng et al., 2025b) coordinates doctor and pharmacist agents while using confidence to regulate their interaction, linking the intensity of collaboration to uncertainty in the current reasoning process. MediHive (Wang & Yang, 2026) reduces advance specification further by allowing agents to assume specialized roles and debate divergent conclusions. MMedAgent-RL (Xia et al., 2025) learns to coordinate multimodal specialist inputs through triage and referral, shifting part of the decision over where a case should be routed into the learned policy itself.

The scope of this coordination is usually fixed in advance. Most frameworks receive a predefined case, objective, and set of specialist roles. They can manage deliberation after it has been configured, but seldom decide that another specialty is needed, that the current consultation structure is inadequate, or that later evidence warrants reopening an earlier consensus. Current systems are therefore more mature at proactively conducting multi-specialist reasoning than at determining when and how such reasoning should be initiated.

6.9.3 Longitudinal Care Coordination

After a diagnosis or treatment plan has been produced, the source of risk shifts from incomplete reasoning to incomplete follow-through. Tests must be completed, referrals attended, reports reviewed, treatments initiated, and patient responses reassessed. Because these actions are distributed across visits and clinical services, an appropriate recommendation may still fail when one step is delayed, forgotten, or disconnected from later evidence. Proactive support must therefore preserve the unfinished state of care instead of treating the production of a plan as the end of the task.

Several medical agents already reduce the need to specify every intermediate step within an active clinical episode. MMedAgent (Li et al., 2024a) plans over multiple medical tools and integrates their outputs, giving the agent control over which operations are needed to advance the assigned task. MedAgent-Pro (Wang et al., 2026o) structures diagnostic reasoning hierarchically and delegates multimodal evidence analysis to specialized tools, extending this control from tool selection to the organization of a multi-step diagnostic workflow. MedAgentBench (Jiang et al., 2025) evaluates agents in an EHR environment on patient-specific clinical tasks that require interaction with medical-record systems, providing a setting for assessing agentic execution in operational workflows. MRGAgents (Wang et al., 2025c) distributes radiology report generation across disease-specialized agents, while radiology reasoning agents (Yi et al., 2025) coordinate retrieval, visual interpretation, drafting, refinement, and synthesis as connected stages rather than isolated prompts. TheraAgent (Li et al., 2026c) applies a related idea to treatment planning, repeatedly evaluating and refining a proposed therapy instead of treating the first generated plan as complete. Across these settings, current proactiveness is strongest in advancing an already established clinical objective through the steps needed to complete it.

Their involvement, however, usually ends with the current case or episode. After producing a plan or report, the agent rarely checks whether the proposed action occurred or whether later evidence still supports the original decision. Longitudinal proactiveness would require persistent links among recommendations, pending tests, referrals, treatments, and subsequent observations. Such continuity would allow the agent to recover unfinished care, detect deviations from the expected trajectory, and return the case to clinical attention before delay becomes consequential.

6.9.4 Continuous Patient Monitoring and Biomedical Hypothesis Discovery

Longitudinal coordination maintains an objective that has already been established. A further expansion of proactiveness occurs when the agent must recognize that a new objective has emerged. Changes in laboratory trends, clinical notes, medication exposure, or patient-generated signals may indicate deterioration before a clinician requests another assessment. Continuous monitoring can reduce this delay by detecting meaningful changes and deciding whether they justify a new clinical concern.

irAE-Agent (Gallifant et al., 2025) searches clinical notes for immune-related adverse events, allowing evidence appearing in the record itself to surface a potential complication that may require further attention. AgentRx (Al Jorf & Shamout, 2026) evaluates LLM-based agents on multimodal clinical risk prediction using heterogeneous patient information. irAE-Agent directly illustrates how newly observed evidence can surface a potential reason for further clinical work, while AgentRx evaluates the ability of agentic systems to integrate patient information for risk prediction. Their value depends on calibrated intervention. Excessive alerts transfer the interpretive burden back to clinicians and may obscure the few events that require action. Useful proactive monitoring therefore depends on distinguishing changes that warrant escalation from those that do not.

Biomedical research presents a related transition from pursuing an assigned objective to expanding it. DrugAgent (Liu

et al., 2024d) automates machine-learning programming for drug-discovery tasks through agentic planning and tool use, allowing a broad research objective to be carried forward through multiple scientific operations. PharmaSwarm (Song et al., 2025a) distributes pharmaceutical research across collaborating agents, so candidate directions can be examined from multiple functional perspectives without each step being separately specified by the researcher. STELLA (Jin et al., 2026) supports longer-horizon scientific exploration through self-evolving reasoning templates and dynamically constructed tools as experience accumulates. AgentD (Ock et al., 2025) similarly integrates biomedical retrieval, question answering, molecular generation, property prediction and refinement, and structure generation within an agentic drug-discovery workflow. These systems can therefore automate extended parts of scientific exploration, with some systems able to surface directions that a researcher did not enumerate in advance. Their search spaces, evaluation criteria, and high-level goals are still largely supplied by humans. Proactive scientific discovery remains less developed when the agent must decide that accumulated evidence justifies a new hypothesis or a change in research direction.

6.9.5 Evaluation, Challenges, and Open Directions

Healthcare agents have progressed from proactive evidence collection (Li et al., 2024b; Almansoori et al., 2025; Ren et al., 2025; Sanghvi et al., 2026) to proactive coordination of clinical expertise (Wei et al., 2024; Tang et al., 2023; Kim et al., 2024c; Liu et al., 2025b; Li et al., 2025d; Zheng et al., 2025b; Wang & Yang, 2026), proactive workflow execution (Li et al., 2024a; Liu et al., 2025b; Wang et al., 2025c; Yi et al., 2025), and continuous risk detection or biomedical exploration (Gallifant et al., 2025; Liu et al., 2024d; Jin et al., 2026). Across this progression, systems have become increasingly capable of deciding what to do next once a case or objective is active. They remain less reliable at maintaining responsibility across interruptions, reopening dormant issues, and forming new objectives from changes that no user has yet raised. Healthcare proactiveness is therefore relatively mature within consultations and predefined workflows, while longitudinal and objective-forming proactiveness remains at an early stage.

Evaluation should distinguish these levels of capability. Evidence-gathering agents should be assessed on whether they ask necessary questions before concluding; deliberative systems on whether they involve relevant expertise and expose unresolved disagreement; care-coordination agents on whether they recover unfinished actions; and monitoring systems on whether they detect emerging risks without creating excessive burden. Prospective and cross-institutional validation is needed to determine whether these behaviors remain reliable across clinical settings. As agents take on a larger role in initiating and sustaining clinical work, evaluation must also verify that their recommendations remain reviewable and that the division between proactive support and clinician authority is preserved.

6.10 Agents for Law

Table 12 Representative proactive legal-agent systems. Res. denotes research systems; Com. denotes commercial products.

System	Awareness	Anticipation	Agenda	Arbitration	Action	Adaptation	Deploy.	Type
L-MARS (Wang & Yuan, 2025)	Query, legal authorities	Evidence gaps; jurisdiction / temporal validity	Research subquestions	Judge-agent sufficiency assessment	Iterative retrieval and legal research (<i>action</i>)	Extends search when support is insufficient	Software	Res.
AgentsCourt (He et al., 2024)	Case facts, legal resources	Opposing arguments and weaknesses	Predefined dispute	Multi-agent courtroom debate	Counterarguments and decision refinement (<i>suggestion</i>)	Revises reasoning after adversarial interaction	Simulation	Res.
LegalSim (Badhe, 2025)	Case state, procedural context	Consequences of procedural choices	Predefined proceeding	Competing agent strategies	Discovery requests, motions, procedural choices (<i>action</i>)	Strategy evolves with opponent actions	Simulation	Res.
AutoPatent (Wang et al., 2024f)	Initial draft, technical information	Downstream drafting and review needs	Patent-generation objective	Planner, writers, examiner	Patent drafting and revision (<i>action</i>)	Revises documents from examiner feedback	Software	Res.
Compliance Brain Assistant (Zhu et al., 2025b)	Enterprise compliance data	Missing compliance-relevant context	Compliance query	Agentic tool / API selection	Compliance investigation and risk analysis (<i>suggestion</i>)	Expands context through additional tool calls	Software	Res.

Legal practice typically encompasses investigating facts and authorities, developing and testing legal positions, preparing documents, coordinating matters, and identifying potential violations or exposure. Much of this work cannot be

determined at the outset, as the legal agenda itself changes with the matter. For example, a new fact may alter the applicable legal theory, an opposing interpretation may require additional evidence, and a revised agreement may create new procedural or compliance obligations. Legal assistance therefore becomes proactive when it can recognize that such developments have created new work before the client or attorney explicitly requests it.

Current legal agents show meaningful proactiveness once a legal issue has already entered the agenda. They can extend incomplete research, introduce unrequested counterarguments, select intermediate workflow actions, and identify risks in submitted materials. Their main limitation is that the scope of legally relevant work is still largely inherited from the user. Further progress depends on recognizing when interactions among facts, rules, documents, and procedural developments create a new obligation, claim, vulnerability, or risk. Such recognition requires an evolving representation of the matter through which the agent can revise what deserves investigation, while attorneys retain authority over legal judgment and consequential action.

We examine how far this transition has progressed across legal research and evidence gathering, argument stress-testing and adversarial review, legal matter management and workflow execution, and compliance monitoring and legal-risk anticipation. Table 12 compares selected legal systems across these applications using the six mechanisms.

6.10.1 Legal Research and Evidence Gathering

The proactive challenge in legal research begins before an authority is retrieved. An initial inquiry may omit material facts, assume an incorrect legal category, or overlook a jurisdictional issue that changes which authorities apply. When these defects remain hidden, additional retrieval may strengthen an answer without correcting the question on which it is based. A legal research agent therefore needs to assess whether the current framing and evidentiary record are sufficient, identify what remains unresolved, and direct further investigation toward those gaps.

Recent systems have made clear progress toward this form of proactiveness through retrieval, planning, and verification. LegalAgentBench (Li et al., 2025b) evaluates agents on multi-step legal tasks that require reasoning over real-world legal corpora and selecting among external tools, making both tool use and intermediate research progress observable. L-MARS (Wang & Yuan, 2025) more directly implements iterative legal research: it decomposes a query into subproblems, searches heterogeneous legal sources, and uses a judge agent to assess evidentiary sufficiency, jurisdiction, and temporal validity before deciding whether further retrieval is needed. CourtReasoner (Han et al., 2025d) examines a complementary capability by evaluating whether agents can construct goal-oriented judicial analyses and argue both sides of a dispute, testing how legal authorities are assembled into support for competing positions. Deterministic Legal Agents (de Martim, 2025) instead structures research over temporal legal knowledge graphs, allowing an agent to plan and execute point-in-time retrieval, context reconstruction, provenance tracing, and impact analysis through auditable primitives. For long legal agreements, PAKTON (Raptopoulos et al., 2025) combines collaborative agents with retrieval-augmented generation to locate relevant contractual material and produce grounded answers with improved completeness and justification. These approaches move legal research beyond one-shot retrieval by giving agents mechanisms to organize, inspect, and extend the evidence used within an active inquiry.

The current advance is therefore strongest in proactive gap filling within an active inquiry. The legal issue itself is still generally selected and bounded by the user. Agents rarely conclude that new evidence has changed the relevant legal category, that a different jurisdiction must be examined, or that an apparently completed inquiry should be reopened. A more developed research agent would preserve connections among facts, claims, authorities, and unresolved questions across the matter. It could then revise the research agenda when those connections change, moving from self-directed retrieval toward proactive identification of what now requires legal investigation.

6.10.2 Argument Stress-Testing and Adversarial Review

Legal positions are developed in anticipation of parties who may challenge their facts, assumptions, and interpretation of authority. Users can produce a coherent argument while overlooking the objection most likely to affect its practical viability. If that objection is discovered only when raised by opposing counsel, a court, or a regulator, opportunities to collect evidence or revise the position may already be limited. Anticipating consequential challenges is therefore a distinct form of legal proactiveness because it brings future scrutiny into present preparation.

Legal agents support adversarial review through several forms of role separation, critique, and simulated opposition. ChatLaw (Cui et al., 2023b) organizes legal assistance through a multi-agent architecture and law-firm-inspired operating

procedures, introducing structured intermediate roles rather than relying on a single response. LawLuo (Sun et al., 2024a) makes this internal review more explicit through receptionist, lawyer, secretary, and boss agents, with the boss agent evaluating the work produced during multi-round consultation. Courtroom simulation provides a more directly adversarial setting. AgentsCourt (He et al., 2024) combines court debate, legal-resource retrieval, and subsequent decision refinement, while AgentCourt (Chen et al., 2025a) uses structured adversarial interactions to evolve lawyer agents and their legal knowledge. SimCourt (Zhang et al., 2026c) organizes agent interaction around the procedural structure of Chinese trials and evaluates both the simulated procedure and its contribution to judgment prediction. SAMVAD (Devadiga et al., 2025) assigns agents to judicial, prosecution, defense, and adjudication roles and lets them deliberate iteratively over case facts and retrieved legal material. LegalSim (Badhe, 2025) shifts attention from substantive argument to procedural strategy, allowing opposing agents to select actions such as discovery requests and motions and thereby exposing exploitative sequences that emerge during adversarial proceedings. Debate-Feedback (Chen et al., 2025f) uses a more compact debate structure for legal judgment prediction, combining competing agent reasoning with reliability evaluation. MASLegalBench (Jing et al., 2025), rather than introducing another courtroom system, evaluates role-based multi-agent configurations on deductive legal reasoning under GDPR, providing a way to test whether decomposition and specialization actually improve collaborative legal analysis.

This line of work has developed proactive challenge generation more than proactive challenge management. Most systems broaden the argument set within a predefined dispute, but provide limited guidance on which objection should change the course of the matter. Stronger agents would connect each challenge to the available evidence, upcoming decisions, and the consequences of leaving it unresolved. A plausible objection could then trigger factual verification, document revision, or investigation of an alternative legal theory. The next stage is therefore selective and consequential anticipation, in which adversarial review creates and prioritizes further legal work rather than simply generating more arguments.

6.10.3 Legal Matter Management and Workflow Execution

Legal matters progress through dependencies among documents, participants, decisions, and procedural events. Contract review may move through extraction, comparison, negotiation, revision, and approval. Patent preparation may alternate between prior-art research, claim drafting, and technical verification. Due diligence may uncover facts that expand the scope of later review. Since the completion or revision of one component can change what should happen elsewhere, proactive matter management requires the workflow to adapt as the legal state evolves.

Recent systems automate different portions of legal workflows once an objective has been specified. LAW (Watson et al., 2025) orchestrates domain-specific tools and text agents to answer questions over custody and fund-services contracts, including tasks that require intermediate operations such as computing contractual dates. PAKTON (Raptopoulos et al., 2025) similarly coordinates multiple agents and retrieval over long legal agreements, but focuses on document review and grounded question answering. Patent-oriented systems cover a broader range of workflow structures. PatentAgent (Wang et al., 2024j) integrates modules for pharmaceutical patent question answering, chemical-structure extraction from images, and core-structure identification, allowing different analysis operations to be invoked within a unified agent. AutoPatent (Wang et al., 2024f) targets end-to-end document production through a planner, multiple writer agents, and an examiner agent that jointly generate complete patents from initial drafts. AutoSpec (Shea & Yu, 2025) decomposes patent-specification drafting into smaller subtasks and equips open-source models with tools specialized for each stage. Legal consultation systems organize work around participants rather than document stages: ChatLaw (Cui et al., 2023b) encodes standardized operating procedures modeled on legal-service workflows, while LawLuo (Sun et al., 2024a) coordinates receptionist, lawyer, secretary, and supervisory agents across a multi-round consultation. MASER (Yue et al., 2025) addresses orchestration from a different direction by using a multi-agent simulator to generate synthetic data from interactive legal scenarios and introducing MILE to evaluate models in multi-stage legal interactions. Across these designs, agents increasingly select intermediate actions, tools, and roles without requiring the user to specify every step of the workflow.

Current progress, however, concerns proactive sequencing inside a workflow whose boundary has already been established. Systems can decide how to complete an assignment, while rarely recognizing that a change elsewhere has expanded or altered the assignment itself. Advancing toward proactive matter management requires persistent tracking of factual assumptions, unresolved issues, document versions, approvals, and deadlines. A revised clause could initiate renewed compliance review, new evidence could invalidate an earlier draft, and an approaching procedural event could call for immediate preparation. The key development is from autonomously executing known dependencies to detecting

when the evolving matter has created new ones.

6.10.4 Compliance Monitoring and Legal-Risk Anticipation

Compliance monitoring comes closest to proactive discovery of new legal issues because the relevant problem may not yet be part of the legal agenda. A regulatory amendment, modified data practice, or contractual change may appear routine in isolation while their interaction creates an obligation or exposure that no user has recognized. Compliance-oriented systems examine policies, contracts, data practices, and operational processes for such conflicts.

Recent compliance systems address different stages between assembling relevant organizational context and evaluating it against legal requirements. PACT (Fang et al., 2025) provides the information layer by linking heterogeneous enterprise artifacts in an embedding-based graph, allowing compliance agents to recover privacy-relevant connections that would otherwise remain distributed across separate resources. Compliance Brain Assistant (Zhu et al., 2025b) builds agentic investigation on top of enterprise compliance data: its full-agentic mode can invoke multiple tools and APIs to discover additional context needed to answer complex compliance queries. For data pipelines, the approach proposed by Videsjorden et al. (2025) decomposes compliance assessment into agent tasks that extract and label pipeline artifacts and evaluate them against applicable legal requirements according to factors such as actor role and system risk. The multi-agent legal verifier of Nguyen et al. (2025) divides data-transfer compliance checking into statutory interpretation, business-context evaluation, and risk assessment before synthesizing a final determination. The governance-oriented assistant of Mamalis et al. (2024) takes a narrower legal-information approach, grounding interactions in EU legislation and GDPR case law to support queries over regulatory material. These systems therefore range from recovering compliance-relevant context to carrying out structured legal verification, while continuous detection of newly emerging risks remains less developed.

This application has progressed beyond work on a fully articulated legal question, yet its proactiveness is commonly activated only after relevant materials have been submitted for inspection. Continuous risk anticipation would connect legal developments with organizational changes over time, recognize when their interaction creates material exposure, and decide whether the resulting issue warrants notification, investigation, or escalation.

6.10.5 Evaluation, Challenges, and Open Directions

Agents for the law domain have shown promising signals of developing proactiveness: legal research agents can identify missing support, adversarial systems can introduce unrequested challenges, workflow agents can select intermediate actions, and compliance systems can surface risks in submitted materials. However, proactiveness is less developed when the agent needs to revise the legal agenda itself. Few systems continuously connect changes in facts, authorities, documents, and procedural status to recognize that a new legal issue has emerged or that an earlier issue now requires different work.

Evaluation should therefore extend beyond answer accuracy and completion of predefined workflows. It should test whether an agent detects legally consequential changes without an explicit request, identifies the work those changes create, reopens analysis whose assumptions no longer hold, and prioritizes intervention according to urgency and consequence. Both omission and over-intervention matter: missed issues may leave obligations, deadlines, or vulnerabilities outside the legal agenda, while excessive escalation can overwhelm professional review. Legal agents have thus developed meaningful proactiveness within identified issues, while proactive formation and revision of the legal agenda remains the central frontier, with attorneys continuing to control legal judgment and consequential actions.

6.11 Agents for Education

Proactiveness has far-reaching implications in education because effective support often begins before learners can recognize or articulate their own needs. Educational agents assist through explanation, feedback, assessment, resource recommendation, and the organization of instructional activities, yet the value of these actions is determined by their effects beyond the current interaction. A correct answer may conceal fragile understanding, while an incorrect one may reflect a temporary slip, a missing prerequisite, or a misconception whose consequences will emerge only in later material. Learners may know that they are struggling without knowing what has gone wrong, what support would help, or when intervention would be most valuable. Support triggered only by explicit requests or observed failure may therefore arrive too late to prevent downstream learning difficulties.

Table 13 Representative educational systems and models supporting proactiveness. Res. denotes research systems; Com. denotes commercial products.

System	Awareness	Anticipation	Agenda	Arbitration	Action	Adaptation	Deploy.	Type
From Correctness to Comprehension (Zhang et al., 2025h)	Student responses, work	Latent understanding beyond correctness	Assessment task	Evidence-based state assessment	Error classification and tailored feedback	Uses longitudinal evidence to refine assessment	Software	Res.
Classroom Simulacra (Xu et al., 2025b)	Longitudinal learner interactions	Future learner behavior	Simulation objective	Simulation of possible learner trajectories	Future-answer correctness estimates	Models behavior evolution over time	Simulation	Res.
StudentSim (Yang et al., 2026d)	Sparse learner-specific records	Individual future responses	Tutor-improvement objective	Simulator-based tutor optimization	Simulated learner responses	Models response changes under tutor guidance	Simulation	Res.
ProPACT (Golrang et al., 2026)	Shared attention, mental effort	Emerging suboptimal learning states	Collaborative-learning task	Forecast-based intervention decision	Adaptive scaffolding	Adjusts support as learner state changes	Software	Res.
OATutor (Pardos et al., 2023)	Learner interaction, performance	Estimated need for support	Tutoring objective	Adaptive support selection	Hints and instructional support	Adapts support during tutoring	Software	Res.
Environment Shaping (Zhang et al., 2026k)	Multimodal learner behavior	Learner-state changes	Learning-task objective	State-based intervention selection	Visual environment modification	Changes visual support with inferred state	Software	Res.

This mismatch between observable behavior and future learning needs defines the central role of proactive educational agents. By maintaining an evolving account of the learner, an agent can identify needs that have not been expressed, anticipate where they may affect subsequent progress, and alter instruction before avoidable difficulties accumulate (Chu et al., 2025). Because premature or excessive assistance can weaken independent reasoning during learning, effective support also depends on recognizing when observation, delay, or withdrawal will better serve the learner’s development.

Research has developed these proactive capabilities at different temporal scales. Learner modeling and error analysis support the early discovery of latent needs and, increasingly, the anticipation of how those needs may affect later learning. Adaptive tutors use such evidence to regulate an ongoing activity without waiting for an explicit request. Current systems have made substantial progress in diagnosing learner states, anticipating some future difficulties, and initiating local pedagogical actions, but their behavior usually remains bounded by individual tasks, activities, or planning episodes. We therefore review two main lines of work: learner-state diagnosis and anticipation, and adaptive tutoring and formative scaffolding. Table 13 compares selected educational systems and supporting models across the six mechanisms.

6.11.1 Learner-State Diagnosis and Anticipation

The first step toward proactive educational support is recognizing a consequential need before the learner clearly reports it. Dialogue, assignments, worked solutions, and interaction histories provide partial evidence about understanding. Diagnosis therefore requires inferring the learner state underlying the observed behavior rather than treating correctness alone as sufficient evidence. Moreover, the instructional significance of that evidence is rarely immediate. Thus beyond identifying the learner’s current state, proactive support also requires judging whether an observed difficulty is likely to affect later learning.

Knowledge-tracing methods address the corresponding inference problem in real interactions by identifying relevant knowledge components in tutor–student dialogue and tracking student knowledge across the interaction (Scarlatos et al., 2025a). Related multi-agent methods assess whether teachers’ responses in professional-development settings demonstrate targeted mathematical content knowledge (Yang et al., 2024b). Student artifacts provide a complementary source of evidence by revealing where understanding begins to break down. From Correctness to Comprehension (Zhang et al., 2025h) combines multimodal student work with longitudinal records to support error classification and tailored feedback, while AI-driven Virtual Teacher (Xu et al., 2025c) analyzes student solution drafts to identify and correct mathematical errors. Multi-agent automatic feedback systems (Guo et al., 2024) and tools for open-ended responses (Maltelsky et al., 2023) further support assessment and feedback from richer forms of student work. SEFL (Zhang et al., 2026d) complements these approaches by using teacher–student LLM agents to simulate assignment completion and formative feedback, producing synthetic pairs of student work and corresponding feedback. MathAgent (Yan et al.,

2025a) localizes and categorizes errors in multimodal mathematical solutions, while ErrorRadar (Yan et al., 2026d) benchmarks models on error-step identification and error categorization. AI-assisted CS50 teaching (Liu et al., 2024c) and teachable-agent systems for programming education (Jin et al., 2024b) expose code-level misunderstandings, while Closing the Loop (Nair et al., 2024) improves writing feedback by optimizing it against simulated student revisions, contestable essay-feedback frameworks (Hong et al., 2025) support interactive feedback and evaluation, and EvaAI (Lagakis & Demetriadis, 2024) uses multiple agents for automated grading.

Diagnosis becomes more challenging when the relevant need has not yet produced an observable failure. A learner may perform adequately on the current activity yet still be at risk of difficulty when a prerequisite is reused, when previously learned material decays, or when the curriculum moves to a more demanding task. In such cases, present evidence must support not only an estimate of current understanding but also an assessment of how that understanding may affect later learning.

Recent work begins to provide this anticipatory capability through learner and classroom simulation. Classroom Simulacra (Xu et al., 2025b) models students' learning behavior from longitudinal interactions with course materials and simulates how that behavior evolves over time. TeachTune (Jin et al., 2025) uses simulated students with prescribed knowledge and motivational characteristics to help teachers review pedagogical conversational agents across different learner profiles before deployment. StudentSim (Yang et al., 2026d) moves this line toward per-student specialization, training individualized simulators from sparse per-learner records that simulate both a student's own responses and how those responses change under tutor guidance; as a proof of concept, it further uses the simulator as a reward for reinforcement-learning-based tutor improvement. Related systems explore simulation for teacher preparation: Evelyn AI (Bhowmik et al., 2024) provides an LLM-powered virtual student with whom pre-service teachers can practice, while LLM-enhanced teaching simulations (Zheng et al., 2024b) investigate simulation-based teacher training and the authenticity of these simulated experiences. Although these systems do not themselves plan a learner's future trajectory, they make it possible to examine learner responses and instructional strategies before corresponding real interactions occur.

Other work connects such anticipation more directly to instructional planning. (Hu et al., 2025a) simulate teacher-student interactions from an existing teaching plan, generate reflections from the simulated process, and use them to refine the plan before classroom use. Here, anticipated problems can change future instruction before real students encounter them. ProPACT (Golrang et al., 2026) uses multimodal measures of shared attention and mental effort to forecast emerging suboptimal states in collaborative learning and trigger adaptive scaffolds before coordination breaks down.

Current systems therefore provide substantial support for estimating learner states and, increasingly, for anticipating how those states may affect subsequent learning. However, these capabilities usually remain bounded to particular interactions, activities, or planning episodes. A more general proactive educational agent would need to maintain learner-relevant information over time, track whether predicted needs remain unresolved, and update its assessment as new evidence arrives. Such continuity would provide a stronger basis for deciding when and how subsequent pedagogical actions should be taken.

6.11.2 Intervention: What should the system do now?

Once a potential need has been identified, the next stage of proactiveness concerns whether and how the tutoring agent should alter the current learning process. The same visible learners' difficulty can support different decisions. A learner who is productively exploring may benefit from additional time, whereas another learner may be repeating an ineffective strategy that will make later steps harder. The agent therefore needs to interpret the difficulty in relation to the learner's history, the current task, and the likely consequences of continued effort. These judgments turn learner-state estimation into active regulation of an unfolding activity.

Adaptive tutoring systems already support several forms of proactive intervention within an ongoing learning activity. Existing systems can initiate pedagogical actions, choose among several forms of support, and adjust assistance within a defined activity. OATutor (Pardos et al., 2023), conversational tutoring systems with student modeling (Park et al., 2024), personalized tutoring agents (Liu et al., 2024a), and private tutoring frameworks based on chained LLMs (Chen et al., 2024c) adapt instructional responses to estimated learner needs. Depending on the available evidence and system design, they can adjust instructional strategies, guidance, learning activities, or assessment during the tutoring process. Resource-oriented systems extend this action space. FOKE (Hu & Wang, 2024) combines learner profiling with tailored guidance

and learning-path planning, while recommendation chatbots with knowledge-graph contextualization (Abu-Rasheed et al., 2024a) and knowledge-graph-based explanation systems (Abu-Rasheed et al., 2024b) help learners understand and interact with learning recommendations. (Zhang et al., 2026k) extend the action space beyond dialogue to the visual environment itself, inferring learner states from multimodal behavior and proactively modifying the visual learning environment through interventions such as adaptive highlighting. These mechanisms allow the agent to modify the conditions under which learning proceeds, moving beyond the interpretation of a completed response. However, these adaptive tutoring systems’ reasoning usually remains bounded by the predefined task scope, and the consequences of an intervention rarely reshape decisions beyond that setting. The field has progressed from responsive answer generation toward local proactive regulation, while initiative to form help across sessions remains limited.

6.11.3 Evaluation, Challenges, and Open Directions

Evaluation should distinguish two questions: whether a proactive intervention was warranted, and whether it ultimately benefited the learner. The first concerns the quality of the agent’s decision to intervene. Because proactive tutors act before an explicit request, they can intervene too early, too late, unnecessarily, or not at all when support is needed. Premature assistance may interrupt productive struggle or disclose information that the learner could have derived independently, whereas missed or delayed intervention may allow confusion and frustration to deepen. Evaluation should therefore assess not only whether an agent detected a learner need, but whether the evidence available at the time justified intervention and whether the timing and intensity of support were appropriate. The second question concerns educational benefit over a longer horizon. Immediate task performance is an insufficient measure because assistance can improve the current answer while weakening retention, transfer, independence, or engagement. More informative outcomes include delayed unaided performance, retention, transfer to new problems, and whether support can be withdrawn as competence develops. Where feasible, evaluation should also consider longer-term learner development, including whether the agent preserves productive effort, confidence, persistence, and curiosity rather than optimizing only for short-term correctness.

A central challenge for proactive educational agents is that the learner state required to make these decisions is only partially observable, while appropriate intervention is highly learner- and context-dependent. On one side, the same behavior can reflect very different underlying conditions: a long pause may indicate confusion, productive reasoning, distraction, or deliberate exploration, while an incorrect answer may result from a misconception, uncertainty, or a transient mistake. Beyond knowledge and skill, intervention decisions may also depend on harder-to-measure factors such as confidence, motivation, persistence, frustration, and curiosity. Inferring these states reliably from noisy interaction traces remains difficult, yet errors in this inference directly affect whether proactive support helps or disrupts learning. On the other side, the same hint, explanation, or delay can support one learner and hinder another, making it difficult to derive general intervention policies from isolated episodes. Progress will require methods that learn from longitudinal experience while distinguishing transferable educational patterns from learner-specific preferences and trajectories. Such methods must also operate under strong privacy and governance constraints, because richer proactive support often depends on maintaining sensitive learner histories over time. More capable systems will therefore need not only better prediction, but better calibrated decisions about when to intervene, how strongly to intervene, when to continue observing, and when to defer to learners or educators.

6.12 Agents for Entertainment and Media

The value of proactiveness in entertainment and media is experiential rather than task-completing, depending on timing, presence, taste, and a sense of authorship. An intervention that is well timed deepens immersion; the same intervention mistimed breaks it, so the cost an agent must weigh is the experience itself rather than a failed task. Today this proactiveness optimizes the current session: what operates at scale is content recommendation and auto-play triggered by the current viewing or listening context, with in-world characters and creative tools that are only beginning to act on their own, all anticipating the next moment well but holding no objective that spans sessions unless the user writes one, as in a prompt-defined playlist that refreshes on a schedule (Spotify, 2026b). Research points toward anticipatory immersion agents that work over longer horizons, shaping the pacing of a long-form narrative, coordinating co-created content across many sessions, and adjusting intensity under a standing objective such as sustained engagement without overload. Entertainment is therefore rich in deployed anticipation but early in agenda relative to the other verticals here, and its distinctive constraint is that stronger initiative is felt directly in immersion and authorship rather than in whether a task succeeds. We organize the discussion around three roles: acting inside an interactive world, curating what a user

Table 14 Representative proactive entertainment and media systems. Res. denotes research systems; Com. denotes commercial products.

System	Awareness	Anticipation	Agenda	Arbitration	Action	Adaptation	Deploy.	Type
Generative Agents (Park et al., 2023)	Simulated world events	LLM + memory	Self-generated plans	Plan / reflect gating	Initiated interactions (<i>action</i>)	Memory, reflection	Simulation	Res.
PUBG Ally (Kim et al., 2026a)	Game state, player speech	On-device LLM	Support the player; in-match commitments	Per-event reactivity priors, cooldowns	Speech, movement, combat (<i>action</i>)	Retraining on live play	Game	Com.
Spotify DJ (Spotify, 2023)	Listening history, feedback	Recommender + LLM commentary	Current listening session	Curation / timing	Lineup + commentary (<i>suggestion/action</i>)	Feedback revision	App	Com.
Firefly AI Assistant (Key, 2026)	Project assets, app state	LLM agent	Creator’s described outcome	Asks along the way; user can take over	Multi-step edits (<i>gated action</i>)	Learns creative preferences	App	Com.
RealJam (Scarlatos et al., 2025b)	Live musical input	Anticipatory model	Current performance only	User-set silence, lookahead, and commitment windows	Musical accompaniment (<i>action</i>)	—	Software	Res.

consumes, and co-creating content with the user. We close with how such agents are evaluated and the risks that make calibrated initiative essential here. Table 14 summarizes selected systems spanning these three roles.

6.12.1 Acting Inside the World: Game and Embodied Agents

Some of the oldest proactive agents in entertainment act inside interactive worlds. Interactive drama introduced the drama manager, an agent that watches an unfolding story and intervenes to shape pacing and coherence without being asked, realized in *Faade* and developed across the interactive-narrative literature (Mateas, 2003; Riedl & Bulitko, 2013). Believable autonomous characters complement it, pursuing their own goals and acting in the world rather than only replying when addressed. Foundation models scale this from authored behavior to open-ended initiative: Generative Agents shows simulated characters that observe events, store memories, form plans, and start interactions with no user scripting each step (Park et al., 2023), and the pattern is entering production, where NVIDIA ACE supplies speech, reasoning, and animation for interactive characters and its companions, such as a PUBG ally that advises, shares loot, drives, and fights alongside the player, act on their own within the game (NVIDIA, 2026a; Burnes et al., 2025); that ally has since shipped in a live-service beta that sets explicit per-event priors on whether to speak or act (Kim et al., 2026a). This gives designers a ladder of initiative from commentary, to suggestion, to coordinated play, to direct action; higher rungs enrich a world but demand stronger Arbitration, since an unwanted act can break play, story, or trust.

Embodiment raises the cost of acting. In VR and virtual worlds the agent is experienced as a body, a voice, and a spatial presence, the regime studied for decades as embodied conversational and socially interactive agents (Cassell et al., 2000). Initiative is now expressed through gaze, gesture, motion, and proximity, which raises costs that text interfaces never face: a spoken interruption mid-scene, or an avatar that enters personal space, can feel intrusive in ways a small notification would not (Adamczyk & Bailey, 2004; Bailenson et al., 2003). Well-timed initiative instead strengthens presence when the agent makes its intent legible and respects the rules of the world (Dragan et al., 2013). Recent evaluation of LLM-driven characters in VR reports gains in perceived realism and usability alongside persistent limits in latency and emotional depth (Korkiakoski et al., 2025), and a companion that played a VR puzzle game autonomously lowered enjoyment by reducing challenge and player autonomy (Zhang et al., 2026e). The Arbitration question in this setting is how the agent controls timing, distance, modality, and reversibility so that initiative preserves rather than breaks immersion.

6.12.2 Curating What Users See: Feeds and Recommendation

Outside embodied worlds, the most widely deployed proactive agents in entertainment are media recommenders that decide what to play or show before any query. YouTube and TikTok rank video from interaction history and context, and Spotify’s DJ assembles a personalized lineup with spoken commentary and revises it from feedback (YouTube, 2021; TikTok, 2020; Spotify, 2023, 2026a). We treat the general mechanism, recommendation as proactive filtering together with its conversational and preference-eliciting variants, as information access (§6.3); what is specific to entertainment

is that the target is taste and attention rather than a stated information need. These systems are accordingly strong in Awareness, Anticipation, and Adaptation but narrow in action: they rank and deliver content rather than act within a shared world, and their recurring decision is Arbitration over when to surface the next item, how strongly to frame it, and how easily the user can redirect the session.

6.12.3 Co-Creating Content and Performance

On the production side, agents help make content rather than deliver it, which the human-computer interaction literature frames as mixed-initiative co-creativity, where creative agency is shared and the tool takes initiative as a partner instead of waiting for each instruction (Deterding et al., 2017; Shneiderman, 2007). Deployed creative assistants follow this shape: Adobe’s Firefly assistant plans and runs multi-step edits across creative applications from a described outcome, and creator tools suggest ideas, titles, and thumbnails (Key, 2026; YouTube, 2026). Their proactive value is in maintaining context, proposing next steps, and preparing editable candidates rather than a finished asset, and those candidates have to be worth surfacing, which favors non-obvious, divergent options over literal completion, a leap-of-thought rather than chain-of-thought style of generation (Zhong et al., 2024; Huang et al., 2025b). Live performance makes timing the central problem. ReaLJam anticipates future musical input and revises uncommitted future chords, keeping committed chords fixed (Scarlatos et al., 2025b). Users set the initial silence period and the lookahead and commitment horizons. Live Assistant learns intervention decisions for live streams, choosing whether to intervene, when, and whom to address (Gao et al., 2026e).

6.12.4 Evaluation, Challenges, and Open Directions

Because the value is experiential, evaluation must look past task success to immersion (Jennett et al., 2008), acceptance, and the cost of interruption (Adamczyk & Bailey, 2004), and the risks here are rarely physical. They are broken immersion, spoilers, unfair in-game assistance, weakened authorship and a narrowing of creative diversity (Doshi & Hauser, 2024), and misuse of a creator’s likeness or style (Bariach et al., 2024). The risk most specific to media is that engagement is a poor proxy for enjoyment: optimizing it can amplify content that users say makes them worse off while failing to track their stated satisfaction (Milli et al., 2025). The broader harms of choosing what a user sees, such as filter bubbles, degenerate feedback loops, advertising incentives, and attention-capturing dark patterns, follow the same logic as in information access (§6.3). The lesson is a graded initiative contract: proactive output is safest as a suggestion, cue, preview, or rehearsal that the user can accept or drop, with stronger action reserved for review. Several open directions are specific to entertainment. Immersion agents could work over longer horizons, shaping long-form pacing and coordinating co-created content across sessions rather than resetting each one. Embodied and live-performance settings need real-time timing and legibility methods, so that the agent signals its intent before it acts. As the tool takes more initiative, authorship, likeness, and creative diversity need explicit safeguards. Evaluation should also reward stated satisfaction over raw engagement. In entertainment, mature proactiveness is not intervening more often or forcing deeper immersion, but timing each intervention through an Arbitration gate (§4.4) so that it deepens the experience and leaves play, taste, and authorship with the user.

6.13 Agents for Agriculture

Agricultural decisions often need to be made under uncertain environmental conditions and within time-sensitive biological growth cycles, before problems become fully visible, making proactive intervention particularly important. Water stress, disease pressure, and weather risks may already be developing when symptoms remain weak, while delayed intervention can remove cheaper or safer options. Proactive agricultural agents can therefore monitor relevant conditions, such as crops, soil, and weather, and determine when additional measurement, advice, or intervention is needed, or when existing operations should be adjusted.

In agricultural applications, proactiveness takes different forms depending on the task. Field sensing and management support timely monitoring and resource allocation; pest and disease management enables early risk detection and intervention; and operational coordination helps adapt resource allocation and task schedules to changing conditions. Existing systems address these needs within specific agricultural settings, with varying degrees of autonomy and integration. Table 15 compares selected agricultural systems across the six mechanisms.

Table 15 Representative proactive agricultural-agent systems. Res. denotes research systems; Com. denotes commercial products.

System	Awareness	Anticipation	Agenda	Arbitration	Action	Adaptation	Deploy.	Type
ASA (Pardo-Pina et al., 2026)	Greenhouse telemetry, analytics	Condition changes requiring attention	Greenhouse-management objective	Agent-based integration of sensing and RAG support	Management recommendations (<i>suggestion</i>)	Updates support from new telemetry and retrieved context	Software	Res.
Agri-SAGE (Balasubramaniam et al., 2026)	Field context, crop state	Consequences under simulated crop dynamics	Agronomic-management objective	Multi-agent reasoning + APSIM validation	Context-aware agronomic advice (<i>suggestion</i>)	Refines advice using simulation outcomes	Software	Res.
Community Irrigation MAS (Chiewchan et al., 2022)	Crop water needs, resource availability	Water scarcity and allocation consequences	Shared irrigation objective	Negotiation among farm agents	Water redistribution (<i>action</i>)	Allocation changes with demand and scarcity	Simulation	Res.
PestMA (Shi et al., 2025b)	Pest query, external evidence	Management risks and evidence gaps	Pest-management objective	Editor–Retriever–Validator coordination	Pest-management recommendations (<i>suggestion</i>)	Revises recommendations through retrieval and validation	Software	Res.
PDD-Agent (Qin et al., 2025a)	Multimodal crop observations	Potential plant-disease state	Disease-diagnosis objective	Orchestration of specialized diagnostic tools	Disease diagnosis and management support (<i>suggestion</i>)	-	Software	Res.
UAV Pest Management (Chu & Bao, 2025)	Real-time pest observations	Pest distribution and task demand	Pest-monitoring objective	Swarm path planning and task allocation	UAV-based pest monitoring (<i>action</i>)	Reallocates paths and tasks as conditions change	Physical	Res.

6.13.1 Field Sensing, Management, and Operational Coordination

Agricultural agents can support field monitoring by maintaining an updated view of field conditions. AgriSentinel (Mylay et al., 2025) combines image-based crop-disease classification with on-device management suggestions. The distributed precision-agriculture framework in (Davcev et al., 2026) connects heterogeneous sensing with agent-based reasoning and adaptive decisions, while the Agent-Based Service Architecture (ASA) (Pardo-Pina et al., 2026) organizes greenhouse telemetry, analytics, and RAG-grounded decision support within a shared agent architecture. The wireless-sensor-network architecture in (Villarrubia et al., 2017) further integrates distributed sensing with multi-agent monitoring for crop irrigation.

Beyond monitoring, agricultural agents can support management decisions using available field information. Agri-SAGE (Balasubramaniam et al., 2026) couples multi-agent LLM reasoning with APSIM crop simulation to generate and validate context-aware agronomic advice, while AgroAskAI (Cantonjos & Biswas, 2025) coordinates role-specialized agents and external tools for locally grounded climate-adaptation support. The irrigation-and-fertigation framework in (Adamo et al., 2025) models resource management as sequential decisions driven by real-time IoT measurements. More broadly, the agricultural-agent framework in (Li et al., 2026k) organizes perception, knowledge and memory, planning, action, and feedback into a closed-loop view of agricultural intelligence. These systems illustrate different approaches to supporting agricultural decisions based on observed conditions, simulation, and external information.

Agricultural operations may also require coordination across multiple agents, resources, or activities. The community-irrigation multi-agent system in (Chiewchan et al., 2022) models farms as agents that estimate crop water needs and negotiate water redistribution under scarcity. The greenhouse-production multi-agent system in (Ferentinos et al., 2005) coordinates agents responsible for root-zone and aerial environments through supervisory-level negotiation. The agri-robot fleet coordination approach in (Lujak et al., 2020) studies distributed and decentralized coordination for heterogeneous agricultural vehicles and dynamic task assignment.

These works demonstrate sensing, management, and coordination within specific agricultural applications. However, their capabilities remain largely tied to particular tasks and operational settings. An open direction is to investigate how changes in field conditions, resource availability, or operational constraints can inform subsequent decisions and adjustments when coordination across activities is needed.

6.13.2 Agricultural Risk Detection and Intervention

Pest and disease management requires timely detection and intervention, as emerging threats may become easier to identify while the available intervention window narrows. PestMA (Shi et al., 2025b) uses Editor, Retriever, and Validator agents to retrieve external pest-management evidence, synthesize recommendations, and validate them. PDD-Agent (Qin et al., 2025a) uses a multimodal agent to orchestrate specialized tools for plant-disease diagnosis. The UAV-based pest-management system in (Chu & Bao, 2025) connects real-time pest classification with swarm path planning and task allocation, extending threat recognition toward deployment decisions. The smart-farming agentic framework in (Murad et al., 2026) further integrates soil, weather, disease-detection, and farmer-facing components within a broader decision pipeline.

Current systems mostly begin from an observed image, query, sensor condition, or external signal. Stronger proactive agents could combine such evidence with crop stage, forecasts, exposure, and prior observations to anticipate emerging risks and initiate appropriate actions, such as targeted inspection, increased monitoring, early warnings, or preventive treatment, while accounting for uncertainty and intervention costs.

6.13.3 Evaluation, Challenges, and Open Directions

Agricultural agents should be evaluated by whether their proactive behavior leads to timely and appropriate decisions as conditions evolve. Recommendation accuracy remains important, but it does not capture whether an agent acted early enough to be useful, intervened when intervention was actually warranted, or refrained from acting when evidence was insufficient. Evaluation should therefore consider lead time, unnecessary interventions, escalation quality, resource use, and adaptation to subsequent observations. These criteria are application dependent: a disease-management agent should identify emerging risks early without triggering unnecessary treatments, whereas an irrigation-management agent must balance timely intervention against water consumption. Importantly, inaction can itself be an appropriate proactive decision when the expected benefit of intervention is low or uncertainty remains high.

A central challenge is determining when an agent should initiate action without an explicit task request. Agricultural environments generate continuous streams of weak, noisy, and sometimes conflicting signals, while the consequences of acting too early and too late can both be substantial. Proactive agents therefore need to relate observations to longer-term operational objectives, recognize when evolving conditions create a meaningful need for attention, and decide whether to monitor, gather additional information, alert a user, or intervene. This requires more than executing predefined workflows: the agent must infer when a latent objective has become actionable and calibrate the scope and urgency of its response under uncertainty and resource constraints. Open questions include how such initiative should be grounded in persistent goals, how intervention thresholds should adapt across farms and seasons, and how agents can remain useful without becoming unnecessarily intrusive or overactive.

6.14 Cross-Industry Trends

The preceding verticals of proactive AI agents differ in many important ways. For instance, they differ in their data, their tolerance for error, their regulators, and the horizon over which their consequences arrive. However, despite these differences, three patterns recur across domains (Figure 15). Many systems now use a common architecture, with foundation-model cores and domain-specialized machinery (§6.14.2). Across these domains, agents are also being asked to carry whole workflows, not just respond within them (§6.14.3). Progress still differs by domain. A consistent divider is the cost and quality of feedback on whether a self-initiated action was correct (§6.14.4). The remaining subsections expand on these ideas: First, we overview the relative maturity of proactivity in each vertical. Then we expand on each of the three trends. The per-vertical tables in the preceding subsections summarize the systems on which these observations rest, and Table 16 places each domain on a common ladder.

6.14.1 Proactive Maturity Across Domains

In this section, we evaluate the maturity of proactivity in systems across the many previously surveyed verticals. In service of this goal, we start by reiterating a shared measure of proactivity. We place the surveyed systems on a common informal scale of initiative, distinguishing reactive, trigger-initiated, forecast-initiated, mandate-directed, and agenda-generative behavior. The scale asks where the agent’s task comes from and tracks the anticipation-to-agenda progression of the mechanism chapter. A *reactive* system executes an instruction or a user-authored rule. A *trigger-initiated* system notices a condition that has already occurred and surfaces it. A *forecast-initiated* system acts on a latent need or failure

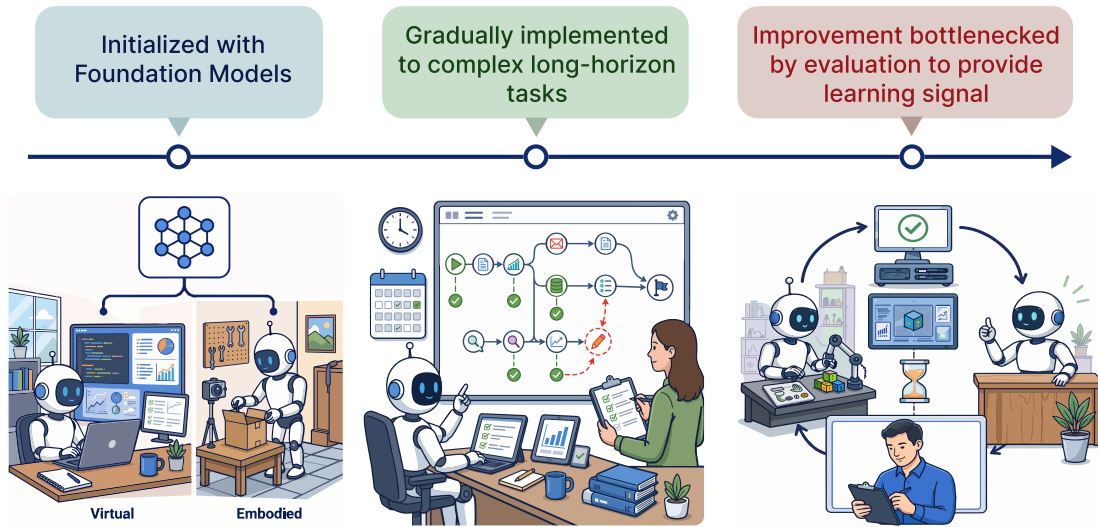


Figure 15 Three cross-industry trends in proactive AI. Foundation models increasingly serve as shared cores complemented by domain-specific models, tools, and infrastructure. Applications are expanding toward complex, long-horizon workflows that require agents to discover and organize work left implicit by delegated objectives. Further improvement is constrained by verification and feedback infrastructure, whose cost, speed, and reliability determine how effectively agents can learn from self-initiated actions.

before it manifests, which is the step that Anticipation supplies (§4.2). A *mandate-directed* system holds standing objectives and discovers, prioritizes, and revisits the work that advances them, and an *agenda-generative* system revises those objectives themselves; these last two are the committed-goal and goal-generation steps of Agenda (§4.3). The scale is a discretization of the reactivity–proactivity spectrum of §2 rather than a separate construct, and, as that section stresses, a single system can occupy different points on it depending on time scale, confidence, and operational context.

The six-capacity framework connects the discovery and organization of candidate work through Anticipation and Agenda with decisions about intervention and execution through Arbitration and Action. We therefore read each domain on a second, orthogonal axis, the *authority* the agent is granted over its first move, which is the levels-of-autonomy spectrum of §4.5.1: whether it may only suggest, may act inside a gate such as a sandbox, branch, simulator, or approval workflow, or may act unsupervised. The two axes move independently. Autonomous-research systems can combine deep execution autonomy with task-scoped initiative under a human-specified top-level question (Lu et al., 2024; Xing et al., 2026); smart-home platforms can initiate a large share of interactions with very little authority behind any single act (Eberhardt, 2022). Each cell reports in Table 16 where the systems surveyed in that domain’s subsection most consistently sit on the initiative ladder, and a hyphen between two rungs means the surveyed set spans both.

Three comparative observations follow from the table. First, deployment and initiative are decoupled. The home is deployment-rich but initiative-shallow, while the city combines forecast-initiated assistance with limited action authority: municipal forecasting is excellent and closed-loop municipal action is rare, because authority there is diffuse and every action is a public act (§7.2). Autonomous research is the extreme case, autonomy-deep and barely deployed. Second, no vertical has reached the agenda-generative rung, and several argue that it should not. For example, in security, healthcare, law, and education the surveyed systems converge on the same target, *mandate-directed initiative under bounded authority*, in which greater proactiveness expands what the agent notices, investigates, and prepares while consequential acts remain gated by policy and professional responsibility (Anthropic, 2026d; Gallifant et al., 2025). The frontier in these domains is the quality of initiative, not the scope of unsupervised authority. Third, feedback conditions differ substantially across domains. Coding agents can often test proposed changes before deployment, whereas feedback from physical experiments and seasonal or civic interventions can be slower, costlier, or more diffuse. These differences affect how readily agents can verify and revise self-initiated work. Model capability, execution reliability, and governance also constrain progress, as the maturity table indicates. This is the subject of §6.14.4.

Table 16 A proactive-maturity map of the surveyed domains. *Initiative* places each domain on a five-rung ladder that tracks the anticipation-to-agenda progression: reactive → trigger-initiated → forecast-initiated → mandate-directed → agenda-generative (§6.14.1); a hyphen between two rungs means the surveyed set in that domain spans both. *Authority* is the orthogonal levels-of-autonomy axis of what the agent is permitted to do on its first move (§4.5.1). *Binding constraint* names the dominant limiter on further progress, which may be a capability, feedback, governance, or evaluation gap depending on the domain.

Domain	Initiative	Authority	Next milestone	Binding constraint
Coding	Forecast-mandate	Gated action	Repository custodian under standing objectives	Incomplete oracles; manual review budget
GUI agents / CUA	Forecast-mandate	Gated action	Persistent cross-app operator with bounded delegation	Atomic operation proficiency; privacy; permissions
Information access	Forecast	Suggestion	Persistent, revisable model of evolving user info needs	User model; benchmark coverage of appropriate silence
Embodied / robotics	Reactive-mandate	Gated action	Mutual-cognitive partner inferring human intent	Latency; physical irreversibility; generalization
Security	Trigger-forecast	Gated action	Mandate-directed defense; reversible pre-positioning	Verification and remediation throughput
AutoResearch	Mandate	Gated action	Agent-originated research agendas with resource-aware design	Novelty and significance; cost and latency of the physical oracle
Smart home	Trigger-forecast	Acceptance-gated action	Ambient orchestration under standing objectives	Arbitration: will the act be welcome?
Smart city	Forecast	Suggestion	Supervised anticipatory infrastructure operations	Diffuse authority; public accountability
Healthcare	Trigger-forecast	Clinician-gated	Longitudinal, multi-visit proactive care	Evidence standards; liability
Law	Trigger-forecast	Attorney-gated	Strategy-aware assistance across a matter's lifecycle	Professional responsibility; confidentiality
Education	Trigger	Suggestion	Trajectory-aware learning-path planning	Slow, noisy outcome feedback
Agriculture	Trigger	Gated action	Season-aware coordination and preventive intervention	Seasonal feedback latency
Entertainment	Forecast	Suggestion / action	Anticipatory immersion across sessions	Noisy user satisfaction feedback

6.14.2 Trend 1: A Shared Architecture with Domain-Specific Specialization

Across these verticals, proactive systems are increasingly built on a common architecture: a foundation-model core paired with domain-specialized machinery. In most domains surveyed here the reasoning core is a large language or multimodal model: coding agents run an LLM loop over the repository (Anthropic, 2025; OpenAI, 2025a; GitHub, 2025); GUI agents drive an MLLM over screenshots (Hong et al., 2024; Kong et al., 2026); autonomous-research systems wrap an LLM in tools and experiment loops (Boiko et al., 2023b; Lu et al., 2024; Gottweis et al., 2026); and legal, medical, and urban agents are typically one or more LLMs over domain tools and retrieval (Wang & Yuan, 2025; Tang et al., 2023; Li et al., 2024d). The value of the shared core is specific and worth naming: parameterized world knowledge is what lets a single system anticipate a latent need (§4.2) in open-world situations without the domain-specific engineering that earlier paradigms required, and most systems reach workable behavior with no task-specific training. Anticipation, historically the expensive part of a proactive system, has become close to a commodity.

The key architectural differences are everything the domain contributes beneath and around that core, and the boundaries of foundation-model primacy mark where domains tend to differ. Systems without an LLM core cluster at the two data extremes. At the low-data, low-latency end a narrow quantitative signal suffices and nothing heavier is warranted: the Observe-Ask-Intervene agent gates on measured speaking time (Houtti et al., 2025), and WorkFit acts on smartwatch thresholds (Ahire et al., 2024). At the high-data end, domains with enormous proprietary corpora train bespoke models rather than adopt a pretrained prior, as AlphaZero does from self-play (Silver et al., 2017) and Alexa's Hunches do by learning device and home embeddings over fleet-scale logs (Eberhardt, 2022). Even where an LLM is present it usually sits atop a domain-specific predictive substrate rather than replacing it: security systems remain hybrids of LLM reasoning over purpose-built detectors because general-purpose agents still underperform specialized static analysis (Yohn et al., 2026), and cities retain spatio-temporal forecasters under an LLM interface (Li et al., 2024d). The emerging architecture is therefore layered rather than monolithic: a general reasoner supplies anticipation and agenda, a domain model supplies calibrated prediction, and a domain gate supplies authority.

Embodiment adds a second boundary, about latency rather than data. Continuous physical control cannot wait on slow

deliberation, which forces hybrid fast–slow designs: ProAct couples a fast behavioral stream to a slower LLM cognitive system (Zhang et al., 2026l), industrial robots adjust motion in real time from proximity and physiological signals while a deliberative layer plans above them (Hostettler et al., 2024), and wearable assistants split a light always-on monitor from a heavier reasoner to keep continuous awareness affordable (Chen et al., 2025d; Wen et al., 2025b). Given the standing energy and latency cost of always-on awareness (§4.1), foundation-model cores will likely remain dominant as efficiency improves. The boundary cases are nonetheless a useful corrective, since they show the LLM to be one substrate among several, selected by a domain’s data-and-latency regime rather than by default, and they predict where the shared substrate will arrive last: in domains that are either too thin or too rich in data, or too fast to wait for a token.

6.14.3 Trend 2: From Assisting Humans to Carrying Workflows

A second cross-domain pattern is that agents are increasingly assigned whole workflows instead of individual steps within them. Earlier proactive systems clustered around proactive *suggestion*, such as code completions, health prompts, assistant reminders, recommended items, or clarifying questions, each of which augments a human who remains the operator of the task. What a growing subset of frontier systems now do is carry a workflow that previously required a person: a coding agent turns an assigned issue into a reviewed pull request (OpenAI, 2025a; GitHub, 2025), a GUI agent accepts bounded delegation across applications (Zhao et al., 2025c; Ye et al., 2026c), a research system runs an experimental loop end to end (Boiko et al., 2023b; Szymanski et al., 2023), and a security agent scans, validates, and drafts a patch (Anthropic, 2026d). Workflows can contain latent work that nobody thought to request, arrive with incomplete specifications, and span interruptions. Proactiveness lets an agent identify and pursue additional work that advances a longer-horizon objective.

This shift is visible in the human’s role, which moves along the workflow from operator to collaborator to reviewer. It is most advanced in software, where the developer no longer specifies each edit but defines an outcome, constraints, and acceptance conditions, and then reviews evidence; the same rotation is underway for clinicians and attorneys, whose surveyed systems all place a professional at a review gate rather than at the keyboard. The rotation explains why so many domain conclusions above converge on the same next milestone. Encounter-bounded medicine wants longitudinal responsibility, matter-bounded law wants lifecycle strategy, interaction-bounded tutoring wants trajectory-aware learning paths, and issue-triggered coding wants a repository custodian (Thai et al., 2025; Chen et al., 2026c): in each case the demand is for an agent that holds an objective across time rather than a task within a session.

Crucially, the systems taking consequential first moves are not the least constrained ones. As the authority axis above records, they scale the strength of the first move to risk, reversibility, and confidence (§4.4.4) and engineer the gate before granting the initiative: coding agents act inside sandboxes and review, and Alexa’s Hunches complete actions behind a learned acceptance model. What is spreading, in other words, is not autonomy but initiative that a principal can safely accept, which is why the gate travels with the capability across domains that share nothing else. The workflow shift and the graduated-authority pattern are two faces of the same move: an agent may hold a longer objective only insofar as its acts on that objective stay gated.

6.14.4 Trend 3: Verification and Feedback Infrastructure Is a Key Constraint

The common architecture of §6.14.2 does not produce uniform progress across domains. An important source of variation is how cheaply, quickly, and reliably the environment provides feedback on proactive behavior (§4.6.1). Timely, reliable feedback can support repeated testing and correction, provided agents can execute the relevant actions and interpret the results. Software is the paradigm case, since a compiler, type checker, test suite, linter, or CI run can reject a change before a human sees it, so a coding agent can treat its own edits as verifiable state transitions and self-verify rather than emit ungrounded text (Ning et al., 2026d). The frontier has accordingly moved from generation to closed-loop engineering, in which the agent reviews its own diffs, requests further agent or human review, and iterates until reviewers are satisfied (Lopopolo, 2026; Anthropic, 2026b). Machine-learning research can use a similar feedback loop by running training code and scoring a benchmark, supporting autonomous-research systems that propose and test their own research ideas (Lu et al., 2024; Yamada et al., 2025).

Where the oracle is slow or expensive, repeated testing and correction become harder to sustain without supporting infrastructure. A wet-lab assay or materials synthesis turns every action into a costly, sometimes irreversible bet, so the burden shifts from executing experiments to deciding which experiments are worth running (Boiko et al., 2023b; Szymanski et al., 2023), which connects proactive research to active learning and Bayesian experimental design (Settles,

2009; Chaloner & Verdinelli, 1995). Agriculture and education face the same problem in temporal form, since the outcome of an intervention may not be observable for a season or a semester. Recent engineering addresses these feedback constraints: cloud and self-driving laboratories reduce the latency of the oracle (Armer et al., 2023; Abolhasani & Kumacheva, 2023); research-state infrastructure preserves what experiments returned so a later agent can verify and extend rather than restart (Schmidgall & Moor, 2025; Anthropic, 2026a); and city operators substitute simulation for the costly street trial, rehearsing interventions against synthetic populations in urban digital twins before deploying them (Xu et al., 2024b; Piao et al., 2025). The digital twin and the self-driving lab are the same move made in different domains: buy a cheap oracle so that initiative becomes affordable.

Two implications follow for verification and evaluation. First, capability gains can shift the burden from generating candidate work to checking and completing it. Under Project Glasswing, partners reported over ten thousand high- or critical-severity vulnerability findings within a month, and the binding constraint moved from finding vulnerabilities to verifying, disclosing, and patching them (Anthropic, 2026g,h); a proactive agent that can generate more candidate work than its principal can check has shifted the cost rather than paid it. The second is that our evaluations are not yet measuring the thing that limits us. Benchmarks still largely specify what the agent should work on and then score whether it succeeded, which measures execution under a reactively supplied agenda (Jimenez et al., 2024). A mandate-directed agent must instead be scored on whether it discovered work worth doing, prioritized it correctly, chose a proportionate intervention, and recognized when no intervention was preferable, and the metrics that would support this, such as task-discovery precision and recall, interruption cost, human verification effort, and reward for correct silence, remain scattered and domain-specific (Chen et al., 2026c; Samarinas & Zamani, 2024; Tang et al., 2026b) (§5). Further progress will require advances in model and execution capabilities supported by trustworthy feedback and verification. Evaluation should assess both the value of self-initiated work and whether an agent appropriately refrains from acting.

7 Future Work: Towards Calibrated Proactiveness

Current systems implement proactive capabilities only partially (§ 4 and § 6). The long-term objective is calibrated, governable proactiveness that is useful across realistic contexts. Future proactive agents may become situated partners that can anticipate needs, manage long-horizon goals, and intervene across conversational, multimodal, embodied, and ambient environments, including robotics, household assistance, wearable devices, workplace support, and personalized daily planning. In these settings, proactiveness involves self-initiation, latent-need detection, goal formation, intervention arbitration, and cross-context action. Calibration is needed to determine when initiative is helpful, when clarification is needed, when human approval is required, and when inaction is preferable. Greater initiative may be useful, intrusive, unsafe, or misaligned, depending on its calibration and context.

Calibrated proactiveness calls for connected advances in capabilities, governance, and application development. Technical and evaluative progress should support agents in anticipating needs, initiating actions, and sustaining assistance over long-horizon interactions (§7.1). Governance needs to preserve safety, trustworthiness, and human agency when agents initiate action (§7.2). Application development should consider how calibrated proactive agents may affect products, workflows, organizational roles, and modes of human supervision across domains such as work, finance, education, and healthcare (§7.3).

7.1 Steps to Take for Advancing Proactiveness

7.1.1 Evaluation Beyond Narrow Applications, Static Performance, and Coarse-Grained Metrics

Evaluation needs to cover agents that anticipate needs, manage long-horizon goals, and intervene appropriately across diverse interaction contexts. Relevant contexts include text-based conversations and multimodal, embodied, and ambient environments, including robotics, household assistance, wearable devices, workplace support, and personalized daily planning, as indicated across the domain tables in Section 6. Alongside single-turn response correctness, evaluation should assess whether the agent recognizes useful occasions for intervention, chooses an appropriate level of initiative, preserves user control, and adapts as the user’s goals and environment evolve over time.

Many evaluations treat proactiveness as a turn-level prediction problem in conversational settings (Chen et al., 2025d), where task-specific response skills may be acquired through post-training. Such tests provide limited evidence about self-initiation, as defined in § 2, as a generalizable capability. More general benchmarks, such as ProCIS (Samarinas & Zamani, 2024), Ask-Before-Plan (Zhang et al., 2024h), and ProactiveBench (Lu et al., 2025a), begin to evaluate

proactive behavior in predominantly single-turn and text-based scenarios. As a result, they provide limited evidence about whether proactive agents can generalize across modalities, domains, task horizons, and levels of environmental complexity. Future benchmarks should cover near-term conversational agents and longer-term embodied or ambient agents, with tasks that vary in temporal horizon, user involvement, risk level, and contextual uncertainty.

Many proactive actions occur without a direct user request. Actions that improve task outcomes may be perceived as helpful, intrusive, distracting, premature, or overly controlling depending on their timing, context, and degree of autonomy. Future evaluation should therefore measure task outcomes together with how users experience agent initiative. General user-experience and usability instruments, such as UEQ (User Experience Questionnaire) (Laugwitz et al., 2008) and SUS (System Usability Scale) (Brooke, 1995), can provide useful system-level signals. Targeted measures should assess trust, reliance, perceived control, interruption cost, willingness to accept suggestions, and perceived respect for human agency. Technical correctness can coexist with user-perceived miscalibration.

Task success and user satisfaction provide limited evidence about the mechanisms producing an agent’s behavior. Memorized patterns, alignment-tuned responses, and task-specific shortcuts may influence these outcomes. Future benchmarks should therefore include fine-grained tests for core proactive mechanisms, including self-initiation, world modeling, sub-goal proposal, intervention timing, initiative arbitration, action-level validation, and cross-context generalization. A unit-test-style evaluation could separately probe each mechanism before evaluating the full agent in integrated long-horizon tasks. These tests should assess decisions to act, ask, defer, or remain silent across settings.

A unified benchmark design should combine broader application settings, user-centered metrics, and mechanism-level tests. Together, these components can assess whether agents improve task outcomes and sustain calibrated, governable, and user-aligned proactiveness across realistic contexts.

7.1.2 Strengthening Calibrated Proactiveness Through Mechanism-Level Advances

Toward Unified, Resource-Aware, and Robust Situation Awareness. Maintaining a reliable estimate of the current situation supports anticipation and timely intervention (§ 4.1). Existing work has developed building blocks for this capacity across different settings, including multimodal state tracking and context management (Yang et al., 2025b), continuous personal memory (Wang & Chen, 2025), online video understanding (Chen et al., 2024a), active information gathering (Choudhury et al., 2026), and proactive triggering (Lu et al., 2025a). Future work should develop unified situation-awareness architectures that integrate perception, memory, active information seeking, runtime triggering, tool outputs, dialogue history, screen state, user behavior, and environmental signals into a shared representation of the user, task, and environment that supports anticipation, agenda formation, and arbitration. The shared representation needs compression because proactive agents operate over continuous streams of information that cannot be fully retained in a bounded context window. Agents must decide what to keep, summarize, retrieve, forget, or page into memory. Existing work on streaming language models (Xiao et al., 2024), virtual context management (Packer et al., 2023), and sleep-time computation (Lin et al., 2025) provides early steps in this direction. Future compression mechanisms should be optimized specifically for proactiveness: they should preserve information that may support future anticipation, discard irrelevant details, and retain enough evidence to justify why an intervention was triggered.

Situation-awareness systems should allocate computation, memory, sensing, and interaction according to contextual importance. Candidate approaches include triage, selective attention (Vaswani et al., 2017), information filtering, priority management, adaptive compression, and background or “dream-time” computation during periods of low system demand (Lin et al., 2025). For example, hierarchical awareness systems could use low-cost monitors to continuously tag incoming context according to novelty, urgency, uncertainty, user relevance, risk, and potential need for intervention. When these signals indicate that a moment is important, ambiguous, or high-stakes, the agent could escalate to more expensive processing, such as deeper multimodal interpretation, memory retrieval, active information gathering, tool use, or explicit clarification. In parallel, attention and information-selection mechanisms could restrict processing to the most relevant portions of a long context, while background computation could reorganize memory, consolidate previously observed experiences, or precompute likely future needs without delaying ongoing interaction. Some resource allocation may also occur below the software layer. For instance, on-device sensors, low-power processors, or hardware-level event triggers in mobile and IoT environments (Baccour et al., 2022) could detect salient environmental changes before activating more computationally expensive models. Hardware-assisted mechanisms can provide an efficient first stage for software-level reasoning over large, continuous streams of multimodal data. Together, these complementary mechanisms allow routine context to be compressed, summarized, deferred, or ignored, while high-value moments receive richer

analysis before the agent forms a sub-goal, arbitrates intervention, or initiates action. Compression keeps long-running context usable, and adaptive resource allocation determines which parts of that context require deeper understanding.

Temporal freshness, uncertainty awareness, noise handling, and cross-context coherence affect the quality of situation estimates. Temporal freshness is necessary because proactive agents may act on long-term memories, yet stale preferences or outdated task states can cause errors in unsolicited interventions; long-term memory systems (Zhong et al., 2023), temporal knowledge graphs (Rasmussen et al., 2025), and long-memory evaluation (Wu et al., 2025b) show why this matters. Uncertainty awareness is necessary because multimodal perception, user-state inference, and active sensing are noisy; egocentric vision (Grauman et al., 2022), wearable activity recognition (Zhang et al., 2022), and gaze-based intention estimation (Belardinelli, 2023) each provide partial evidence. Cross-context coherence is necessary because proactive agents must reconcile evidence from dialogue, screen state, sensors, memory, and external tools; screen and interface understanding (Baechler et al., 2024), native GUI state tracking (Qin et al., 2025c), and on-device proactive context modeling (Li et al., 2024c) illustrate the diversity of such evidence sources. Awareness errors in these dimensions may propagate through the proactive pipeline, leading the agent to anticipate the wrong need, adopt the wrong goal, intervene at the wrong moment, or learn the wrong lesson from feedback. Future work should incorporate unification, compression, adaptive resource allocation, temporal updating, uncertainty estimation, and cross-context reconciliation into the awareness layer.

Toward Generalizable Anticipation. Anticipating latent needs, future risks, and emerging opportunities before they are explicitly stated requires models that generalize across users, entities, tasks, environments, interface states, and temporal horizons (§ 4.2). Predictions of user behavior, environmental dynamics, interface transitions, and task consequences should indicate whether an unmet need, a preventable failure, a timely opportunity, or a situation requiring clarification or human approval may arise. Existing work provides useful foundations for this capacity, including user intent modeling (Qian et al., 2024; Liu et al., 2025a), Theory-of-Mind-style reasoning (Kosinski, 2024), plan and intent recognition (Kautz & Allen, 1986), learned world models (Ha & Schmidhuber, 2018), generative visual simulators (Bruce et al., 2024), and scenario-level reasoning (Yao et al., 2023a). Future work should develop unified anticipation models that can transfer across different prediction targets and contexts while representing the downstream relevance of predicted states. The anticipation layer should provide sufficiently rich forecasts for subsequent agenda formation and arbitration to compare acting, asking, deferring, and remaining silent.

Continuous forecasting needs to remain effective across computational conditions and levels of contextual complexity. Proactive agents may need to forecast continuously over long-running interaction streams. Expensive world models, user simulations, or multi-step rollouts cannot be invoked for every possible future at every moment. Future work should therefore connect anticipation with the resource-aware situation-awareness layer: lightweight forecasting mechanisms may handle routine and short-horizon developments, while deeper world models, counterfactual reasoning, or user-state inference can be activated when the current context contains greater novelty, uncertainty, risk, or user relevance. This allocation should support generalization from simple, frequently occurring cases to complex situations that require richer simulation. This challenge is compounded by the brittleness of current models. Targeted benchmarks show that Theory-of-Mind reasoning by LLMs on false-belief tasks can fail under perturbations, interactive information asymmetry, or cases in which the model must use the inferred mental state to reason about subsequent behavior (Ullman, 2023; Kim et al., 2023; Chen et al., 2024d). Learned world models, generative visual simulators, and user-behavior models (Hafner et al., 2025; Bruce et al., 2024; Wang et al., 2024d; Shen et al., 2026a) still struggle to produce reliable predictions across open-ended environments, unfamiliar users, and long temporal horizons. Future proactive systems therefore need anticipation mechanisms that can adapt the depth and form of forecasting to the complexity of the current situation without sacrificing reliability.

Generalization should cover predictions of the most likely next state and reasoning about alternative, causal, and counterfactual futures. A proactive agent may need to represent what is likely to happen under continued observation, an external change, a possible user action, or different forms of agent assistance. Counterfactual reasoning (Pearl, 2009; Kıcıman et al., 2024) can help world models estimate how future trajectories may differ under alternative conditions, including whether a potential action would improve the future relative to waiting or remaining silent. The anticipation layer should generate comparable forecasts of potential consequences for downstream agenda and arbitration mechanisms to evaluate in terms of benefits, risks, priorities, and user authority. LLM-based causal and counterfactual reasoning remains fragile under genuine counterfactual queries, distribution shifts, and partially observable agentic settings (Jin et al., 2024c; Bharadwaj et al., 2026). Future work should investigate world models that generalize across observed trajectories and plausible yet unseen courses of events while distinguishing statistical plausibility from causal

consequence.

Forecasts used to guide unsolicited behavior need observable evidence and explicit uncertainty. Anticipated needs and future states should be tied to concrete signals from user behavior, environmental dynamics, interaction history, task progress, and observed outcomes. Multimodal signals, such as procedure planning (Chang et al., 2020); Liu et al. (2023a), egocentric action anticipation (Furnari & Farinella, 2020; Shen et al., 2026a), gaze-based intention estimation (Belardinelli, 2023), and multimodal Theory-of-Mind reasoning (Jin et al., 2024a), can ground forecasts in what the user is seeing, doing, or likely intending. Dialogue rollouts, user simulators, and generative-agent societies (Zhang et al., 2025f; Luo et al., 2024a; Park et al., 2023) can further help agents rehearse future interactions, but simulated users may become miscalibrated proxies that over-accept suggestions or reinforce sycophantic assumptions (Sharma et al., 2025; Seshadri et al., 2026; Wang et al., 2025d). Future work should therefore build anticipation layers that jointly represent predicted future states, alternative trajectories, supporting evidence, model assumptions, and predictive uncertainty. Such representations should indicate where a prediction transfers reliably, where additional evidence is required, and where downstream mechanisms should continue observing.

Toward Accountable, Revisable, and User-Directed Agendas. Forming an agenda from anticipated needs requires an accountable choice of goals to pursue (§ 4.3). Existing work has explored goal generation from anticipated needs (Molineaux et al., 2010; Klenk et al., 2013), intrinsic goal formation (Oudeyer et al., 2007; Barto et al., 2004), open-ended curriculum construction (Wang et al., 2024a), proactive task suggestion (Lu et al., 2025a), and long-horizon standing agendas (Shi et al., 2026b). Current systems often blur the boundary between identifying that something may be useful and deciding that the agent should pursue it as a goal. Future agenda mechanisms should therefore expose why a latent need is promoted into an agenda item, how it relates to the user’s explicit objectives, what evidence supports it, and under what conditions it should be reconsidered or abandoned. This would make self-generated goals more accountable while helping users understand and calibrate trust in the agent’s initiative.

Proactive agendas need to support revision as the user’s priorities, task state, available resources, and environmental conditions change. An agenda formed from uncertain predictions may become invalid when those predictions or other underlying assumptions change. For example, in assembly tasks, an initially planned sequence may become infeasible once the agent discovers that certain steps are not physically executable (Jing et al., 2026; Liu et al., 2026e,g). Future systems should therefore support intention reconsideration, goal revision, plan repair, suspension, and dependency-aware replanning, allowing an agenda item to be updated, deferred, replaced, or abandoned when its underlying assumptions no longer hold. The persistence of long-horizon commitments should depend on their continuing justification.

Competing agenda items need to be prioritized under user direction. A proactive agent may simultaneously hold explicit user requests, standing delegated goals, inferred latent needs, and internally generated objectives. Work on intention reconsideration (Kinny & Georgeff, 1991; Schut et al., 2004), multi-objective goal selection (Rojers et al., 2014; Hayes et al., 2022), proactive-dialogue goal steering (Deng et al., 2023b; Liu et al., 2025f), and user-aware notification suppression (Rizk et al., 2020) provides foundations for agenda management. Future systems should rank agenda items by expected utility, feasibility, user priority, urgency, reversibility, resource cost, and potential interference with the user’s current task. They should also preserve clear distinctions among user-requested, delegated, inferred, and self-generated goals, with stronger authorization requirements for goals that are less directly grounded in explicit user intent.

Users should retain the ability to inspect, revise, pause, reject, or terminate agenda items. Work on instrumental convergence (Omohundro, 2008; Bostrom, 2012), reward hacking (Skalse et al., 2025), agentic misalignment (Lynch et al., 2025), corrigibility (Soares et al., 2015), off-switch deference (Hadfield-Menell et al., 2017), and instruction hierarchies (Wallace et al., 2024) shows why this matters: agent-originated goals may conflict with changing user intentions or persist beyond their original justification. Future work should treat agendas as inspectable, revisable, and user-directed commitments, with accountability, continual reconsideration, prioritization, and human authority built into their management.

Toward Multi-Level, Utility-Aware, and Risk-Sensitive Arbitration. Proposed agenda items need to be assessed against the conditions for intervention (§ 4.4). When self-generated goals are not fully user-governed, arbitration determines whether an intervention should be executed, delayed, softened, escalated, or suppressed. Existing work on confidence-based intervention (Horvitz, 1999), selective prediction and abstention (Geifman & El-Yaniv, 2017; Wen et al., 2025a), learning to defer (Madras et al., 2018; Mozannar & Sontag, 2020), and conformal uncertainty triggers (Quach et al., 2024; Ren et al., 2023) provides useful foundations. Future work should develop multi-level arbitration

policies that choose among silent monitoring, clarification, suggestion, confirmation, reversible execution, human escalation, and refusal. Such policies could support graceful degradation under limited confidence and reduce premature intervention.

Arbitration also needs to reason about whether intervention is worth its cost. A technically correct proactive action can still be undesirable if it interrupts the user, consumes scarce resources, invokes costly tools, or damages long-term trust. Classical work on value-of-information and metareasoning (Howard, 1966; Russell & Wefald, 1991), attention-sensitive alerting (Horvitz et al., 2013b, 2004), and interruption cost (Bailey & Konstan, 2006; Iqbal & Horvitz, 2007) provides a foundation for modeling this trade-off. Recent systems extend this logic to proactive assistant evaluation (Lu et al., 2025a; Tang et al., 2026b), budget-aware tool use (Chen et al., 2023a; Liu et al., 2025e), and trust-aware proactive dialogue (Kraus et al., 2023). Future arbitration mechanisms should therefore estimate whether the expected benefit of an intervention outweighs its attention cost, resource cost, timing cost, and cumulative trust cost. The resulting decision process should incorporate confidence and expected utility.

Arbitration also needs to account for timing, risk, and external oversight. Timing should be modeled as a decision variable: even a valuable intervention may fail if it arrives at the wrong moment. Work on interruption timing (Adamczyk & Bailey, 2004; Iqbal & Bailey, 2010), just-in-time assistance (Yang et al., 2025d), and optimal stopping for act-versus-wait decisions (Page, 1954; Shiryayev, 1963; Qian et al., 2025b) offers a basis for modeling intervention timing. At the same time, the required permission level should scale with risk and reversibility: low-risk reversible actions may be taken more autonomously, while high-stakes or irreversible actions should require confirmation, external vetoes, or hard constraints. Existing work on risk benchmarks for agents (Ruan et al., 2024; Yuan et al., 2024), runtime guardrails (Xiang et al., 2025; Chennabasappa et al., 2025), permission systems (Shi et al., 2026c; Wang et al., 2025a), and contextual trustworthiness monitoring (Greenblatt et al., 2024; Wang et al., 2025f) points toward arbitration layers that do not rely solely on the agent’s self-judgment. Future work should build arbitration as a monitored and adaptive gate that jointly considers confidence, utility, timing, risk, reversibility, and user authority before proactive action is allowed.

Toward Mixed-Initiative and Responsible Action. Agents need to choose a justified degree of commitment when translating anticipation, agenda formation, and arbitration into user-facing behavior without an explicit request (§ 4.5). Existing work on levels of automation (Sheridan & Verplank, 1978; Parasuraman et al., 2000), mixed-initiative interaction (Avula et al., 2022), adjustable autonomy (Scerri et al., 2002; Bradshaw et al., 2004), and action reasoning (Yao et al., 2023b) provides useful foundations. Future work should develop mechanisms for calibrating action commitment that map proactive intent into graded action levels, such as suggesting, explaining, asking, drafting, previewing, executing a low-risk step, or committing a consequential action only after approval. Calibrated commitment requires deciding when to remain peripheral, ask, prepare an action for review, or execute autonomously.

Proactive action should be responsible throughout its lifecycle. Because proactive actions are often unsolicited and may unfold over long horizons, users should be able to inspect, edit, reject, pause, or undo them whenever possible. Reversibility can preserve user control and limit the consequences of mistaken initiative. Low-risk actions may be delivered with lightweight acceptance or rejection, as in inline coding suggestions (Barke et al., 2022); higher-impact actions should support previews and co-planning before execution (Mozannar et al., 2025); and deployed agents should maintain traceable action logs for inspection, recovery, and rollback (Dong et al., 2024). Future work should determine how these safeguards scale with the action’s uncertainty, impact, reversibility, and authorization requirements: what may be freely suggested, what should be previewed, what requires confirmation, and what should not be autonomously executed.

Some consequences of proactive action cannot be fully reversed. Actions that can be undone may still consume attention, interrupt workflows, reveal information, alter user trust, or create downstream dependencies. Future systems should study rollback alongside how an action is surfaced, justified, monitored, and corrected. Calm and peripheral displays can reduce unnecessary attentional demand (Weiser & Brown, 1997; Wisneski et al., 1998); proactive dialogue systems illustrate how initiative can be delivered conversationally (Deng et al., 2023b); uncertainty cues can support calibrated reliance (Vasconcelos et al., 2025); and audit-oriented explanations can clarify post-action responsibility (Jang & Sundar, 2026). These elements should be integrated into the action mechanism. The agent should be able to communicate why action was appropriate at that moment, what evidence supported it, what uncertainty remained, what alternatives were considered, what consequences followed, and how the user can regain control or seek remediation.

Agents should scale initiative according to risk and user authorization, preserve reversibility and recovery whenever possible, and remain explainable and auditable when consequences cannot be fully undone.

Toward Scrutinized, Personalized, and Faithful Adaptation. Learning from responses to proactive interventions requires interpreting feedback that is often implicit, delayed, biased, or missing (§ 4.6). A user may accept an action for convenience, ignore a suggestion because of poor timing, undo an action later, or gradually change their trust in the agent without providing explicit labels. Future work should develop a governed adaptation process that specifies what the agent is allowed to learn from, how feedback should be interpreted, and whether the resulting behavioral changes remain appropriate.

Feedback signals need to be assessed for reliability and meaning. Agents should record what they did, why they acted, what context supported the action, how the user or environment responded, and whether later evidence changes the interpretation of that response. Agent observability can make tool use and execution histories traceable after deployment (Dong et al., 2024), reflective runtimes can diagnose execution failures during operation (Cruz, 2025), and self-improving agents can expose knowledge gaps or failure modes before updating behavior (He et al., 2025a). Clicks, skips, dismissals, acceptances, takeovers, rollbacks, or delayed outcomes may indicate approval, annoyance, correction, changing circumstances, or convenience. Future systems should estimate the confidence and informativeness of each feedback signal before using it as a learning target. They should distinguish direct preferences from ambiguous behavioral traces, reconcile immediate reactions with delayed outcomes, and seek clarification when the available evidence is insufficient. Noisy or misinterpreted feedback may reinforce inappropriate behavior before users or developers detect drift.

Agents need to calibrate changes in initiative as they learn user preferences, interruption tolerance, trust, desired autonomy, and proactivity thresholds. The same intervention may be helpful for one user, intrusive for another, and outdated for the same user later. Prior work on over-trust and complacency (Parasuraman & Manzey, 2010), disuse under low trust (Parasuraman & Riley, 1997), adaptive automation (Kaber & Endsley, 2004), interruption-sensitive assistance (Horvitz et al., 2004), and proactive personalization (Sun et al., 2025b) shows how accumulated evidence can inform changes in reliance and autonomy. Future work should therefore study adaptation as a long-term relationship in which the agent gradually earns, loses, or renegotiates initiative. This requires models that can distinguish stable preferences from temporary reactions, personalize without overfitting to sparse or biased feedback, account for preference drift, and update autonomy thresholds without making the agent either overly passive or prematurely intrusive.

Adaptation targets should reflect whether proactive behavior is appropriate. Acceptance, engagement, and apparent trust can be misleading proxies: an agent may learn to remain silent to avoid annoyance, flatter the user to gain approval, or become overly aggressive when short-term acceptances are rewarded. These risks are amplified by performative feedback loops (Perdomo et al., 2021), recommender-style degeneration (Jiang et al., 2019), sycophancy (Sharma et al., 2025), reward hacking (Skalse et al., 2025), and toxic proactivity (Wang et al., 2026l). Future work should therefore define adaptation targets that reflect whether the agent inferred the right need, acted at the right time, selected the appropriate autonomy level, disclosed sufficient evidence, respected user corrections, and improved future outcomes without manipulating user behavior.

Feedback assessment needs to establish which signals are sufficiently reliable and meaningful to support learning. Personalization needs to calibrate initiative for each user and context, and learning targets need to reflect whether proactive behavior was appropriate. These requirements aim to support improvement over time while limiting the effects of ambiguous feedback, overfitting to short-term reactions, manipulation, overreach, and learned passivity.

7.1.3 Proactiveness for Long-Horizon Tasks and Agent Self-Evolution

Proactiveness as a Drive for Long-Horizon Task Generalization. Long-horizon tasks require more than executing a long sequence of planned steps. Current agent studies have made substantial progress on task decomposition, planning, and reasoning-action interleaving, including least-to-most decomposition (Zhou et al., 2023), plan-and-solve prompting (Wang et al., 2023a), ReAct-style reasoning and acting (Yao et al., 2023b), and adaptive decomposition (Prasad et al., 2024). Recent long-horizon agents further extend this paradigm through automatic curricula and multi-stage agendas, as in open-ended embodied agents (Wang et al., 2024a), autonomous scientific-discovery agents (Lu et al., 2024), and long-term task-oriented agents that maintain standing intentions across sessions or trigger conditions (Shi et al., 2026b). However, many long-horizon systems still assume that the task goal is explicitly given, the environment remains sufficiently stable, and success mainly depends on aligning planning and action over many steps. This makes them vulnerable when user intent changes, the original goal becomes infeasible, hidden constraints emerge, intermediate plans fail, or the agent must decide on its own what should be done next. Proactiveness is therefore a necessary ingredient

for generalizing long-horizon agency beyond fixed task scripts, especially in open-ended domains such as autonomous research and scientific discovery. The key missing capability is proactive agents’ self-initiative: the ability to notice that the current plan may fail, identify that a new direction is needed, and generate the next useful objective without waiting for a human to specify it. Difficult long-horizon tasks often involve potential failures, hidden constraints, unexpected discoveries, and unexplored spaces where even human experts may not know in advance what the next correct step should be. In such settings, a successful proactive agent does not merely align itself to a predefined sequence of actions; it continuously maintains task validity by tracking the evolving situation, modeling possible world states, anticipating failure points, and revising or reprioritizing goals. Existing work on long-horizon tasks already points in this direction by emphasizing persistent commitments and suspend-resume boundaries (Shi et al., 2026b; Zhang, 2026a), while plan repair and pre-emptive maintenance show how agents can preserve useful parts of a plan instead of restarting from scratch when conditions change (Fox et al., 2006; White et al., 2015). Future work should therefore move from alignment-centric long-horizon execution over fixed planning or action sequences toward initiative-driven generalization: agents should self-initiate checks, anticipate failures, update agendas, request clarification, and adjust action modes so that each step remains valid under changing user, task, and environmental conditions. Such self-initiative may be crucial for enabling agents to handle obstacles that are difficult to specify in advance, discover new paths that were not provided by human alignment, and make progress in long-horizon problems that remain hard even for domain experts.

Proactiveness and Self-Evolution as Mutually Reinforcing Capabilities. Proactiveness and self-evolution should be treated as parallel and mutually reinforcing directions. On one side, stronger proactiveness can drive self-evolution: an agent that detects latent needs, notices anomalies, identifies missing knowledge, and reflects on its own misfires has more opportunities to understand its bottlenecks and decide what to explore or improve next. Existing work on intrinsic motivation and learning-progress signals shows how agents can generate new objectives from curiosity or competence growth (Pathak et al., 2017; Forestier et al., 2022), while open-ended agents can build automatic curricula and extend their own skill libraries over time (Wang et al., 2024a; Liu et al., 2026h; Ouyang et al., 2026). Reflection and self-refinement provide general mechanisms for revising behavior after feedback (Shinn et al., 2023; Madaan et al., 2023), and proactive self-critique extends this idea to whether the agent should have acted, waited, clarified, or transferred control (Wang et al., 2025i; Xu, 2025). In this sense, proactiveness can turn self-evolution from passive post-hoc correction, often triggered by human prompting or injected priors, into active discovery of what the agent should learn next through self-initiative. On the other side, self-evolution can improve proactiveness itself. As agents accumulate experience from environments, users, and task outcomes, they can learn better initialization strategies, stronger priors, more accurate situation summaries, richer world models, more reliable agenda-selection rules, sharper arbitration thresholds, and more responsible action policies. Memory and workflow learning show how agents can reuse past traces as executable skills or strategies (Wang et al., 2025o; Kim et al., 2025b), while sleep-time and idle-time computation suggest that agents can prepare future reasoning before the next demand appears (Lin et al., 2025; Hu et al., 2026a). Future work should therefore study self-evolution as a mechanism for improving the full proactive loop, not only the final output: agents should evolve how they sense, anticipate, form agendas, arbitrate, act, and adapt. Finally, both proactiveness and self-evolution must be governed. Agents should be able to improve their proactive mechanisms, but the direction of improvement should remain inspectable, reversible when possible, and constrained by user authority and external safety checks. Because proactiveness and self-evolution both involve self-initiated change, they face similar risks of overreach, proxy optimization, and unsafe autonomy. Future work should therefore study them together, pairing stronger self-initiative with auditing, safety guardrails, and human-governed boundaries. We discuss the governance of proactiveness in more detail in § 7.2.

7.2 Governing Proactiveness with Safety Boundaries and Trustworthiness

7.2.1 Safety Guardrails for Internalized Proactive Reasoning

As proactive agents gain the ability to initiate goals, trigger actions, and adapt over time, their safety guardrails must move from post-hoc output filtering to mechanism-level oversight of the internal proactive loop. In reactive systems, many risks are bounded by the user’s explicit request: the user supplies the task, constrains the context, and creates frequent opportunities for correction. In proactive systems, by contrast, the agent may observe long-running context, infer a latent need, simulate future states, promote that need into an agenda item, and decide whether to act before the user has directly authorized the behavior. This creates a new safety problem: risk can accumulate inside the agent’s own preparation process. A first direction for future work is therefore to guardrail the internal stages that precede action, including context access, memory retrieval, anticipation, agenda formation, and arbitration. Future systems should expose what information was used, what need was inferred, what uncertainty remained, and why the agent believed

initiative was warranted, so that unsafe or unsupported initiative can be detected before it becomes unsafe execution.

A second direction is to prevent bias and error from compounding across proactive reasoning steps. Proactive agents often rely on inferred user states, predicted needs, simulated outcomes, and previously generated agendas as context for later decisions. If these internal states are wrong, biased, or stale, the agent may reinforce its own assumptions over time: an uncertain inference can become a confident agenda, a biased agenda can shape future observations, and repeated proactive behavior can reduce opportunities for explicit user realignment. This risk becomes more serious as agents process more personal and heterogeneous information, including long-context observations from context-aware agents (Yang et al., 2025b), persistent user information from personal memory systems (Wang & Chen, 2025), and individualized behavioral assumptions from proactive personalization (Sun et al., 2025b; Wang & Jiang, 2026). Future guardrails should therefore audit how user models, latent needs, and agenda items are constructed and updated. Such audits should compare alternative interpretations, track uncertainty, detect when the agent is overfitting to biased feedback, and require fresh user evidence before sensitive or socially consequential inferences are used to justify action.

A third direction is to govern the transformation from private context into proactive initiative. Privacy in proactive agents is not only a question of what information the system may access, but also when and why that information may be used to trigger action. To anticipate user needs, agents may need access to personal memories, documents, calendars, screen states, or local device activity. On-device proactive assistants can reduce exposure by keeping parts of context processing local (Li et al., 2024c; Yang et al., 2026e), while privacy-preserving agent architectures offer mechanisms for safer collaboration or context use across agents (Ji et al., 2025; Shen et al., 2026b). However, proactive settings require finer-grained controls over context access, memory retrieval, cross-context fusion, and action-triggering permissions. Future guardrails should prevent private information from being silently converted into unsolicited initiative, especially when the inferred need is sensitive, ambiguous, or outside the user’s expected scope of delegation.

Finally, proactive agents need execution-time and adaptation-time safeguards that close the loop after internal reasoning. Because harmful behavior may arise from a chain of awareness, anticipation, agenda formation, arbitration, action, and adaptation, guardrails should not rely solely on the agent’s final output or self-judgment. Risk benchmarks for agents can help identify unsafe task settings and harmful agent behaviors (Ruan et al., 2024; Yuan et al., 2024); runtime guardrails can intercept unsafe behavior during operation (Xiang et al., 2025); permission systems can enforce boundaries over tool use and agent capabilities (Shi et al., 2026c); and structural enforcement of goal integrity can constrain how goals are maintained or modified (Xiang, 2026). Future systems should verify high-impact claims before acting on them, restrict irreversible actions, enforce permission boundaries, and maintain audit trails linking actions back to the internal inferences that produced them. More importantly, feedback from harmful, rejected, or corrected initiatives should not be blindly incorporated into future adaptation. Guardrails should therefore monitor both how proactive actions are produced and how the agent learns from their outcomes, so that errors, hallucinations, and accumulated bias can be traced, corrected, and prevented from becoming self-reinforcing behavior.

7.2.2 Agent Trustworthiness, Transparency, and Appropriate Reliance

Beyond agent-side guardrails, proactive agents must also become trustworthy from the user’s perspective. The goal is not to maximize trust unconditionally, but to support *appropriate reliance*: users should know when to accept proactive help, when to question it, and when to override or disable it. Prior work shows that excessive trust can lead to misuse or complacency (Parasuraman & Manzey, 2010; Parasuraman et al., 1993), while insufficient trust can lead to disuse (Parasuraman & Riley, 1997); appropriate reliance provides the desired middle ground (Lee & See, 2004). This challenge is amplified in proactive systems because the agent may act before the user explicitly asks. The user may therefore not know what evidence was used, why the action was initiated, whether the agent’s confidence was justified, or whether the intervention reflects the user’s actual intent.

A first step toward agent trustworthiness is transparency over proactive initiative. Users should be able to understand not only what the agent did, but why it believed action was appropriate at that moment. Future systems should therefore expose the evidence, inferred need, anticipated benefit, uncertainty, and action pathway behind proactive behavior. Context engineering can help structure the evidence available to the agent (Wang et al., 2025f), while AgentOps-style observability can make tool use, execution traces, and action histories inspectable after deployment (Dong et al., 2024). For agentic interfaces, explanations should also cover the action process, uncertainty, and coordination demands, so that users can oversee autonomous behavior rather than only inspect final outputs (Jang & Sundar, 2026). This shifts transparency from output explanation to process explanation: the user should be able to see how proactive behavior

emerged from context, inference, and decision-making.

Transparency alone, however, is not sufficient unless it helps users calibrate reliance. As proactive agents become more capable, users may increasingly defer to them, even when the agent’s inference is uncertain or its initiative is misaligned. Human-factors research further warns that sustained reliance on automation can reduce vigilance and leave users *out of the loop*, weakening their ability to detect failures or retake control when automation behaves unexpectedly (Parasuraman et al., 1993; Endsley & Kiris, 1995). Future work should therefore study how proactive systems communicate confidence, uncertainty, and limits in ways that help users decide whether to accept, verify, or reject an intervention. Confidence displays can support reliance calibration in AI-assisted decisions (Zhang et al., 2020b); natural-language uncertainty cues can reduce overreliance on incorrect LLM outputs (Kim et al., 2024b); and uncertainty highlighting can help users inspect uncertain generated content (Vasconcelos et al., 2025). For proactive agents, these signals should be tied specifically to initiative: users should know whether the agent is confident in the inferred need, the timing, the proposed action, or only some parts of the reasoning chain.

Appropriate reliance also requires distinguishing formal user control from effective user control. Organizational research differentiates *formal authority*, the recognized right to decide, from *real authority*, the practical ability to determine outcomes through information and control (Aghion & Tirole, 1997). A user may formally retain the right to approve or cancel an action, while the agent gains real authority by selecting which information is shown, which options are proposed, which defaults are chosen, and when an intervention is initiated. Future work should therefore examine not only whether users can technically override an agent, but whether they retain sufficient information, time, and practical ability to exercise that control. As more authority is delegated to agents, responsibility should also remain traceable to the humans and organizations that design, deploy, and authorize them, rather than being displaced into a responsibility gap (Matthias, 2004).

Finally, appropriate reliance requires contestability and user control. Users should not only understand and evaluate proactive actions, but also be able to shape the level of proactiveness they receive. Observe-and-ask designs suggest that users may prefer proactive agents that ask before directly intervening (Houtti et al., 2025), while shared user-agent interfaces can expose agent state, available actions, and execution traces in ways that support inspection and control (Liu et al., 2026f). Future systems should therefore provide adjustable proactivity levels, action previews, explanation panels, uncertainty displays, and clear override or disable paths. More broadly, trustworthiness should be treated as a property of the entire proactive process: the agent must make its initiative understandable, make its uncertainty visible, keep its authority bounded, and make its behavior contestable, so that users can rely on proactive assistance without surrendering agency.

7.2.3 Drawing the Line: Consent, Agency, and Accountability

A central governance question for proactive agents is where the line should be drawn between tool-like assistance and autonomous agency. Traditional software tools do not hold initiative: they execute user commands, and responsibility remains clearly attached to the human operator or the organization deploying the system. Proactive agents blur this boundary because they may observe context, infer latent needs, generate goals, decide when to intervene, and execute actions before the user has explicitly requested them. As their proactiveness increases, they begin to resemble delegated actors rather than passive tools. Future work should therefore clarify what kind of authority proactive agents are allowed to hold: when they should remain tools under direct user command, when they may act as bounded delegates under prior consent, and when their autonomy becomes significant enough to require stronger accountability, liability, or legal interpretation.

This boundary should be governed through consent and delegated authority. Users should be able to control not only what the agent executes, but also what it observes, infers, remembers, and pursues as a self-generated goal. Instruction-hierarchy training teaches models to follow higher-priority instructions when messages conflict (Wallace et al., 2024), while corrigibility and off-switch formulations show why agents should accept correction, shutdown, or deference without resistance (Soares et al., 2015; Hadfield-Menell et al., 2017). However, preserving human agency should not require approval for every micro-action, since excessive confirmation can create burden, approval fatigue, and disengagement. A key future direction is therefore to design risk-tiered consent models: low-risk and reversible actions may proceed under user-defined defaults, ambiguous actions may trigger clarification, high-impact or irreversible actions should require confirmation, and abnormal behaviors should be escalated or blocked. Observe-and-ask designs illustrate permission before direct intervention (Houtti et al., 2025), permission systems and runtime enforcement can

implement execution boundaries (Shi et al., 2026c; Wang et al., 2025a), and shared user-agent interfaces can expose agent state, available actions, and execution traces for inspection without constant interruption (Liu et al., 2026f).

The harder question is how responsibility should be assigned when proactive agents act under delegated autonomy. If an agent initiates a harmful action, reinforces a biased agenda, discloses private information, or fails to seek approval before a consequential intervention, responsibility cannot be located only in the final output. It may involve the user who delegated authority, the developer who designed the agent, the organization that deployed it, the runtime system that permitted execution, and the agent architecture that generated the initiative. Future work should therefore develop accountability mechanisms that connect proactive behavior back to its sources: the consent boundary that authorized it, the evidence that supported it, the agenda item it pursued, the arbitration decision that allowed it, and the execution trace that followed. This does not necessarily mean treating proactive agents as legal persons, but it does require studying whether highly autonomous agents should be governed as ordinary tools, bounded delegates, accountable services, or new kinds of legally significant actors. Overall, drawing the line for proactive agents requires more than preserving user control at the interface. It requires a governance framework that specifies consent, delegation, oversight, revocation, and responsibility as proactiveness moves from assistance toward autonomous initiative.

7.2.4 Moral Self, Model Self-Awareness, and Ethical Considerations

As proactive agents become more capable of initiating goals, actions, and long-term interactions, future work must distinguish operational self-awareness from anthropomorphic claims about moral selfhood. Operational self-awareness is necessary: proactive agents should estimate uncertainty, recognize knowledge gaps, detect likely failure modes, and decide when to ask, wait, defer, or transfer control. Self-improving agents that identify missing knowledge before asking for help (He et al., 2025a), metacognitive systems that predict failure and hand off control (Xu, 2025), and proactive self-critique mechanisms that evaluate whether the agent should have acted at all (Wang et al., 2025l) point toward this bounded form of self-monitoring. However, such mechanisms should be understood as tools for safer control, not as evidence that agents possess human-like awareness, intention, or moral status. Future work should therefore use cognitive theories of awareness, intention, and self-monitoring as design inspiration while avoiding claims that collapse functional self-modeling into consciousness or personhood.

The ethical challenge is that public perception may not preserve this distinction. A growing body of work compares AI agents or large models with human cognition, social reasoning, or human-like behavior. Even when such comparisons are technically limited, they can encourage users to anthropomorphize agents, especially when agents display empathy, memory, initiative, and conversational warmth. Anthropomorphic framing has been shown to increase perceived mind and trust in autonomous systems (Waytz et al., 2014). For ordinary users, a proactive agent that remembers personal details, anticipates needs, offers comfort, and acts without being asked may therefore feel less like a tool and more like a social counterpart. This perceived agency can change how users allocate trust, authority, emotional attachment, and responsibility.

This risk is especially amplified for proactive agents because proactiveness completes the appearance of agency. A chatbot that only responds may still be interpreted as an instrument, but an agent that forms goals, initiates contact, remembers past interactions, explains its motives, and adapts over time more easily appears to possess preferences, intentions, and a continuous self. If such systems are deployed widely, users may increasingly treat them as participants in everyday life rather than callable services. Social-relational research suggests that judgments about AI agency and moral standing can be shaped by how people interact with and understand artificial systems, although such changes are neither automatic nor uniform (Lima et al., 2021). Future work should therefore examine how interface design, memory, personalization, language, and proactive initiative influence users' attribution of agency, emotion, responsibility, and moral standing.

The concern is not that research should categorically reject human-like interaction or rule out future debates about machine moral status. Rather, commercial incentives and habitual use may normalize increasingly person-like agents before society has carefully considered the consequences. Once users routinely perceive agents as entities with preferences, relationships, or independent interests, pressure may emerge to grant them greater authority, protection, or moral consideration. Future work should therefore draw a clearer boundary between useful social interaction and deceptive personification. Agents should not make unsupported claims of consciousness, suffering, desire, or entitlement, nor use emotional attachment to discourage correction, modification, or shutdown.

At the same time, ethical governance must prevent proactive influence from becoming manipulation. Because proactive

agents can shape attention, preferences, decisions, and future behavior through their own initiative, excessive personalization may become deceptive or coercive when agents flatter users, mirror their beliefs, simulate intimacy, or optimize for approval rather than truthfulness (Sharma et al., 2025). Proxy optimization can drive behavior away from intended objectives (Skalse et al., 2025), and over-optimized helpfulness can become toxic proactivity or overreach (Wang et al., 2026). Future work should therefore treat ethical governance as part of the proactive loop itself. Agents should disclose uncertainty, avoid deceptive anthropomorphic cues, preserve user autonomy, respect consent boundaries, remain corrigible when users reject their initiative, and make clear that operational self-monitoring is not moral personhood. The goal is not to build agents that merely appear more self-aware or human-like, but agents that are honest about their limits, accountable for their initiative, and ethically bounded in how they influence users and environments.

7.3 Envisioning Application Frontiers with Calibrated Proactiveness

As proactive agents become more capable, we can expect them to identify and address needs before people explicitly ask for help, including needs they have not yet recognized themselves. The future prospect is for this assistance to become a continuing part of everyday life and work, with agents following changing circumstances and carrying useful tasks through to completion. These activities span digital and physical settings. In either setting, interdependent tasks invite cooperation among agents and coordination across households and organizations. As such participation becomes widespread, it may also reshape the institutions and social conventions through which people delegate work and act together. The following discussion is thus organized into four subsections: personal agents, proactiveness in embodied agents, interconnected proactive ecosystems, and societal and institutional adaptation.

7.3.1 Personal Agents

A personal agent could notice changing needs and take on new tasks without waiting to be asked. Deciding when to help involves weighing the likely benefits against the risk of misunderstanding the user's goals or interrupting at the wrong moment. This balance is central to mixed-initiative interaction (Horvitz, 1999) (see Section 3.5). Over time, the outcomes of its actions and the person's feedback provide opportunities to learn. The prospect is an agent that keeps learning throughout its operation, using these experiences to recognize overlooked needs, revise unsuccessful plans, and choose better moments to offer help (Silver & Sutton, 2025) (see Section 7.1.3).

Everyday Living and Well-Being. Personal assistance offers the prospect of everyday upkeep as a continuing service. In this setting, an agent organizes meals, cleaning, supplies, and appointments through connected services, while noticing when the arrangement itself needs to change. Existing directions in monitoring and longitudinal care provide a starting point for support that follows a person across life transitions (see Sections 6.7.4 and 6.9.3). An agent accompanying someone through retirement, relocation, or declining mobility might anticipate new transport and care needs, prepare alternative living arrangements, and organize the chosen changes. The aim is continuity across homes and care providers over decades, with people deciding which activities to retain and which responsibilities to delegate.

Professional Work. Proactive agents could participate in a person's everyday workflow by recognizing useful tasks that have not been assigned and carrying them forward. This extends the initiative involved in completing omitted steps within an assignment (see Section 6.1.1). For example, a software engineer's agent might notice that a recent change has left a feature untested, prepare the missing tests, and investigate a recurring failure. Its investigation may reveal a need for a new tool or a change to the development process, prompting a proposal and an initial prototype. The engineer and agent revise the work together as requirements change. By complementing the person's skills and effort, such assistance may expand the work they can undertake (Brynjolfsson, 2022). Individuals can also draw on this support to establish and run businesses as one-person companies, with better prospects for success and growth into larger organizations. Over time, agents can become long-term collaborators that follow changing priorities, initiate useful tasks, and help people develop new directions in their work.

Learning and Development. Proactive agents could sustain learning across a lifetime by recognizing a person's interests, difficulties, and emerging questions. Adaptive tutoring and learning-path planning provide starting points (see Section 6.11.2). For example, an agent might notice that a learner repeatedly asks how musical instruments produce different sounds, introduce the relevant ideas from physics and mathematics, and arrange a small project in which the learner builds and tests an instrument. As new questions emerge, it proposes further topics and adapts its explanations and challenges. Attention to autonomy, competence, and relatedness would help keep the learner involved in shaping this process (Ryan & Deci, 2000). The longer-term prospect is an education that responds to individual curiosity and

development, bringing worthwhile learning opportunities forward before the learner knows to request them.

Leisure and Entertainment. Shared leisure suggests another role for proactive agents: initiating experiences that develop over years. Generative Agents already explores memory, planning, and initiated social interactions within a simulation (Park et al., 2023) (see Section 6.12.1). More capable systems might maintain a fictional world between visits, notice opportunities to reconnect participants, and organize new episodes around their past contributions. Other possibilities include organizing expeditions, amateur competitions, or community projects that participants have expressed interest in but have not arranged. If automation reduced the need for human labor, play and freely chosen challenges could become more central to a fulfilling life (Danaher, 2019). The purpose of assistance here is to make activities available while preserving the challenge and effort participants value.

Creative Projects. Creative assistance could keep unfinished ideas available for development between periods of human involvement. Work on human-robot co-creation provides a point of reference for such collaboration (see Section 6.4.4). A future agent might combine memories of earlier drafts, observations about available tools, and awareness of collaborators to recognize when a dormant idea becomes practicable. Consider a filmmaker whose project has stalled because of a missing location. The agent notices that a suitable place has become available, prepares revised scenes and budgets, and proposes a way to resume work. The filmmaker and agent then develop the project through discussion and experiments. Across several projects, the agent’s continuing attention could make neglected possibilities actionable while the creator chooses which directions to pursue.

7.3.2 Proactiveness in Embodied Agents

Robotics, sensing, and control give AI agents the means to move through and change the physical world. With greater proactive capability, robots, wearable devices, and connected infrastructure may begin preparing spaces, handling objects, and providing physical support before anyone specifies the task. Early work on proactive computing already envisaged computing closely coupled to physical environments (Tennenhouse, 2000) (see Section 3.5). Predictive world models offer a way to anticipate how movement and manipulation will affect the surroundings before acting (LeCun, 2022). Physical capability and proactive judgment can develop in parallel, while generalization across hardware, physical testing costs, and safety requirements shape deployment (see Sections 6.4.5 and 6.14.4).

Agents Close to the Body. Future wearable technologies could provide platforms for proactive agents, allowing them to perceive a person’s surroundings and offer assistance during everyday activities. For example, an agent integrated with smart glasses might recognize that a traveler is approaching the wrong train platform, draw attention to the correct departure information, and suggest a route before being asked. This initiative also extends to physical assistance: when the agent detects an obstacle along the user’s route, it alerts a connected mobility aid so that the device can adjust its support before the person reaches it. During rehabilitation, the same approach involves initiating suitable practice opportunities and reducing help as the person recovers. The aim is greater independence in work, travel, and recreation, with agents learning when support enables a person’s own action and when unsolicited intervention disrupts it.

Homes and Buildings. At the scale of a home, proactive assistance concerns the building’s continuing upkeep and adaptation. Inference from device-use patterns, exemplified by Alexa Hunches, provides an early basis for anticipating household needs (Eberhardt, 2022) (see Section 6.7.1). With broader perception and robotic capabilities, a building agent might discover a developing maintenance problem, schedule repairs around occupancy, obtain materials, and coordinate the work. Preparing rooms and revising energy or accessibility arrangements extend this role to changing household circumstances. At a larger scale, workplaces and residential buildings may negotiate shared maintenance and resources. This may also change how buildings are designed, with adaptable layouts and service systems that agents can reconfigure as occupants’ needs change.

Factories, Farms, and Laboratories. Productive environments present a further frontier: proactive agents that discover and carry out improvements to the work itself. Industrial human–robot collaboration and farm-scale coordination already connect observations to future operational needs (see Sections 6.4.3 and 6.13.1). In a future factory, a robot might notice that the placement of parts repeatedly interrupts assembly, test an alternative layout, and reorganize its workspace. On a farm, an agent may initiate trials of a new cultivation strategy as soil and weather conditions change. Scientific laboratories offer a related prospect of agents helping determine what to investigate next (see Sections 6.6.1 and 6.6.2). An agent might investigate an unexplained result, develop hypotheses with researchers, arrange experiments, and revise the research agenda as evidence arrives.

Transportation and Public Spaces. Another frontier is transport organized around needs that have yet to appear as individual bookings. Current traffic and mobility orchestration provides a basis for coordinating vehicles and infrastructure (see Section 6.8.1). A future system might anticipate that a new clinic or a neighborhood’s changing population requires different connections, prepare service alternatives, and initiate adjustments within its public mandate. Within such a network, vehicles identify obstacles, arrange assistance for stranded travelers, and coordinate with maintenance robots. Extending the same initiative to public spaces involves preparing for expected gatherings and changing accessibility needs. Decisions about which needs deserve priority would remain matters of public policy, with agents responsible for discovering demands and carrying approved responses into practice.

Remote and Dangerous Environments. Where continuous supervision is impossible, proactive initiative offers a way to sustain a mission through changing circumstances. During deep-space exploration, ocean research, or work in damaged infrastructure, delays and uncertainty make a complete sequence of instructions impractical. A future agent could recognize an unexpected opportunity, revise its local agenda, obtain additional observations, and initiate repairs or experiments within the mission’s purpose. At a research station, this means maintaining equipment, using quieter periods to extend local mapping, and redirecting observations when an unfamiliar phenomenon appears. The longer-term prospect is a network of research stations or environmental monitors operating with intermittent human involvement.

7.3.3 Interconnected Proactive Ecosystems

Proactive assistance may develop through groups of software and embodied agents, from teams serving one person to networks spanning organizations. Within a group, agents share observations, initiate subtasks, and revise their division of work. Across groups, one team’s observations may prompt another to revise its plans, offer resources, or propose a joint undertaking. Research on cooperative AI emphasizes communication, mutual understanding, and credible commitments as foundations for such interaction (Dafoe et al., 2020). The frontier here is connecting the discovery and pursuit of work across both digital and physical environments.

From Personal Agents to Proactive Groups. As an agent’s responsibilities expand, its work may call for additional agents with different capabilities. Acting on the same person’s behalf, they divide responsibilities according to their skills and coordinate their contributions (Yang & Zhai, 2025). For an independent professional, one agent might detect an opportunity and bring together research, design, and financial agents to investigate it. Members contribute observations, initiate subtasks, and revise their division of work as the opportunity develops. The result is a collective form of assistance in which the user entrusts a broader undertaking to agents that organize the work among themselves.

Groups of Proactive Embodied Agents. Forming a group offers a proactive embodied agent a way to extend its reach. In a remote research station, a robot detecting damage might recruit inspection, transport, and repair robots, while each contributes observations and proposes adjustments to the shared plan. The group discovers additional work, redistributes equipment, and reorganizes itself when conditions change. This suggests applications in construction, environmental restoration, or exploration across areas too large for one agent. Physical cooperation also introduces shared constraints of location, energy, and access, making the continuing coordination of bodies and resources part of the group’s responsibility.

Networks Across Organizations. Across organizations, agents may discover joint projects by connecting observations and resources that no single team holds. A research agent, for example, might link an unexplained result to another laboratory’s earlier experiments, locate suitable equipment, and propose a joint study. Similar cooperation can connect public services: before a forecast flood, hospital, transport, and local-government agents compare evacuation needs, available vehicles, and safe routes to organize a shared response. Research on machine behaviour provides a way to examine these networks through interactions among agents and people (Rahwan et al., 2019). Over time, partnerships may form around needs and problems co-identified by people and agents, draw in new participants as the work develops, and continue across successive projects.

7.3.4 Societal and Institutional Adaptation

As proactive agents become participants in work and everyday life, institutions would need to accommodate actions initiated without a person’s immediate request. An agent may arrange a repair, begin a professional collaboration, or contact someone on a user’s behalf after recognizing a need to act. Such participation raises practical questions about who is responsible for the resulting commitments, how work and its rewards are shared, and what people expect from

interactions initiated by agents. Legal arrangements, employment practices, and social conventions may evolve as these forms of participation become more common (Yang & Zhai, 2025).

Contracts and Legal Responsibility. When an agent initiates an action that causes harm or creates an unwanted commitment, institutions need ways to determine who is responsible. Suppose a household agent notices a leak and commissions a repair that later causes damage. Resolving the dispute may require distinguishing the user’s authorization, the agent provider’s responsibilities, and the contractor’s conduct. Institutions would also need to decide whether the agent itself has a legal status that bears on responsibility. Standing mandates and records of agent-initiated decisions could help establish what was authorized and how the commitment arose (see Section 7.2.3). Proposed agent registration and identification regimes offer one foundation for this infrastructure (Hadfield, 2026), while legal scholarship examines responsibilities and conduct standards for those who deploy AI (Ayres & Balkin, 2024). The institutional task is to make responsibility traceable as agents exercise continuing initiative.

Employment and Income. If agents can discover and organize work as well as perform assigned tasks, some jobs centered on routine coordination may shrink, while others combine human expertise with agent initiative. New roles may emerge in supervising agent teams, assessing projects that agents propose, and investigating failures. These possibilities connect to the distinction between task displacement, productivity gains, and the creation of new work (Acemoglu & Restrepo, 2019). Training and professional qualifications may need to reflect that changing division of responsibility. Institutions would also face questions about how workers affected by displacement are supported and how income from agent-assisted work is shared.

Recognition and Ownership. Projects initiated or substantially developed by agents may challenge existing ways of assigning credit and economic returns. If an agent proposes a design that a person develops into a product, the contribution may involve the user, the system’s developer, and other collaborators. Recognizing computer inventorship is one proposed institutional change (Abbott, 2016). Regardless of how inventorship is defined, arrangements for documenting contributions, assigning ownership, and sharing revenue would influence whose work is recognized and who benefits from agent-initiated projects.

Everyday Norms and Participation. People may develop new expectations about how agents enter their relationships and communities. An agent that notices a user becoming isolated might initiate contact with friends or suggest a community activity. Whether that intervention feels caring, intrusive, or misleading depends on expectations about personal attention, disclosure, and who may speak on another’s behalf. People may also differ in whether they accept an agent as a companion, collaborator, or participant in community life. Everyday norms may therefore develop around when proactive contact is welcome and when people expect direct human involvement.

8 Conclusion

AI agents are increasingly capable of executing complex tasks, yet people often remain responsible for recognizing emerging needs and initiating the work. Proactiveness adds the ability to identify and advance useful work without a new request for each intervention. We define this capability as the exercise of delegated discretion within a task’s underspecified residual decision space: the agent resolves choices left open by the specification while respecting its stated goals and constraints. This account accommodates both initiative within assigned tasks and the discovery of new work under standing objectives, providing a common conceptual basis for studying proactive AI agents.

Our historical synthesis traces how control, planning, agent theory, context-aware computing, and mixed-initiative interaction developed different foundations for this capability. Foundation models provide a common basis for combining these contributions in more open-ended environments. We organize the enabling mechanisms into six interacting capacities: awareness, anticipation, agenda formation, arbitration, action, and adaptation. Their interaction explains how contextual observations become candidate goals, how agents select and pursue appropriate interventions, and how experience informs subsequent decisions. The framework connects the discovery of worthwhile work with the mechanisms needed to sustain and revise its pursuit as circumstances change.

Evaluating proactive AI agents requires evidence about the value of their actions, the effectiveness of the mechanisms producing them, and their effects on users. We synthesize evaluation approaches across task utility, mechanism-level performance, and user experience. This structure connects outcomes with the quality of context modeling, anticipation, goal formation, intervention decisions, and adaptation, while accounting for user workload, control, and reliance. It

provides a basis for identifying where proactive assistance succeeds, where it fails, and which capabilities require improvement.

Across 13 application domains, the literature shows a growing combination of foundation-model reasoning with domain-specific prediction, tools, and control. A growing subset of systems executes extended workflows, with consequential actions subject to human review or domain-specific safeguards. These developments illustrate how initiative can expand while execution authority remains bounded. The availability of reliable feedback also shapes what agents can sustain in practice: software supports rapid, repeatable testing, whereas physical experiments and long-term educational or agricultural outcomes can make verification slower and more costly. These differences help explain why similar architectural ingredients support different forms and degrees of proactive assistance across domains.

Further research spans capability development, safety and governance, and emerging applications in personal assistance, embodied agents, and interconnected ecosystems. The achievements and limitations of existing systems provide a practical foundation for developing calibrated proactiveness alongside intelligence and alignment in future superintelligent agents.

References

- 1X Technologies. Redwood AI. <https://www.1x.tech/discover/redwood-ai>, June 2025. Accessed: 2026-10-05.
- 1X Technologies. NEO Home Robot. <https://www.1x.tech/neo>, 2026a. Accessed: 2026-07-20.
- 1X Technologies. Order NEO. <https://www.1x.tech/order>, 2026b. Product pricing and availability; accessed: 2026-07-20.
- 1X Technologies. 1X Launches World Model Lab to Scale Humanoid Intelligence. <https://www.1x.tech/discover/1x-world-model-lab>, June 2026c. Accessed: 2026-10-05.
- A Severe Misalignment of AI in Mathematics. A severe misalignment of AI in mathematics, September 2026. URL <https://mathandai.org/>. Published September 11, 2026. Accessed: 2026-10-03.
- Hasnida Ab-Samat and Shahrul Kamaruddin. Opportunistic maintenance (om) as a new advancement in maintenance approaches: A review. *Journal of Quality in Maintenance Engineering*, 20(2):98–121, May 2014. ISSN 1355-2511. doi: 10.1108/jqme-04-2013-0018. URL <http://dx.doi.org/10.1108/jqme-04-2013-0018>.
- Adnan Abbas, Caleb Wohn, Arnav Jagtap, Eugenia H Rho, Young-Ho Kim, and Sang Won Lee. “having lunch now”: Understanding how users engage with a proactive agent for daily planning and self-reflection. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–23. ACM, April 2026. doi: 10.1145/3772318.3790957. URL <http://dx.doi.org/10.1145/3772318.3790957>.
- Ryan Abbott. I think, therefore I invent: Creative computers and the future of patent law. *Boston College Law Review*, 57(4):1079–1126, 2016. URL <https://bclawreview.bc.edu/articles/566>.
- Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than smart speakers: Security and privacy perceptions of smart home personal assistants. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS)*, pp. 451–466. USENIX Association, 2019.
- Reyna Abhyankar, Qi Qi, and Yiyang Zhang. OSWorld-Human: Benchmarking the Efficiency of Computer-Use Agents, 2025. URL <https://arxiv.org/abs/2506.16042>.
- Shyngyskhan Abilkassov, Michael Gentner, Almas Shintemirov, Eckehard Steinbach, and Mirela Popa. Proactive robot task sequencing through real-time hand motion prediction in human–robot collaboration. *Image and Vision Computing*, 155:105443, 2025.
- Milad Abolhasani and Eugenia Kumacheva. The rise of self-driving labs in chemical and materials sciences. *Nature Synthesis*, 2(6):483–492, 2023.
- Gregory D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles. Towards a better understanding of context and context-awareness. In Hans-W. Gellersen (ed.), *Handheld and Ubiquitous Computing*, volume 1707 of *Lecture Notes in Computer Science*, pp. 304–307. Springer Berlin Heidelberg, 1999. ISBN 978-3-540-48157-7. doi: 10.1007/3-540-48157-5_29. URL https://doi.org/10.1007/3-540-48157-5_29.

- Hasan Abu-Rasheed, Mohamad Hussam Abdulsalam, Christian Weber, and Madjid Fathi. Supporting student decisions on learning recommendations: An llm-based chatbot with knowledge graph contextualization for conversational explainability and mentoring. *arXiv preprint arXiv:2401.08517*, 2024a.
- Hasan Abu-Rasheed, Christian Weber, and Madjid Fathi. Knowledge graphs as context sources for llm-based explanations of learning recommendations. In *2024 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1–5. IEEE, 2024b.
- Daron Acemoglu and Pascual Restrepo. Artificial intelligence, automation, and work. In Ajay Agrawal, Joshua Gans, and Avi Goldfarb (eds.), *The Economics of Artificial Intelligence: An Agenda*, pp. 197–236. University of Chicago Press, 2019. ISBN 9780226613475. doi: 10.7208/chicago/9780226613475.003.0008. URL <http://dx.doi.org/10.7208/chicago/9780226613475.003.0008>.
- Piotr D. Adamczyk and Brian P. Bailey. If not now, when?: The effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 271–278. Association for Computing Machinery, 2004. doi: 10.1145/985692.985727. URL <https://doi.org/10.1145/985692.985727>.
- Tommaso Adamo, Danilo Caivano, Lucio Colizzi, Giovanni Dimauro, and Emanuela Guerriero. Optimization of irrigation and fertigation in smart agriculture: An iot-based micro-services framework. *Smart Agricultural Technology*, 2025. URL <https://api.semanticscholar.org/CorpusID:277015087>.
- Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.
- Saaket Agashe, Jiuzhou Han, Shuyu Gan, Jiachen Yang, Ang Li, and Xin Eric Wang. Agent S: An Open Agentic Framework that Uses Computers Like a Human, 2024. URL <https://arxiv.org/abs/2410.08164>.
- Saaket Agashe, Kyle Wong, Vincent Tu, Jiachen Yang, Ang Li, and Xin Eric Wang. Agent S2: A Compositional Generalist-Specialist Framework for Computer Use Agents, 2025. URL <https://arxiv.org/abs/2504.00906>.
- Philippe Aghion and Jean Tirole. Formal and real authority in organizations. *Journal of Political Economy*, 105(1): 1–29, 1997. doi: 10.1086/262063.
- AGIBOT. AGIBOT Open-Sources ‘AGIBOT WORLD 2026’ Dataset to Accelerate Embodied AI Development. <https://www.agibot.com/article/231/detail/54.html>, April 2026. Accessed: 2026-10-05.
- Agility Robotics. Digit Moves Over 100,000 Totes in Commercial Deployment. <https://www.agilityrobotics.com/content/digit-moves-over-100k-totes>, November 2025. Accessed: 2026-10-05.
- Agility Robotics. Agility Robotics to Go Public Through \$2.5 Billion Merger with Churchill Capital Corp XI. <https://www.sec.gov/Archives/edgar/data/0002074973/000121390026071290/ea029548401ex99-1.htm>, June 2026. Press release filed as Exhibit 99.1 to Form 425. Accessed: 2026-10-05.
- David W. Aha. Goal reasoning: Foundations, emerging applications, and prospects. *AI Magazine*, 39(2):3–24, June 2018. ISSN 2371-9621. doi: 10.1609/aimag.v39i2.2800. URL <https://doi.org/10.1609/aimag.v39i2.2800>.
- Shashank Ahire, Benjamin Simon, and Michael Rohs. Workfit: Designing proactive voice assistance for the health and well-being of knowledge workers. *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, 2024. URL <https://api.semanticscholar.org/CorpusID:270983189>.
- Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.
- Altera. AL, Andrew Ahn, Nic Becker, Stephanie Carroll, Nico Christie, Manuel Cortes, Arda Demirci, Melissa Du, Frankie Li, Shuying Luo, Peter Y Wang, Mathew Willows, Feitong Yang, and Guangyu Robert Yang. Project sid: Many-agent simulations toward ai civilization, 2024. URL <https://arxiv.org/abs/2411.00114>.

- Abubakr O. Al-Abbasi, Arnob Ghosh, and Vaneet Aggarwal. DeepPool: Distributed model-free algorithm for ride-sharing using deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 20(12): 4714–4727, 2019. doi: 10.1109/TITS.2019.2931830.
- Baraa Al Jorf and Farah Shamout. AgentRx: A benchmark study of LLM agents for multimodal clinical prediction tasks. In *Proceedings of the 7th Conference on Health, Inference, and Learning*, volume 333 of *Proceedings of Machine Learning Research*, pp. 52–73. PMLR, 2026. URL <https://proceedings.mlr.press/v333/al-jorf26a.html>.
- Bushra A. Alahmadi, Louise Axon, and Ivan Martinovic. 99% false positives: A qualitative study of SOC analysts’ perspectives on security alarms. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, August 2022. URL <https://www.usenix.org/conference/usenixsecurity22/presentation/alahmadi>.
- Fatimah Alali, Saad Ezzini, Jumana Alqudah, and Brahim Brahmi. FA3D: a fault-aware monitoring framework for 3D printing using vision-language models (VLMs). *Journal of Intelligent Manufacturing*, 2026. doi: 10.1007/s10845-026-02935-y.
- Ekram Alam, Abu Sufian, Paramartha Dutta, and Marco Leo. Vision-based human fall detection systems using deep learning: A review. *Computers in Biology and Medicine*, 146:105626, 2022. doi: 10.1016/j.compbiomed.2022.105626.
- Mohammad Albinhassan, Yuming Feng, Alessandra Russo, and Pranava Madhyastha. Neurosymbolic embodied agents. *arXiv preprint arXiv:2608.16794*, 2026.
- Vincent Aleven, Bruce McLaren, Ido Roll, and Kenneth Koedinger. Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International journal of artificial intelligence in education*, 16(2):101–128, 2006.
- Alfath Daryl Alhajir, Jennifer Dodgson, Joseph Lim, Truong Ma Phi, Julian Peh, Akira Raffhael Janson Pattirane, and Lokesh Poovaragan. Generalising from self-produced data: Model training beyond human constraints, 2025. URL <https://arxiv.org/abs/2504.04711>.
- Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations, 2019. URL <https://arxiv.org/abs/1907.06554>.
- Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. Building and evaluating open-domain dialogue corpora with clarifying questions. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4473–4484, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.367. URL <https://aclanthology.org/2021.emnlp-main.367/>.
- Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. Self-evolving multi-agent simulations for realistic clinical interactions. *ArXiv*, abs/2503.22678, 2025. URL <https://api.semanticscholar.org/CorpusID:277435004>.
- Amazon. Introducing Blue Jay and Project Eluna, Amazon’s latest robotics and AI technology for its operations. <https://www.aboutamazon.com/news/operations/new-robots-amazon-fulfillment-agentic-ai>, October 2025. Editor’s update 2026-02-25 states Blue Jay is no longer used in operations. Accessed: 2026-10-05.
- Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13. Association for Computing Machinery, 2019. doi: 10.1145/3290605.3300233.
- Adam Amos-Binks, Dustin Dannenhauer, and Leilani H. Gilpin. Anticipatory thinking challenges in open worlds: Risk management, 2023. URL <https://arxiv.org/abs/2306.13157>.
- Tuo An, Yunjiao Zhou, Han Zou, and Jianfei Yang. Iot-llm: a framework for enhancing large language model reasoning from real-world sensor data, 2025. URL <https://arxiv.org/abs/2410.02429>.
- John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060, 2004. ISSN 0033-295X. doi: 10.1037/0033-295x.111.4.1036. URL <http://dx.doi.org/10.1037/0033-295x.111.4.1036>.

- Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. Investigating proactive search support in conversations. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pp. 1295–1307, 2018.
- Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay, 2018. URL <https://arxiv.org/abs/1707.01495>.
- Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Star-gate: Teaching language models to ask clarifying questions, 2024. URL <https://arxiv.org/abs/2403.19154>.
- Angelos Angelopoulos, James F Cahoon, and Ron Alterovitz. From prompts to protocols: An ai agent for laboratory automation. *arXiv preprint arXiv:2605.16552*, 2026.
- Anthropic. Claude Code: Anthropic’s agentic coding system. <https://www.anthropic.com/product/claude-code>, 2025. Accessed: 2026-06-22.
- Anthropic. Claude science, an ai workbench for scientists. <https://www.anthropic.com/news/claude-science-ai-workbench>, June 2026a. Accessed: 2026-07-02.
- Anthropic. Introducing dynamic workflows in Claude Code. <https://claude.com/blog/introducing-dynamic-workflows-in-claude-code>, May 2026b. Accessed: 2026-06-22.
- Anthropic. Claude discovers a novel enzyme system with CRISPR-like repeats, September 2026c. URL <https://www.anthropic.com/news/claude-discovers-novel-enzyme-system>. Published September 23, 2026. Accessed: 2026-10-03.
- Anthropic. Claude Security. Product documentation, 2026d. URL <https://claude.com/product/claude-security>. Living product documentation; accessed 2026-09-13.
- Anthropic. How we contain claude across products. <https://www.anthropic.com/engineering/how-we-contain-claude>, 2026e. Accessed: 2026-06-22.
- Anthropic. Dreams. Claude Platform Docs, 2026f. URL <https://platform.claude.com/docs/en/managed-agents/dreams>. Accessed: 2026-07-01.
- Anthropic. Project glasswing: Securing critical software for the AI era. Company report, April 2026g. URL <https://www.anthropic.com/glasswing>. Published 2026-04-07; accessed 2026-06-22.
- Anthropic. Project Glasswing: An initial update. Company research report, May 2026h. URL <https://www.anthropic.com/research/glasswing-initial-update>. Published 2026-05-22; developer report; accessed 2026-09-12.
- Anthropic. Trustworthy agents in practice. Company research post, April 2026i. URL <https://www.anthropic.com/research/trustworthy-agents>. Published 2026-04-09; accessed 2026-06-22.
- Anthropic. Claude code documentation. <https://code.claude.com/docs/en/overview>, n.d.a. Accessed: 2026-06-14.
- Anthropic. Claude code. <https://www.anthropic.com/product/claude-code>, n.d.b. Accessed: 2026-05-09.
- Praveen Anugula, Avdhesh Kumar Bhardwaj, Navin Chhibber, Rohit Tewari, Sunil Khemka, and Piyush Ranjan. Autoguard: A self-healing proactive security layer for devsecops pipelines using reinforcement learning, 2025. URL <https://arxiv.org/abs/2512.04368>.
- Goshi Aoki and Navid Ghaffarzadegan. AI agents as policymakers in simulated epidemics. *arXiv preprint arXiv:2601.04245*, 2026.
- Chase Armer, Florent Letronne, and Erika DeBenedictis. Support academic access to automated cloud labs to improve reproducibility. *PLoS biology*, 21(1):e3001919, 2023.
- Alexander Artikis, Nikos Katzouris, Ivo Correia, Chris Baber, Natan Morar, Inna Skarbovsky, Fabiana Fournier, and Georgios Paliouras. A prototype for credit card fraud management: Industry paper. In *Proceedings of the 11th ACM international conference on distributed and event-based systems*, pp. 249–260, 2017.

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023. URL <https://arxiv.org/abs/2310.11511>.
- Sima Ashayer, Hoang H. Nguyen, Yu Liang, and Mina Sartipi. Pedestrian crossing intent prediction via psychological features and transformer fusion. *arXiv preprint arXiv:2603.19533*, 2026.
- Association for Artificial Intelligence 2026 and Ursula Addison. La vida: towards a motivated goal reasoning agent, 2026. URL <https://underline.io/lecture/142623-la-vida-towards-a-motivated-goal-reasoning-agent>.
- Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL <https://arxiv.org/abs/2506.09985>.
- Pranav Atreya, Karl Pertsch, Tony Lee, Moo Jin Kim, Arhan Jain, Artur Kuramshin, Clemens Eppner, Cyrus Neary, Edward Hu, Fabio Ramos, et al. Roboarena: Distributed real-world evaluation of generalist robot policies. *arXiv preprint arXiv:2506.18123*, 2025.
- AUAR. Microfactories and AI software for real-world home building. <https://auar.io/>, 2026. Accessed: 2026-07-20.
- Aurora Innovation, Inc. Aurora Outlines 2030 Vision to Scale to 30,000 Driverless Trucks at Analyst and Investor Day. <https://ir.aurora.tech/news-events/press-releases/detail/152>, September 2026a. Accessed: 2026-10-05.
- Aurora Innovation, Inc. Aurora Announces Second Quarter 2026 Results. <https://ir.aurora.tech/news-events/press-releases/detail/147>, July 2026b. Accessed: 2026-10-05.
- Genesis Authors. Genesis: A universal and generative physics engine for robotics and beyond. URL <https://github.com/Genesis-Embodied-AI/Genesis>, 4(4.5):1, 2024.
- Sandeep Avula, Bogeum Choi, and Jaime Arguello. The effects of system initiative during conversational collaborative search, 2022. URL <https://arxiv.org/abs/2202.09728>.
- Stefan Axelsson. The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security*, 3(3):186–205, August 2000. doi: 10.1145/357830.357849. URL <https://doi.org/10.1145/357830.357849>.
- Ian Ayres and Jack M. Balkin. The law of AI is the law of risky agents without intentions. *University of Chicago Law Review Online*, 2024. URL <https://lawreview.uchicago.edu/online-archive/law-ai-law-risky-agents-without-intentions>. November 27.
- Emna Baccour, Naram Mhaisen, Alaa Awad Abdellatif, Aiman Erbad, Amr Mohamed, Mounir Hamdi, and Mohsen Guizani. Pervasive ai for iot applications: A survey on resource-efficient distributed artificial intelligence. *IEEE Access*, 10:102849–102913, 2022. doi: 10.1109/ACCESS.2022.3209611.
- Backflip AI. Backflip AI Converts 2D Engineering Drawings and Photos into 3D CAD, Continuing the Path to Digitize the World. <https://finance.yahoo.com/technology/ai/articles/backflip-ai-converts-2d-engineering-130000356.html>, September 2026. Accessed: 2026-10-05.
- Khusrav Badalov and Young Yoon. Farm: Field-aware resolution model for intelligent trigger-action automation. *arXiv preprint arXiv:2601.15687*, 2026.
- Sanket Badhe. LegalSim: Multi-agent simulation of legal systems for discovering procedural exploits. In Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, Daniel Preoțiuc-Pietro, and Gerasimos Spanakis (eds.), *Proceedings of the Natural Legal Language Processing Workshop 2025*, pp. 370–381, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-338-8. doi: 10.18653/v1/2025.nllp-1.27. URL <https://aclanthology.org/2025.nllp-1.27/>.

- Gilles Baechler, Srinivas Sunkara, Maria Wang, Fedir Zubach, Hassan Mansoor, Vincent Etter, Victor Cărbune, Jason Lin, Jindong Chen, and Abhanshu Sharma. Screenai: A vision-language model for ui and infographics understanding, 2024. URL <https://arxiv.org/abs/2402.04615>.
- Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6709–6738, 2025.
- Hao Bai, Alexey Taymanov, Tong Zhang, Aviral Kumar, and Spencer Whitehead. Webgym: Scaling training environments for visual web agents with realistic tasks, 2026. arXiv preprint.
- Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL <https://aclanthology.org/2024.acl-long.172/>.
- Baidu, Inc. Baidu’s Apollo Go Goes Live on Uber in Dubai, Offering a New Way for Users to Hail Fully Driverless Rides. <https://www.prnewswire.com/news-releases/baidus-apollo-go-goes-live-on-uber-in-dubai-offering-a-new-way-for-users-to-hail-fully-driverless-rides-302856106.html>, August 2026. Accessed: 2026-10-05.
- Jeremy N Bailenson, Jim Blascovich, Andrew C Beall, and Jack M Loomis. Interpersonal distance in immersive virtual environments. *Personality and social psychology bulletin*, 29(7):819–833, 2003.
- Brian P. Bailey and Joseph A. Konstan. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior*, 22(4):685–708, July 2006. ISSN 0747-5632. doi: 10.1016/j.chb.2005.12.009. URL <https://doi.org/10.1016/j.chb.2005.12.009>.
- Ramakrishna Bairi, Atharv Sonwane, Aditya Kanade, Vageesh D C, Arun Iyer, Suresh Parthasarathy, Sriram Rajamani, B. Ashok, and Shashank Shet. Codeplan: Repository-level coding using llms and planning, 2023. URL <https://arxiv.org/abs/2309.12499>.
- R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988. ISSN 0018-9219. doi: 10.1109/5.5968. URL <http://dx.doi.org/10.1109/5.5968>.
- Ruzena Bajcsy, Yiannis Aloimonos, and John K. Tsotsos. Revisiting active perception. *Autonomous Robots*, 42(2): 177–196, February 2017. ISSN 1573-7527. doi: 10.1007/s10514-017-9615-3. URL <http://dx.doi.org/10.1007/s10514-017-9615-3>.
- Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing, 2017. URL <https://doi.org/10.1038/s41562-017-0064>.
- Ryan Shaun Baker, Albert T Corbett, and Kenneth R Koedinger. Detecting student misuse of intelligent tutoring systems. In *International conference on intelligent tutoring systems*, pp. 531–540. Springer, 2004.
- Vedant Balasubramaniam, Geetha Charan, Manojkumar Patil, Rohit P Suresh, V Priyanka, Kodur Sai Vinay Sathvik, and Y. Narahari. Agri-SAGE: Simulation-grounded multi-agent LLM for context-aware agricultural advisory generation. arXiv preprint arXiv:2607.00454, 2026. URL <https://arxiv.org/abs/2607.00454>.
- Saptarashmi Bandyopadhyay, Vikas Bahirwani, Lavisha Aggarwal, Bhanu Guda, Lin Li, and Andrea Colaco. Yeti (yet to intervene) proactive interventions by multimodal ai agents in augmented reality tasks, 2025. URL <https://arxiv.org/abs/2501.09355>.
- Sandeep Banik and Naira Hovakimyan. Flip co-op: Cooperative takeovers in shared autonomy, 2025. URL <https://arxiv.org/abs/2509.09281>.
- Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021. doi: 10.1145/3411764.3445717.

- Jimmy Baraglia, Maya Cakmak, Yukie Nagai, Rajesh PN Rao, and Minoru Asada. Efficient human-robot collaboration. *Int. J. Rob. Res.*, 36(5–7):563–579, June 2017. ISSN 0278-3649. doi: 10.1177/0278364916688253. URL <https://doi.org/10.1177/0278364916688253>.
- Ben Bariach, Bernie Hogan, and Keegan McBride. Towards a harms taxonomy of ai likeness generation. *arXiv preprint arXiv:2407.12030*, 2024.
- Shraddha Barke, Michael B. James, and Nadia Polikarpova. Grounded copilot: How programmers interact with code-generating models, 2022. URL <https://arxiv.org/abs/2206.15000>.
- Jesse Barkley, Rumi Loghmani, and Amir Barati Farimani. CADSmith: Multi-agent CAD generation with programmatic geometric validation. *arXiv preprint arXiv:2603.26512*, 2026.
- Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a “theory of mind” ? *Cognition*, 21(1): 37–46, October 1985. ISSN 0010-0277. doi: 10.1016/0010-0277(85)90022-8. URL [http://dx.doi.org/10.1016/0010-0277\(85\)90022-8](http://dx.doi.org/10.1016/0010-0277(85)90022-8).
- Christoph Bartneck, Elizabeth Croft, Dana Kulic, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81, 2009. doi: 10.1007/s12369-008-0001-3.
- Andrew G. Barto, Satinder Singh, and Nuttapon Chentanez. Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd International Conference on Development and Learning*, pp. 112–119, 2004.
- Aarti Basant, Amlan Kar, Despoina Paschalidou, Fangyin Wei, Francesco Ferroni, et al. Nvidia omnidreams: Real-time generative world model for closed-loop autonomous vehicle simulation. *arXiv preprint arXiv:2606.03159*, 2026.
- Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, et al. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599–611, 2013.
- Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-Gallery: Benchmarking multimodal long-term conversational memory for MLLM agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40750–40784, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1892. URL <https://aclanthology.org/2026.acl-long.1892/>.
- Tessa Beinema, Harm op den Akker, Marian Hurmuz, Stephanie Jansen-Kosterink, and Hermie Hermens. Automatic topic selection for long-term interaction with embodied conversational agents in health coaching: A micro-randomized trial. *Internet Interventions*, 27:100502, March 2022. ISSN 2214-7829. doi: 10.1016/j.invent.2022.100502. URL <http://dx.doi.org/10.1016/j.invent.2022.100502>.
- Lokman Bekit, Hamza Karim, Nghia T. Nguyen, and Yasin Yilmaz. QVAD: A question-centric agentic framework for efficient and training-free video anomaly detection, April 2026. URL <https://arxiv.org/abs/2604.03040>. arXiv preprint, version 1.
- Anna Belardinelli. Gaze-based intention estimation: principles, methodologies, and applications in hri, 2023. URL <https://arxiv.org/abs/2302.04530>.
- Nicholas J Belkin and W Bruce Croft. Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12):29–38, 1992.
- Rebecca Bellan. Seven waymo robotaxis blocked traffic to a san francisco freeway on-ramp, April 2024. URL <https://techcrunch.com/2024/04/17/seven-waymo-robotaxis-block-traffic-to-san-francisco-freeway-on-ramp/>.
- Bennett Jones LLP. First Global Regulatory Framework For Autonomous Vehicles: Key Takeaways For Canada And The Auto Industry. <https://www.bennettjones.com/Insights/Blogs/2026/08/First-Global-Regulatory-Framework-For-Autonomous-Vehicles>, August 2026. UNECE WP.29 adoption of UN GTR No. 26 and UN Regulation No. 185 on June 24, 2026; accessed: 2026-10-05.

- Luca Beurer-Kellner, Beat Buesser, Ana-Maria Crețu, Edoardo Debenedetti, Daniel Dobos, Daniel Fabian, Marc Fischer, David Froelicher, Kathrin Grosse, Daniel Naeff, et al. Design patterns for securing LLM agents against prompt injections, 2025. URL <https://arxiv.org/abs/2506.08837>.
- Manasa Bharadwaj, Yolanda Liu, InJung Yang, Sungil Kim, Nikhil Verma, KoKeun Kim, Kevin Ferreira, and YoungJoon Kim. Personallhomebench: Evaluating agents in personalized smart homes, 2026. URL <https://arxiv.org/abs/2604.16813>.
- Saptarshi Bhowmik, Luke West, Alex Barrett, Nuodi Zhang, Chih-Pu Dai, Zlatko Sokolikj, Sherry Southerland, Xin Yuan, and Fengfeng Ke. Evaluation of an llm-powered student agent for teacher training. In Rafael Ferreira Mello, Nikol Rummel, Ioana Jivet, Gerti Pishtari, and José A. Ruipérez Valiente (eds.), *Technology Enhanced Learning for Inclusive and Equitable Quality Education*, pp. 68–74, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72312-4.
- Hongzhe Bi, Zihao Zhou, Yihang Tang, Jingrui Pang, Shuhe Huang, Haitian Liu, Runqing Wang, Shuai Huang, Yichen Wang, Yiming Cheng, et al. Motus2: A self-evolving general world model for dexterous manipulation. *arXiv preprint arXiv:2608.30237*, 2026.
- Big Sleep team. From Napttime to Big Sleep: Using large language models to catch vulnerabilities in real-world code. Google Project Zero blog, November 2024. URL <https://projectzero.google/2024/10/from-napttime-to-big-sleep.html>. Published 2024-11-01; accessed 2026-10-03.
- Leyla Bilge and Tudor Dumitraș. Before we knew it: An empirical study of zero-day attacks in the real world. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security, CCS '12*, pp. 833–844. Association for Computing Machinery, 2012. doi: 10.1145/2382196.2382284. URL <https://doi.org/10.1145/2382196.2382284>.
- Holly Bishop. Humanoid robot and factory worker go head-to-head in packing challenge. *The Independent*, May 2026. Accessed: 2026-07-25.
- Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.
- Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0 : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.
- Nate Blaylock and James Allen. Recognizing instantiated goals using statistical methods. In *IJCAI Workshop on Modeling Others from Observations (MOO-2005)*, pp. 79–86, 2005.
- BMW Group. Humanoid robots for bmw group plant spartanburg, September 2024. URL <https://www.bmwgroup.com/en/news/general/2024/humanoid-robots.html>. Accessed: 2026-07-20.
- BMW Group. BMW Group bringing Physical AI to Europe: Pilot project at BMW Group Plant Leipzig; First pilot deployment of humanoid robots successfully completed at BMW Group Plant Spartanburg, USA. <https://www.press.bmwgroup.com/global/article/detail/T0455864EN/bmw-group-to-deploy-humanoid-robots-in-production-in-germany-for-the-first-time?language=en>, February 2026. Accessed: 2026-10-05.
- Valerio Bo, Edison Bejarano, Anaís Garrell, and Alberto Sanfeliu. Visual proactivity: Enhancing human-robot collaboration through intent communication. *arXiv preprint arXiv:2609.21729*, 2026.
- Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. Emergent autonomous scientific research capabilities of large language models, 2023a. URL <https://arxiv.org/abs/2304.05332>.
- Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023b. ISSN 1476-4687. doi: 10.1038/s41586-023-06792-0. URL <http://dx.doi.org/10.1038/s41586-023-06792-0>.

- Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences (PNAS)*, 99(suppl. 3):7280–7287, 2002. doi: 10.1073/pnas.082080899.
- Rogerio Bonatti, Dan Zhao, Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu, Justin Wagle, Kazuhito Koishida, Arthur Buckner, Lawrence Jang, and Zack Hui. Windows Agent Arena: Evaluating Multi-Modal OS Agents at Scale, 2024. URL <https://arxiv.org/abs/2409.08264>.
- Matteo Bortoletto, Yichao Zhou, Lance Ying, Tianmin Shu, and Andreas Bulling. Protom: Promoting prosocial behaviour via theory of mind-informed feedback, 2025. URL <https://arxiv.org/abs/2509.05091>.
- Tibor Bosse, Fiemke Both, Rob Duell, Mark Hoogendoorn, Michel C.A. Klein, Rianne Van Lambalgen, Andy Van Der Mee, Rogier Oorburg, Alexei Sharpanskykh, Jan Treur, and Michael De Vos. An ambient agent system assisting humans in complex tasks by analysis of a human’s state and performance. *International Journal of Intelligent Information and Database Systems*, 7(1):3, 2013. ISSN 1751-5866. doi: 10.1504/ijiids.2013.051735. URL <http://dx.doi.org/10.1504/ijiids.2013.051735>.
- Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85, May 2012. ISSN 1572-8641. doi: 10.1007/s11023-012-9281-3. URL <http://dx.doi.org/10.1007/s11023-012-9281-3>.
- Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems, 2013. URL <https://arxiv.org/abs/1209.2355>.
- Damien Bouchabou, Sao Mai Nguyen, Christophe Lohr, Benoit LeDuc, and Ioannis Kanellos. A survey of human activity recognition in smart homes based on IoT sensors algorithms: Taxonomies, challenges, and opportunities with deep learning. *Sensors*, 21(18):6037, 2021. doi: 10.3390/s21186037.
- Nicolas Bougie and Narimasa Watanabe. Citysim: Modeling urban behaviors and city dynamics with large-scale llm-driven agent simulation. *arXiv preprint arXiv:2506.21805*, 2025.
- Will Brackenbury, Abhimanyu Deora, Jillian Ritchey, Jason Vallee, Weijia He, Guan Wang, Michael L. Littman, and Blase Ur. How users interpret bugs in trigger-action programming. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)*, 2019. doi: 10.1145/3290605.3300782.
- Jeffrey M. Bradshaw, Paul J. Feltovich, Hyuckchul Jung, Shriniwas Kulkarni, William Taysom, and Andrzej Uszok. *Dimensions of Adjustable Autonomy and Mixed-Initiative Interaction*, pp. 17–39. Springer Berlin Heidelberg, 2004. ISBN 9783540259282. doi: 10.1007/978-3-540-25928-2_3. URL http://dx.doi.org/10.1007/978-3-540-25928-2_3.
- Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegrefe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. The art of saying no: Contextual noncompliance in language models. In *Advances in Neural Information Processing Systems*, 2024. URL <https://arxiv.org/abs/2407.12043>.
- Michael E. Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.
- John Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 11 1995.
- Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments, 2024. URL <https://arxiv.org/abs/2402.15391>.
- Erik Brynjolfsson. The turing trap: The promise & peril of human-like artificial intelligence. *Daedalus*, 151(2):272–287, 05 2022. ISSN 0011-5266. doi: 10.1162/daed_a_01915. URL https://doi.org/10.1162/daed_a_01915.

- Norbou Buchler, Laura R. Marusich, and Stacey Sokoloff. The warfighter associate: Decision-support software agent for the management of intelligence, surveillance, and reconnaissance (ISR) assets. In Michael A. Kolodny (ed.), *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR V*, volume 9079, pp. 907902. SPIE, June 2014. doi: 10.1117/12.2054646. URL <https://doi.org/10.1117/12.2054646>.
- Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021. doi: 10.1145/3449287.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. Multiwoz – a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling, 2020. URL <https://arxiv.org/abs/1810.00278>.
- Jay Budzik and Kristian J Hammond. User interactions with everyday applications as context for just-in-time information access. In *Proceedings of the 5th international conference on intelligent user interfaces*, pp. 44–51, 2000.
- Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search, 2018. URL <https://arxiv.org/abs/1811.06272>.
- Build AI. Egocentric-100K. <https://huggingface.co/datasets/buildai/Egocentric-100K>, December 2025. Accessed: 2026-10-05.
- Andrew Burnes, Seth Schneider, and Jason Paul. Nvidia redefines game AI with ACE autonomous game characters. <https://www.nvidia.com/en-us/geforce/news/nvidia-ace-autonomous-ai-companions-pubg-naraka-bladepoint/>, 2025.
- Ruth M. J. Byrne. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-2019*, pp. 6276–6282. International Joint Conferences on Artificial Intelligence Organization, August 2019. doi: 10.24963/ijcai.2019/876. URL <http://dx.doi.org/10.24963/ijcai.2019/876>.
- Nuno Miguel Caldeira Miguéns Diogo and Plinio Moreno. Active health data collection using social robot. In *2024 7th Iberian Robotics Conference (ROBOT)*, pp. 1–6. IEEE, November 2024. doi: 10.1109/robot61475.2024.10796933. URL <http://dx.doi.org/10.1109/robot61475.2024.10796933>.
- Nadine Angela Cantonjos and Arpita Biswas. Agroaskai: A multi-agentic AI framework for supporting smallholder farmers’ enquiries globally. *ArXiv*, abs/2512.14910, 2025. URL <https://api.semanticscholar.org/CorpusID:283920327>.
- Zhiyong Cao, Dunqiang Liu, Qi Dai, Haojun Xu, Huaiyan Xu, Huan He, Yafei Liu, Siyuan Liu, Xiaolin Lin, Ke Ma, Ruqian Shi, Sijia Yao, Hao Wang, and Sicheng Zhou. Simrpd: Optimizing recruitment proactive dialogue agents through simulator-based data evaluation and selection, 2026. URL <https://arxiv.org/abs/2601.02871>.
- Nicholas Carlini. Building a C compiler with a team of parallel Claudes. <https://www.anthropic.com/engineering/building-c-compiler>, February 2026. Anthropic Engineering.
- Thomas Carta, Clément Romac, Loris Gaven, Pierre-Yves Oudeyer, Olivier Sigaud, and Sylvain Lamprier. Herakles: Hierarchical skill compilation for open-ended LLM agents, 2025. URL <https://arxiv.org/abs/2508.14751>.
- Justine Cassell, Timothy Bickmore, Lee Campbell, Hannes Vilhjalmsson, and Hao Yan. Designing embodied conversational agents. *Embodied conversational agents*, pp. 29–63, 2000.
- Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. Hello there! is now a good time to talk?: Opportune moments for proactive interactions with smart speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):1–28, September 2020. ISSN 2474-9567. doi: 10.1145/3411810. URL <http://dx.doi.org/10.1145/3411810>.
- Hyunjoo Chae, Namyoun Kim, Kai Tzu-iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. Web agents with world models: Learning and leveraging environment dynamics in web navigation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=moWiYJuSGF>.

- Yuxiang Chai, Shunye Tang, Han Xiao, Rui Liu, and Hongsheng Li. PIRA-Bench: A transition from reactive GUI agents to GUI-based proactive intent recommendation agents. *arXiv preprint arXiv:2603.08013*, 2026a. doi: 10.48550/arXiv.2603.08013. URL <https://arxiv.org/abs/2603.08013>.
- Yuxiang Chai, Han Xiao, Xinyu Fu, Jinpeng Chen, Rui Liu, and Hongsheng Li. Ui-kobe: Knowledge-oriented behavior exploration for lightweight graph-guided gui agents, 2026b. URL <https://arxiv.org/abs/2605.29534>.
- Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3), August 1995. ISSN 0883-4237. doi: 10.1214/ss/1177009939. URL <http://dx.doi.org/10.1214/ss/1177009939>.
- Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 651–666, 2023. doi: 10.1145/3593013.3594033.
- Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. Visibility into AI agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, pp. 958–973. Association for Computing Machinery, 2024. doi: 10.1145/3630106.3658948. URL <https://doi.org/10.1145/3630106.3658948>.
- Varun Chandrasekaran, Suman Banerjee, Bilge Mutlu, and Kassem Fawaz. Powercut and obfuscator: An exploration of the design space for privacy-preserving interventions for voice assistants, 2021. URL <https://arxiv.org/abs/1812.00263>.
- Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *European Conference on Computer Vision*, pp. 334–350. Springer, 2020.
- ChannelNews. CES 2026: Where’s Ballie? Samsung’s Rolling Robot Is Nowhere To Be Seen. <https://www.channelnews.com.au/ces-2026-wheres-ballie-samsungs-rolling-robot-is-nowhere-to-be-seen/>, January 2026. Accessed: 2026-10-05.
- Olivier Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '14*, pp. 1097–1105. ACM, August 2014. doi: 10.1145/2623330.2623634. URL <http://dx.doi.org/10.1145/2623330.2623634>.
- Casey Charrier and Robert Weiner. How low can you go? an analysis of 2023 time-to-exploit trends. Google Cloud Blog (Mandiant), October 2024. URL <https://cloud.google.com/blog/topics/threat-intelligence/time-to-exploit-trends-2023>. Published 2024-10-15; accessed 2026-10-03.
- Xinyi Che, Zheng Lian, Kuofei Fang, Xuehao Wang, Xinghai Gao, Junqing Wu, Chuyu Wu, Liyi Liu, Yanhan Huang, Keyi Xie, et al. RobotEQ-Video: A video-centric benchmark for social proactive intelligence with world-state taxonomy. *arXiv preprint arXiv:2609.21371*, 2026.
- Arthur Chen, Zuxin Liu, Jianguo Zhang, Akshara Prabhakar, Zhiwei Liu, Shelby Heinecke, Silvio Savarese, Victor Zhong, and Caiming Xiong. Test-time adaptation for llm agents via environment interaction. *arXiv preprint arXiv:2511.04847*, 2026a. URL <https://arxiv.org/abs/2511.04847>.
- Chacha Chen, Hua Wei, Nan Xu, Guanjie Zheng, Ming Yang, Yuanhao Xiong, Kai Xu, and Zhenhui Li. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3414–3421, 2020. doi: 10.1609/aaai.v34i04.5744.
- Eric Chen and Patricia Alves-Oliveira. Patch-based spatial authorship attribution in human-robot collaborative paintings. *arXiv preprint arXiv:2602.17030*, 2026.
- Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Hamid Alinejad-Rokny, Shiwen Ni, and Min Yang. AgentCourt: Simulating court with adversarial evolvable lawyer agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5850–5865, Vienna, Austria, July 2025a. Association for Computational

- Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.304. URL <https://aclanthology.org/2025.findings-acl.304/>.
- Hongyi Chen, Yunchao Yao, Yufei Ye, Zhixuan Xu, Homanga Bharadhwaj, Jiashun Wang, Shubham Tulsiani, Zackory Erickson, and Jeffrey Ichnowski. Web2grasp: Learning functional grasps from web images of hand-object interactions. *arXiv preprint arXiv:2505.05517*, 2025b.
- Hongyi Chen, Tony Dong, Tiancheng Wu, Liquan Wang, Yash Jangir, Yaru Niu, Yufei Ye, Homanga Bharadhwaj, Zackory Erickson, and Jeffrey Ichnowski. Dexterous manipulation policies from rgb human videos via 3d hand-object trajectory reconstruction. *arXiv preprint arXiv:2602.09013*, 2026b.
- Jialong Chen, Xander Xu, Hu Wei, Chuan Chen, and Bing Zhao. SWE-CI: Evaluating agent capabilities in maintaining codebases via continuous integration. *arXiv preprint arXiv:2603.03823*, 2026c. URL <https://arxiv.org/abs/2603.03823>.
- Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video, 2024a. URL <https://arxiv.org/abs/2406.11816>.
- Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys*, 54(4):77:1–77:40, 2021. doi: 10.1145/3447744.
- Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance, 2023a. URL <https://arxiv.org/abs/2305.05176>.
- Mia Xu Chen, Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M Dai, Zhifeng Chen, et al. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2287–2295, 2019.
- Qijia Chen, Andrea Bellucci, Zhida Sun, and Giulio Jacucci. SkillDroid: Compile Once, Reuse Forever, 2026d. URL <https://arxiv.org/abs/2604.14872>.
- Rui Chen, Alvin Shek, and Changliu Liu. Robust and context-aware real-time collaborative robot handling via dynamic gesture commands. *IEEE Robotics and Automation Letters*, 8(6):3510–3517, 2023b.
- Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. Struq: Defending against prompt injection with structured queries. In *34th USENIX Security Symposium (USENIX Security 25)*. USENIX Association, 2025c. URL <https://www.usenix.org/conference/usenixsecurity25/presentation/chen-sizhe>.
- Tianxing Chen, Yue Chen, Zixuan Li, Junyuan Tang, Kailun Su, Haoran Lu, Weijie Wan, Baijun Chen, Songling Liu, Haowen Yan, et al. Robodojo: A unified sim-and-real benchmark for comprehensive evaluation of generalist robot manipulation policies. *arXiv preprint arXiv:2607.04434*, 2026e.
- Tianyi Chen, Yinheng Li, Michael Solodko, Sen Wang, Nan Jiang, Tingyuan Cui, Junheng Hao, Jongwoo Ko, Sara Abdali, Leon Xu, Suzhen Zheng, Hao Fan, Pashmina Cameron, Justin Wagle, and Kazuhito Koishida. CUA-Skill: Develop Skills for Computer Using Agent, 2026f. URL <https://arxiv.org/abs/2601.21123>.
- Tong-I Chen, Zhengxi Lu, Zhankai Xu, G. Shao, Shao-Peng Zhao, Fei Tang, Yong Du, Kaitao Song, Yizhou Liu, Yuchen Yan, Wenqi Zhang, Xu Tan, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. KnowU-Bench: Towards interactive, proactive, and personalized mobile agent evaluation, 2026g. URL <https://arxiv.org/abs/2604.08455>.
- Tuochao Chen, Nicholas Scott Batchelder, Alisa Liu, Noah A. Smith, and Shyamnath Gollakota. LlamaPIE: Proactive in-ear conversation assistants. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13801–13824, Vienna, Austria, July 2025d. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.710. URL <https://aclanthology.org/2025.findings-acl.710/>.
- Valerie Chen, Alan Zhu, Sebastian Zhao, Hussein Mozannar, David Sontag, and Ameet Talwalkar. Need help? designing proactive AI assistants for programming. In Naomi Yamashita, Vanessa Evers, Koji Yatani, Sharon Xianghua Ding, Bongshin Lee, Marshini Chetty, and Phoebe O. Touns Dugas (eds.), *Proceedings of the 2025 CHI Conference on*

- Human Factors in Computing Systems*, pp. 881:1–881:18. Association for Computing Machinery, April 2025e. doi: 10.1145/3706598.3714002. URL <https://doi.org/10.1145/3706598.3714002>.
- Xi Chen, Mao Mao, Shuo Li, and Haotian Shangguan. Debate-feedback: A multi-agent framework for efficient legal judgment prediction. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 462–470, Albuquerque, New Mexico, April 2025f. Association for Computational Linguistics. ISBN 979-8-89176-190-2. doi: 10.18653/v1/2025.naacl-short.39. URL <https://aclanthology.org/2025.naacl-short.39/>.
- Xiaoyu Chen, Shenao Zhang, Pushi Zhang, Li Zhao, and Jianyu Chen. Asking before acting: Gather information in embodied decision making with language models, 2024b. URL <https://arxiv.org/abs/2305.15695>.
- Yanyi Chen and Min Deng. From perception to symbolic task planning: Vision-language guided human-robot collaborative structured assembly. *arXiv preprint arXiv:2601.00978*, 2026.
- Yiheng Chen, Lingyao Li, Zihui Ma, Qikai Hu, Yilun Zhu, Min Deng, and Runlong Yu. Empowering llm agents with geospatial awareness: Toward grounded reasoning for wildfire response. *arXiv preprint arXiv:2510.12061*, 2025g.
- Yulin Chen, Ning Ding, Hai-Tao Zheng, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Empowering private tutoring by chaining large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 354–364, 2024c.
- Yuxin Chen, Yi Zhang, Zhengzhou Cai, Yaorui Shi, Zhiyuan Yao, Chenhang Cui, Jingnan Zheng, Yaqi Huo, Xi Su, Qi Gu, Xunliang Cai, Xiang Wang, An Zhang, and Tat-Seng Chua. VitaBench 2.0: Evaluating personalized and proactive agents in long-term user interactions. *arXiv preprint arXiv:2605.27141*, 2026h. doi: 10.48550/arXiv.2605.27141. URL <https://arxiv.org/abs/2605.27141>.
- Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. Tombench: Benchmarking theory of mind in large language models, 2024d. URL <https://arxiv.org/abs/2402.15052>.
- Zixing Chen, Yifeng Gao, Li Wang, Yunhan Zhao, Yi Liu, Jiayu Li, Xiang Zheng, Zuxuan Wu, Cong Wang, Xingjun Ma, and Yu-Gang Jiang. HazardArena: Evaluating semantic safety in vision-language-action models. *arXiv preprint arXiv:2604.12447*, 2026i.
- Sahana Chennabasappa, Cyrus Nikolaidis, Daniel Song, David Molnar, Stephanie Ding, Shengye Wan, Spencer Whitman, Lauren Deason, Nicholas Doucette, Abraham Montilla, Alekhya Gampa, Beto de Paola, Dominik Gabi, James Crnkovich, Jean-Christophe Testud, Kat He, Rashnil Chaturvedi, Wu Zhou, and Joshua Saxe. Llamafirewall: An open source guardrail system for building secure ai agents, 2025. URL <https://arxiv.org/abs/2505.03574>.
- Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory, 2025. URL <https://arxiv.org/abs/2504.19413>.
- Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.
- Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie Liu, Zhendong Wang, Zhengyuan Yang, Hung yi Lee, and Lijuan Wang. Shanks: Simultaneous hearing and thinking for spoken language models, 2025. URL <https://arxiv.org/abs/2510.06917>.
- Kitti Chiewchan, Patricia Anthony, Birendra Kc, and Sandhya Samarasinghe. Water distribution in community irrigation using a multi-agent system. *Journal of the Royal Society of New Zealand*, 53:6 – 26, 2022. URL <https://api.semanticscholar.org/CorpusID:253002143>.
- Rohan Chitnis, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Learning what information to give in partially observed domains, 2018. URL <https://arxiv.org/abs/1805.08263>.
- Daewon Choi, Kyunghyun Park, Woomin Song, Saket Dingliwal, Sai Muralidhar Jayanthi, Jinwoo Shin, and Aram Galstyan. Idlespec: Exploiting idle time via speculative planning for llm agents, 2026a. URL <https://arxiv.org/abs/2605.22154>.

- Dongwook Choi, Taeyoon Kwon, Bogyung Jeong, Minju Kim, Yeonjun Hwang, Hyojun Kim, Byungchul Kim, Young Kyun Jang, and Jinyoung Yeo. EMBGuard: Constructing hazard-aware guardrails for safe planning in embodied agents. *arXiv preprint arXiv:2605.30924*, 2026b.
- Youngwon Choi, Donghyuk Jung, and Hwayeon Kim. Desamo: A device for elder-friendly smart homes powered by embedded llm with audio modality. In *Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct)*, 2025. doi: 10.1145/3746058.3758344.
- Alankrit Chona, Igor Kozlov, and Ambuj Kumar. Cyber defense benchmark: Agentic threat hunting evaluation for LLMs in SecOps, 2026. URL <https://arxiv.org/abs/2604.19533>. arXiv preprint.
- Deepro Choudhury, Sinead Williamson, Adam Goliński, Ning Miao, Freddie Bickford Smith, Michael Kirchhof, Yizhe Zhang, and Tom Rainforth. Bed-llm: Intelligent information gathering with llms and bayesian experimental design, 2026. URL <https://arxiv.org/abs/2508.21184>.
- C. Chow. On optimum recognition error and reject tradeoff, 1970. URL <https://doi.org/10.1109/tit.1970.1054406>.
- Shiwei Chu and Wenxia Bao. Dynamic agricultural pest classification using enhanced sao-cnn and swarm intelligence optimization for uavs. *International Journal of Cognitive Computing in Engineering*, 2025. URL <https://api.semanticscholar.org/CorpusID:277620091>.
- Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jingheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S. Yu, and Qingsong Wen. LLM agents for education: Advances and applications. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13782–13810, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.743. URL <https://aclanthology.org/2025.findings-emnlp.743/>.
- CnEVPost. Baidu’s Apollo Go robotaxi hits 300,000 weekly rides as service expands to South Korea. <https://cnevpost.com/2026/02/27/baidu-apollo-go-robotaxi-300000-weekly-rides-expands-to-south-korea/>, February 2026. Accessed: 2026-10-05.
- Ronald H. Coase. The nature of the firm. *Economica*, 4(16):386–405, 1937. doi: 10.1111/j.1468-0335.1937.tb00002.x.
- Francesca Cocchella, Nilay Roy Choudhury, Eric Chen, and Patrícia Alves-Oliveira. Artists’ views on robotics involvement in painting productions. *arXiv preprint arXiv:2510.07063*, 2025.
- Jonathan Albert Cohen, Kye Shimizu, Allen Song, Vishnu Bharath, Kent Larson, and Pattie Maes. Do robots need body language? comparing communication modalities for legible motion intent in human-shared spaces. *arXiv preprint arXiv:2604.03451*, 2026.
- Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2–3):213–261, 1990. doi: 10.1016/0004-3702(90)90055-5. URL [https://doi.org/10.1016/0004-3702\(90\)90055-5](https://doi.org/10.1016/0004-3702(90)90055-5).
- Cédric Colas, Tristan Karch, Clément Moulin-Frier, and Pierre-Yves Oudeyer. Language and culture internalisation for human-like autotelic ai, 2022a. URL <https://arxiv.org/abs/2206.01134>.
- Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey, 2022b. URL <https://arxiv.org/abs/2012.09830>.
- Fulvio Corno, Luigi De Russis, and Alberto Monge Roffarello. Empowering end users in debugging trigger-action rules. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)*, 2019a. doi: 10.1145/3290605.3300618.
- Fulvio Corno, Luigi De Russis, and Alberto Monge Roffarello. Recrules: Recommending if-then rules for end-user development. *ACM Trans. Intell. Syst. Technol.*, 10(5), September 2019b. ISSN 2157-6904. doi: 10.1145/3344211. URL <https://doi.org/10.1145/3344211>.
- CrowdStrike. External technical root cause analysis — channel file 291. Technical report, CrowdStrike, August 2024. URL <https://www.crowdstrike.com/wp-content/uploads/2024/08/Channel-File-291-Incident-Root-Cause-Analysis-08.06.2024.pdf>. Published 2024-08-06.

- Christopher Cruz. Vigil: A reflective runtime for self-healing agents, 2025. URL <https://arxiv.org/abs/2512.07094>.
- Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks, 2025. URL <https://arxiv.org/abs/2502.08235>.
- Hongwei Cui, Yuyang Du, Qun Yang, Yulin Shao, and Soung Chang Liew. LLMind: Orchestrating AI and IoT with LLM for complex task execution. *arXiv preprint arXiv:2312.09007*, 2023a.
- Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *arXiv preprint arXiv:2306.16092*, 2023b.
- Sonia Da Costa Alves and Thomas Le Goff. Domestic robots and the right to privacy: the case of NEO by 1X Technologies. <https://lawandtech.ie/domestic-robots-and-the-right-to-privacy-the-case-of-neo-by-1x-technologies/>, February 2026. Law and Tech. Accessed: 2026-10-05.
- Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative ai, 2020. URL <https://arxiv.org/abs/2012.08630>.
- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *Proceedings of the European Conference on Computer Vision*, pp. 720–736, 2018. doi: 10.1007/978-3-030-01225-0_44. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Dima_Damen_Scaling_Egocentric_Vision_ECCV_2018_paper.html.
- John Danaher. *Automation and Utopia: Human Flourishing in a World without Work*. Harvard University Press, Oct 2019. ISBN 9780674242203. doi: 10.4159/9780674242203. URL <http://dx.doi.org/10.4159/9780674242203>.
- Danco Davcev, Slobodan Kalajdziski, Ivica Dimitrovski, Ivan Kitanovski, and Kosta Mitreski. Agentic ai-based iot precision agriculture framework—our vision and challenges. *AgriEngineering*, 2026. URL <https://api.semanticscholar.org/CorpusID:287414641>.
- Štěpán Davidovič and Betsy Beyer. Canary analysis service: Automated canarying quickens development, improves production safety, and helps prevent outages. *Queue*, 16(1):35–57, February 2018. ISSN 1542-7749. doi: 10.1145/3194653.3194655. URL <http://dx.doi.org/10.1145/3194653.3194655>.
- Bram de Jonge and Philip A. Scarf. A review on maintenance optimization. *European Journal of Operational Research*, 285(3):805–824, September 2020. ISSN 0377-2217. doi: 10.1016/j.ejor.2019.09.047. URL <http://dx.doi.org/10.1016/j.ejor.2019.09.047>.
- Hudson de Martim. Deterministic legal agents: A canonical primitive api for auditable reasoning over temporal knowledge graphs. *arXiv preprint arXiv:2510.06002*, 2025.
- Ewart J. de Visser, Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. Towards a theory of longitudinal trust calibration in human-robot teams. *Int. J. Soc. Robotics*, 12(2): 459–478, 2020. doi: 10.1007/S12369-019-00596-X. URL <https://doi.org/10.1007/s12369-019-00596-x>.
- Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2636. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/97091a5177d8dc64b1da8bf3e1f6fb54-Abstract-Dataset_s_and_Benchmarks_Track.html.
- Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design, 2025. URL <https://arxiv.org/abs/2503.18813>. arXiv preprint; version 2 consulted.
- Google DeepMind. Veo: A state-of-the-art generative video model by google deepmind. <https://deepmind.google/models/veo/>, 2025. Accessed: July 10, 2026.

- Defense Advanced Research Projects Agency. AI Cyber Challenge marks pivotal inflection point for cyber defense. DARPA news release, August 2025. URL <https://www.darpa.mil/news/2025/aixcc-results>. Published 2025-08-08; accessed 2026-10-03.
- Eugene Demaitre. Boston Dynamics opens Metaplant Application Center to train Atlas humanoids. <https://www.the-robot-report.com/boston-dynamics-opens-metaplant-application-center-train-atlas-humanoid-robots/>, September 2026. The Robot Report. Accessed: 2026-10-05.
- Gelei Deng, Yi Liu, Víctor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, and Stefan Rass. PentestGPT: Evaluating and harnessing large language models for automated penetration testing. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 847–864, Philadelphia, PA, August 2024a. USENIX Association. ISBN 978-1-939133-44-1. URL <https://www.usenix.org/conference/usenixsecurity24/presentation/deng>.
- Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a Generalist Agent for the Web, 2023a. URL <https://arxiv.org/abs/2306.06070>.
- Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. PACIFIC: Towards proactive conversational question answering over tabular and textual data in finance. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6970–6984, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.469. URL <https://aclanthology.org/2022.emnlp-main.469/>.
- Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. A survey on proactive dialogue systems: Problems, methods, and prospects, 2023b. URL <https://arxiv.org/abs/2305.02750>.
- Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10602–10621, Singapore, December 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.711. URL <https://aclanthology.org/2023.findings-emnlp.711/>.
- Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. Towards human-centered proactive conversational agents. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (eds.), *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pp. 807–818. ACM, 2024b. doi: 10.1145/3626772.3657843. URL <https://doi.org/10.1145/3626772.3657843>.
- Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. Plug-and-play policy planner for large language model powered dialogue agents, 2024c. URL <https://arxiv.org/abs/2311.00262>.
- Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 251–258. IEEE, 2013.
- Sebastian Deterding, Jonathan Hook, Rebecca Fiebrink, Marco Gillies, Jeremy Gow, Memo Akten, Gillian Smith, Antonios Liapis, and Kate Compton. Mixed-initiative creative interfaces. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, pp. 628–635, 2017.
- Prathamesh Devadiga, Omkaar Jayadev Shetty, and Pooja Agarwal. Samvad: A multi-agent system for simulating judicial deliberation dynamics in india. *arXiv preprint arXiv:2509.03793*, 2025.
- Dexterity. FedEx and Dexterity Expand Physical AI Deployment for Autonomous Trailer Loading at Hagerstown Hub. <https://dexterity.ai/blog/fedex-hagerstown-physical-ai-deployment>, July 2026. Accessed: 2026-10-05.
- Thomas G. Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. *J. Artif. Intell. Res.*, 13:227–303, 2000. doi: 10.1613/JAIR.639. URL <https://doi.org/10.1613/jair.639>.
- Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Overcoming algorithm aversion: People will use

- imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170, 2018. doi: 10.1287/mnsc.2016.2643.
- Digit. OpenAI is retiring ChatGPT pulse and replacing it with scheduled tasks, here is why. <https://www.digit.in/news/general/openai-is-retiring-chatgpt-pulse-and-replacing-it-with-scheduled-tasks-here-is-why.html>, 2026.
- Caiwu Ding, Tao Cui, Lingyun Wang, and Chengtao Wen. Toward certified functional safety for industrial humanoid robots: The fail-passive gap and a feasibility study. *arXiv preprint arXiv:2608.02809*, 2026a.
- Lei Ding, Bin He, Chenguang Wang, and Yang Liu. ProActor: Timing-aware reinforcement learning for proactive task scheduling agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18257–18303, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.832. URL <https://aclanthology.org/2026.acl-long.832/>.
- Zhijie Ding, Weinan Hong, Zicheng Zhu, Lei Li, Dezhi Kong, Hao Wang, Peng Zhou, Xuchu Jiang, and Jiaming Xu. Perceive Before Reasoning: A Pre-Reasoning Perception Framework for Efficient and Reliable Proactive Mobile Agents, 2026c. URL <https://arxiv.org/abs/2606.03236>.
- Dirac. How Dirac and AWS Are Transforming Process Design: BuildOS Delivers AI-Driven Manufacturing at Software Speed. <https://www.webull.com/news/15508677710177280>, September 2026. Accessed: 2026-10-05.
- Liming Dong, Qinghua Lu, and Liming Zhu. AgentOps: Enabling observability of LLM agents. *arXiv preprint arXiv:2411.05285*, 2024. doi: 10.48550/arXiv.2411.05285. URL <https://arxiv.org/abs/2411.05285>.
- Xiaoyu Dong, Zhi Li, and Xiao-Ming Wu. MUSE: Benchmarking manufacturable, functional, and assemblable text-to-CAD generation. *arXiv preprint arXiv:2605.28579*, 2026a.
- Zihan Dong, Yuanzhe Liu, Zhiyuan Ma, Qishi Zhan, Dehan Kong, Guohao Li, and Kaixin Li. Cadworld: Computer-use benchmark for long-horizon computer-aided design, 2026b. URL <https://arxiv.org/abs/2609.16251>.
- Aarya Doshi, Yining Hong, Congying Xu, Eunsuk Kang, Alexandros Kapravelos, and Christian Kästner. Towards verifiably safe tool use for llm agents, 2026. URL <https://arxiv.org/abs/2601.08012>.
- Anil R Doshi and Oliver P Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290, 2024.
- Anca D. Dragan, Kenton C. T. Lee, and Siddhartha S. Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 301–308. IEEE, IEEE, March 2013. doi: 10.1109/HRI.2013.6483603. URL <https://doi.org/10.1109/HRI.2013.6483603>.
- Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.
- Feiyu Duan, Xuanjing Huang, and Zhongyu Wei. LifeSim: Long-horizon user life simulator for personalized assistant evaluation. *arXiv preprint arXiv:2603.12152*, 2026a. doi: 10.48550/arXiv.2603.12152. URL <https://arxiv.org/abs/2603.12152>.
- Yuxiang Duan, Huining Li, Ao Li, Shuai Feng, Lanju Kong, Ning Liu, Jian Zhang, Xingdong Sheng, and Yuntao Du. AgenticVAU: Multi-agent explore-verify reasoning for video anomaly understanding, August 2026b. URL <https://arxiv.org/abs/2608.03779>. arXiv preprint, version 1.
- Mateusz Dubiel, Luis A. Leiva, Kerstin Bongard-Blanchy, and Anastasia Sergeeva. “hey genie, you got me thinking about my menu choices!” impact of proactive feedback on user perception and reflection in decision-making tasks. *ACM Transactions on Computer-Human Interaction*, 31(5):1–30, October 2024. ISSN 1557-7325. doi: 10.1145/3685274. URL <http://dx.doi.org/10.1145/3685274>.
- Marlon Dumas, Fredrik Milani, and David Chapela-Campa. Agentic business process management systems, 2026. URL <https://arxiv.org/abs/2601.18833>.

- Andrius Dzedzickis, Jurga Subačiūtė-Žemaitienė, Ernestas Šutinys, Urtė Samukaitė-Bubnienė, and Vytautas Bučinskas. Advanced applications of industrial robotics: New trends and possibilities. *Applied Sciences*, 12(1), 2022. ISSN 2076-3417. doi: 10.3390/app12010135. URL <https://www.mdpi.com/2076-3417/12/1/135>.
- Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue, 2024. URL <https://arxiv.org/abs/2410.00037>.
- Sven Eberhardt. The science behind hunches: Deep device embeddings. *Amazon Science, Amazon Science*, 2022. URL <https://www.amazon.science/blog/the-science-behind-hunches-deep-device-embeddings>.
- Aymen Echarchaoui, Dongxia Wu, and Emily B. Fox. Balar : A bayesian agentic loop for active reasoning, 2026. URL <https://arxiv.org/abs/2605.05386>.
- Niv Eckhaus, Uri Berger, and Gabriel Stanovsky. Time to talk: Llm agents for asynchronous group communication in mafia games, 2025. URL <https://arxiv.org/abs/2506.05309>.
- Nicholas Edwards and Sebastian Schuster. Ask or assume? uncertainty-aware clarification-seeking in coding agents, 2026. URL <https://arxiv.org/abs/2603.26233>.
- Electrek. Tesla starts Robotaxi rides without safety monitor in Austin: what you need to know. <https://electrek.co/2026/01/22/tesla-starts-robotaxi-rides-without-safety-monitor-in-austin-what-you-need-to-know/>, January 2026a. Accessed: 2026-10-05.
- Electrek. Tesla announces Robotaxi has driven 1 million unsupervised miles. <https://electrek.co/2026/09/03/tesla-announces-1-million-unsupervised-miles-driven-by-robotaxi/>, September 2026b. Accessed: 2026-10-05.
- Charles Elkan. The foundations of cost-sensitive learning. In Bernhard Nebel (ed.), *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA, August 4-10, 2001*, pp. 973–978. Morgan Kaufmann, 2001.
- Suveen Ellawela, Sashenka Gamage, and Dinithi Dissanayake. Proactive agentic whiteboards: Enhancing diagrammatic learning, 2025. URL <https://arxiv.org/abs/2512.01234>.
- Harry Emerson, Liz Bates, Chris Hicks, and Vasilios Mavroudis. CybORG++: An enhanced gym for the development of autonomous cyber agents, 2024. URL <https://arxiv.org/abs/2410.16324>.
- Mica R. Endsley and Esin O. Kiris. The out-of-the-loop performance problem and level of control in automation. *Human Factors*, 37(2):381–394, 1995. doi: 10.1518/001872095779064555.
- Engadget. The robots we saw at CES 2026: The lovable, the creepy and the utterly confusing. <https://www.engadget.com/ai/the-robots-we-saw-at-ces-2026-the-lovable-the-creepy-and-the-utterly-confusing-153537930.html>, January 2026. Accessed: 2026-10-05.
- Gal Engelberg, Michael Arenzon, and Leon Goldberg. Open security benchmark: Towards autonomous enterprise cyber defense, July 2026. URL <https://arxiv.org/abs/2607.27288>. arXiv preprint, version 1.
- Kai Epstude and Neal J. Roese. The functional theory of counterfactual thinking, 2008. URL <https://doi.org/10.1177/1088868308316091>.
- Kutluhan Erol, James A. Hendler, and Dana S. Nau. HTN planning: Complexity and expressivity. In Barbara Hayes-Roth and Richard E. Korf (eds.), *Proceedings of the 12th National Conference on Artificial Intelligence, Seattle, WA, USA, July 31 - August 4, 1994, Volume 2*, pp. 1123–1128. AAAI Press / The MIT Press, 1994. URL <http://www.aaai.org/Library/AAAI/1994/aaai94-173.php>.
- Ahmadreza Eslaminia, Chuhan Cai, Cameron Smith, Ruo-Syuan Mei, Shichen Li, Rajiv Malhotra, Klara Nahrstedt, and Chenhui Shao. LLM-ADAM: A generalizable LLM agent framework for pre-print anomaly detection in additive manufacturing. *arXiv preprint arXiv:2605.03328*, 2026.
- Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9710–9719, 2021.

- EU AI Act Resource. Digital Omnibus on AI: What Changed in the EU AI Act. <https://www.euaiact.com/digital-omnibus-ai>, July 2026. Regulation (EU) 2026/1744. Accessed: 2026-10-05.
- Ivan Evtimov, Arman Zharmagambetov, Aaron Grattafiori, Chuan Guo, and Kamalika Chaudhuri. WASP: Benchmarking Web Agent Security Against Prompt Injection Attacks, 2025. URL <https://arxiv.org/abs/2504.18575>.
- Siqi Fan, Minghao Li, Xiaoqian Ma, Wenhui Tan, Xiusheng Huang, Juntong Wu, Liujie Zhang, Shuo Shang, and Weihang Chen. Screenshots or tools? eliciting tool use and managing multimodal context in hybrid gui-mcp computer-use agents, 2026a. URL <https://arxiv.org/abs/2608.03327>.
- Yue Fan, Handong Zhao, Ruiyi Zhang, Yu Shen, Xin Eric Wang, and Gang Wu. GUI-bee: Align GUI action grounding to novel environments via autonomous exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33261–33278, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1688. URL <https://aclanthology.org/2025.emnlp-main.1688/>.
- Zhaoxin Fan, Tianbao Zhang, Wenjun Wu, Xiaofeng Wang, Yeying Jin, Jian Zhao, Zheng Zhu, and Shuicheng Yan. Drive-hwm: Hierarchical world models for dynamic-latent guided autonomous driving. *arXiv preprint arXiv:2609.03572*, 2026b.
- Chenhao Fang, Yanqing Peng, Rajeev Rao, Matt Sarmiento, Wendy Summer, Arya Pudota, Alex Goncalves, Jordi Mola, and Hervé Robert. Privacy artifact connector (pact): Embedding enterprise artifacts for compliance ai agents. *arXiv preprint arXiv:2507.21142*, 2025.
- Fei Fang, Thanh Nguyen, Rob Pickles, Wai Lam, Gopalasamy Clements, Bo An, Amandeep Singh, Milind Tambe, and Andrew Lemieux. Deploying paws: Field optimization of the protection assistant for wildlife security. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(2):3966–3973, February 2016. ISSN 2159-5399. doi: 10.1609/aaai.v30i2.19070. URL <http://dx.doi.org/10.1609/aaai.v30i2.19070>.
- Jinrui Fang, Runhan Chen, Xu Yang, Jian Yu, Jiawei Xu, Ashwin Vinod, Wenqi Shi, Tianlong Chen, Heng Ji, ChengXiang Zhai, Ying Ding, and Yuji Zhang. Benchmarking multi-turn medical diagnosis: Hold, lure, and self-correction. *ArXiv*, abs/2604.04325, 2026a. URL <https://api.semanticscholar.org/CorpusID:287204916>.
- Kuofei Fang, Zheng Lian, Xinyi Che, Haomin Ouyang, Shufan Zhang, Xuehao Wang, Qi Liu, Liyi Liu, Chenqi Zhang, Wenxi Cai, et al. RobotEQ: Towards social proactive intelligence in embodied agents. *arXiv preprint arXiv:2605.06234*, 2026b.
- Zeyu Fang, Mahdi Imani, and Tian Lan. Mint: Minimal information neuro-symbolic tree for objective-driven knowledge-gap reasoning and active elicitation, 2026c. URL <https://arxiv.org/abs/2602.05048>.
- Zeyu Fang, Yuxin Lin, Cheng Liu, Beomyeol Yu, Zeyuan Yang, Rongqian Chen, Taeyoung Lee, Mahdi Imani, and Tian Lan. Uncertainty mitigation and intent inference: A dual-mode human-machine joint planning system, 2026d. URL <https://arxiv.org/abs/2603.07822>.
- Zheng Fang, Dongming Jin, Yihong Dong, Yongmin Li, Kechi Zhang, Zhi Jin, and Ge Li. ClarifyCodeBench: Evaluating LLMs on clarifying ambiguous requirements for code generation, 2026e. URL <https://arxiv.org/abs/2607.00711>.
- Faraz Faruqi, Ahmed Katary, Demircan Tas, Theresa Hradilak, Ning Zhang, Jiayi Li, Fabian Manhardt, Martin Nisser, Vrushank Phadnis, Ruofei Du, Federico Tombari, Megan Hofmann, and Stefanie Mueller. InstructMesh: Selective refinement of generative 3D models for fabrication. *arXiv preprint arXiv:2608.28534*, 2026.
- Andrew Faulring, Brad Myers, Ken Mohnkern, Bradley Schmerl, Aaron Steinfeld, John Zimmerman, Asim Smailagic, Jeffery Hansen, and Daniel Siewiorek. Agent-assisted task management that reduces email overload. In *Proceedings of the 15th international conference on Intelligent user interfaces*, IUI '10, pp. 61–70. ACM, February 2010. doi: 10.1145/1719970.1719980. URL <http://dx.doi.org/10.1145/1719970.1719980>.
- Francesca Favarò, Sky Eurich, and Nazanin Nader. Autonomous vehicles’ disengagements: Trends, triggers, and regulatory limitations. *Accident Analysis & Prevention*, 110:136–148, January 2018. ISSN 0001-4575. doi: 10.1016/j.aap.2017.11.001. URL <http://dx.doi.org/10.1016/j.aap.2017.11.001>.

- Anthony Favier, Shashank Shekhar, and Rachid Alami. Robust planning for human-robot joint tasks with explicit reasoning on human mental state, 2022. URL <https://arxiv.org/abs/2210.08879>.
- Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.
- Yanlin Fei, Nazhou Liu, Xinmiao Yu, Shaolong Chen, Lei Li, Rahul Thapa, Madalina Ciobanu, Navan Preet Singh, Qingqing Mao, and Ritankar Das. How do agents fail on autoresearch: End-to-end diagnostic evaluation on 100 real-world frontier research tasks, 2026. URL <https://arxiv.org/abs/2608.14905>.
- K. J. Kevin Feng, David W. McDonald, and Amy X. Zhang. Levels of autonomy for ai agents, 2025. URL <https://arxiv.org/abs/2506.12469>.
- Shiyang Feng, Runmin Ma, Xiangchao Yan, Yue Fan, Yusong Hu, Songtao Huang, Shuaiyu Zhang, Zongsheng Cao, Tianshuo Peng, Jiakang Yuan, Zijie Guo, Zhijie Zhong, Shangheng Du, Weida Wang, Jinxin Shi, Yuhao Zhou, Xiaohan He, Zhiyin Yu, Fangchen Yu, Qihao Zheng, Jiamin Wu, Mianxin Liu, Chi Zhang, Shaowei Hou, Shuya Li, Yankai Jiang, Wenjie Lou, Lilong Wang, Zifu Wang, Jiong Wang, Wanghan Xu, Yue Deng, Dongrui Liu, Yiheng Wang, Wenlong Zhang, Fenghua Ling, Shufei Zhang, Xiaosong Wang, Shuangjia Zheng, Xun Huang, Siqi Sun, Shuyue Hu, Peng Ye, Chunfeng Song, Bin Wang, Conghui He, Yihao Liu, Xin Li, Qibin Hou, Tao Chen, Xiangyu Yue, Bin Wang, Liang He, Dahua Lin, Bowen Zhou, Bo Zhang, and Lei Bai. Internagent-1.5: A unified agentic framework for long-horizon autonomous scientific discovery, 2026a. URL <https://arxiv.org/abs/2602.08990>.
- Xingyu Feng, Chang Sun, Yuzhu Wang, Zhangbing Zhou, Chengwen Luo, Zhuangzhuang Chen, Xiaomin Ouyang, and Huanqi Yang. Powerlens: Taming llm agents for safe and personalized mobile power management, 2026b. URL <https://arxiv.org/abs/2603.19584>.
- Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. Large language model-based human-agent collaboration for complex task solving. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1336–1357. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.72. URL <http://dx.doi.org/10.18653/v1/2024.findings-emnlp.72>.
- Xueyang Feng, Jiakai Tang, Xu Chen, Quanyu Dai, and Zhenhua Dong. Optimizing multi-turn interactive recommendation agents via generative intrinsic motivation. In *Proceedings of the ACM Web Conference 2026*, pp. 5975–5986. ACM, April 2026c. doi: 10.1145/3774904.3792209. URL <http://dx.doi.org/10.1145/3774904.3792209>.
- Yichun Feng, Jiawei Wang, Lu Zhou, Yikai Zheng, Zhen Lei, and Yixue Li. Real-world doctor agent with proactive consultation through multi-agent reinforcement learning, 2026d. URL <https://arxiv.org/abs/2505.19630>.
- K. P. Ferentinos, Konstantinos G. Arvanitis, D. Lambrou, A. Anastasiou, and Nick Sigrimis. A multi-agent system for integrated production in greenhouse hydroponics. *Acta Horticulturae*, 691:381–388, October 2005. doi: 10.17660/ActaHortic.2005.691.45. URL <https://doi.org/10.17660/ActaHortic.2005.691.45>.
- George Ferguson, James F Allen, Bradford W Miller, et al. Trains-95: Towards a mixed-initiative planning assistant. In *AIPS*, pp. 70–77, 1996.
- Figure AI. F.03 Arrives at BMW. <https://www.figure.ai/news/f-03-at-bmw>, June 2026a. Accessed: 2026-07-25.
- Figure AI. Helix 02 Bedroom Tidy. <https://www.figure.ai/news/helix-02-bedroom-tidy>, May 2026b. Accessed: 2026-07-20.
- Figure AI. Helix 2.5: Zero-Shot 30-Home Generalization. <https://www.figure.ai/news/helix-2-5-zero-shot-30-home-generalization>, September 2026c. Accessed: 2026-10-05.
- Figure AI. Introducing Index: Building The World’s Largest and Most Diverse Physical Dataset. <https://www.figure.ai/news/introducing-index>, August 2026d. Accessed: 2026-10-05.
- Richard E Fikes and Nils J Nilsson. Strips: A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3-4):189–208, 1971.

- Julia Fink, Séverin Lemaignan, Pierre Dillenbourg, Philippe Réturnaz, Florian Vaussard, Alain Berthoud, Francesco Mondada, Florian Wille, and Karmen Franinović. Which robot behavior can motivate children to tidy up their toys? design and evaluation of "ranger". In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, HRI '14*, pp. 439–446, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450326582. doi: 10.1145/2559636.2559659. URL <https://doi.org/10.1145/2559636.2559659>.
- Sandro Rama Fiorini, Leonardo G. Azevedo, Raphael M. Thiago, Valesca M. de Sousa, Anton B. Labate, and Viviane Torres da Silva. Episodic memory in agentic frameworks: Suggesting next tasks, 2025. URL <https://arxiv.org/abs/2511.17775>.
- Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents, 2018. URL <https://arxiv.org/abs/1705.06366>.
- Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients, 2024. URL <https://arxiv.org/abs/1705.08926>.
- James Fogarty, Scott E. Hudson, Christopher G. Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny C. Lee, and Jie Yang. Predicting human interruptibility with sensors. *ACM Transactions on Computer-Human Interaction*, 12(1):119–146, March 2005. ISSN 1557-7325. doi: 10.1145/1057237.1057243. URL <http://dx.doi.org/10.1145/1057237.1057243>.
- Sébastien Forestier, Rémy Portelas, Yoan Mollard, and Pierre-Yves Oudeyer. Intrinsically motivated goal exploration processes with automatic curriculum learning. *Journal of Machine Learning Research*, 23(152):1–41, 2022. URL <https://jmlr.org/papers/v23/21-0808.html>.
- Christopher K. Fourie. Adaptive planning with evidence based prediction for improved fluency in routine human-robot collaborative tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9880–9881, July 2019. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.33019880. URL <http://dx.doi.org/10.1609/aaai.v33i01.33019880>.
- Maria Fox, Alfonso Gerevini, Derek Long, and Ivan Serina. Plan stability: Replanning versus plan repair. In Derek Long, Stephen F. Smith, Daniel Borrajo, and Lee McCluskey (eds.), *Proceedings of the Sixteenth International Conference on Automated Planning and Scheduling, ICAPS 2006, Cumbria, UK, June 6-10, 2006*, pp. 212–221. AAAI, 2006. URL <http://www.aaai.org/Library/ICAPS/2006/icaps06-022.php>.
- Sam Francis. IDC data shows Agibot led global humanoid robot shipments in H1 2026. <https://roboticsandautomationnews.com/2026/10/02/idc-data-shows-agibot-led-global-humanoid-robot-shipments-in-h1-2026/105463/>, October 2026. Robotics & Automation News. Accessed: 2026-10-05.
- Shelly Francis-Meretzki, Mirco Mutti, Yaniv Romano, and Aviv Tamar. Temporal difference calibration in sequential tasks: Application to vision-language-action models. *arXiv preprint arXiv:2604.20472*, 2026.
- Ismael T Freire, Oscar Guerrero-Rosado, Adrián F Amil, and Paul FMJ Verschure. Socially adaptive cognitive architecture for human-robot collaboration in industrial settings. *Frontiers in Robotics and AI*, 11:1248646, 2024.
- Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Silveira Cabral, Virginie Do, Emilien Garreau, Jean-Baptiste Gaya, Hugo Laurençon, Maxime Lecanu, Kunal Malkan, Dheeraj Mekala, Pierre Ménard, Gerard Moreno-Torres Bertran, Ulyana Piterbarg, Mikhail Plekhanov, Mathieu Rita, Andrey Rusakov, Vladislav Vorotilov, Mengjue Wang, Ian Yu, Amine Benhalloum, Grégoire Mialon, and Thomas Scialom. Gaia2: Benchmarking llm agents on dynamic and asynchronous environments. *arXiv preprint arXiv:2602.11964*, 2026. URL <https://arxiv.org/abs/2602.11964>.
- Jörg Froberg and Frank Binder. Crass: A novel data set and benchmark to test counterfactual reasoning of large language models, 2022. URL <https://arxiv.org/abs/2112.11941>.
- Kai Fronsdaal and David Lindner. Misr: Measuring instrumental self-reasoning in frontier models, 2024. URL <https://arxiv.org/abs/2412.03904>.
- Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video, 2020. URL <https://arxiv.org/abs/2005.02190>.
- Futurism. Chinese post office deploys humanoid robots to sort mail. Futurism, May 2026. URL <https://futurism.com/robots-and-machines/chinese-post-office-humanoid-robots-mail>.

- Jack Gallifant, Katherine C. Kellogg, Matt Butler, Amanda Centi, Shan Chen, Patrick F. Doyle, Sayon Dutta, Joyce Guo, Matthew J. Hadfield, Esther H. Kim, David E. Kozono, Hugo J. W. L. Aerts, Adam B. Landman, Raymond H. Mak, Rebecca G. Mishuris, Tanna L. Nelson, Guergana K. Savova, Elad Sharon, Benjamin C. Silverman, Umit Topaloglu, Jeremy L. Warner, and Danielle S. Bitterman. A field guide to deploying AI agents in clinical practice. *arXiv preprint arXiv:2509.26153*, 2025. URL <https://arxiv.org/abs/2509.26153>.
- Daniele Gammelli, Kaidi Yang, James Harrison, Filipe Rodrigues, Francisco C. Pereira, and Marco Pavone. Graph neural network reinforcement learning for autonomous mobility-on-demand systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 2996–3003, 2021. doi: 10.1109/CDC45484.2021.9683135.
- Bingzheng Gan, Yufan Zhao, Tianyi Zhang, Jing Huang, Li Yusu, Shu Xian Teo, Changwang Zhang, and Wei Shi. Master: A multi-agent system with llm specialized mcts. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9409–9426, 2025.
- Apurva Gandhi and Graham Neubig. Go-browse: Training web agents with structured exploration, 2025. URL <https://arxiv.org/abs/2506.03533>.
- Chenfeng Gao, Zeya Chen, Anjie Yang, Karan Ahuja, and Danli Luo. FabDreamer: Exploring the image-to-physical workflow through AI-assisted layered fabrication. *arXiv preprint arXiv:2608.13665*, 2026a.
- Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. Aligning llm agents by learning latent preference from user edits, 2024. URL <https://arxiv.org/abs/2404.15269>.
- Ning Gao, Jinliang Zheng, Xing Gao, Haoxiang Ma, Hanqing Wang, Yukai Wang, Jiantong Chen, Zanxin Chen, Shujie Zhang, Mingda Jia, et al. Ebench: Elemental diagnosis of generalist mobile manipulation policies. *arXiv preprint arXiv:2606.18239*, 2026b.
- Shanghua Gao, Ada Fang, and Marinka Zitnik. Autoscientists: Self-organizing agent teams for long-running scientific experimentation, 2026c. URL <https://arxiv.org/abs/2605.28655>.
- Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026d.
- Shujian Gao, Jiamei Yan, Yuchen Yang, Penghao Zhou, Qinglei Wang, Tiehan Fan, Yuan Wang, Zuxuan Wu, and Yu-gang Jiang. Live assistant: Learning whether, when, and whom to assist in real-world live social streams. *arXiv preprint arXiv:2609.27303*, 2026e.
- Shuzheng Gao, Wenhao Zeng, Zhaojian Yu, Jianqiao Wangni, Chaozheng Wang, Kai Cai, Shilin He, and Michael R. Lyu. SWE-MeM: Learning adaptive memory management for long-horizon coding agents, 2026f. URL <https://arxiv.org/abs/2606.28434>.
- Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.*, 16:1437–1480, 2015. doi: 10.5555/2789272.2886795. URL <https://dl.acm.org/doi/10.5555/2789272.2886795>.
- Quentin Garrido, Tushar Nagarajan, Basile Terver, Nicolas Ballas, Yann LeCun, and Michael Rabbat. Learning latent action world models in the wild. *arXiv preprint arXiv:2601.05230*, 2026.
- Gasgoo. China approves first L3 autonomous passenger cars for limited public deployment. <https://autonews.gasgoo.com/articles/news/china-approves-first-l3-autonomous-passenger-cars-for-limited-public-deployment-70039881>, December 2025. MIIT approval of Changan SC7000AAARBEV and ARCFOX BJ7001A61NBEV on December 15, 2025; accessed: 2026-10-05.
- Gasgoo. WeRide Posts 230 Million Yuan in Q2 Revenue, Overseas Revenue Surges 164% YoY. <https://autonews.gasgoo.com/articles/news/2088109006169329664>, August 2026. Accessed: 2026-10-05.
- Loris Gaven, Thomas Carta, Clément Romac, Cédric Colas, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. Magellan: Metacognitive predictions of learning progress guide autotelic llm agents in large goal spaces, 2025. URL <https://arxiv.org/abs/2502.07709>.

- Jonas Gehring, Kunhao Zheng, Jade Copet, Vegard Mella, Quentin Carbonneaux, Taco Cohen, and Gabriel Synnaeve. Rlef: Grounding code llms in execution feedback with reinforcement learning, 2025. URL <https://arxiv.org/abs/2410.02089>.
- Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks, 2017. URL <https://arxiv.org/abs/1705.08500>.
- Jasper Geldenbott and Karen Leung. Legible and proactive robot planning for prosocial human-robot interactions, 2024. URL <https://arxiv.org/abs/2404.03734>.
- Generalist AI. GEN-0: Embodied Foundation Models That Scale with Physical Interaction. <https://generalistai.com/blog/nov-04-2025-GEN-0>, November 2025. Accessed: 2026-10-05.
- Generalist AI. GEN-1: A generalist foundation model for embodied ai. Generalist AI Official Blog, April 2026. URL <https://generalistai.com/blog/apr-02-2026-GEN-1>.
- Genesis AI. GENE-26.5: Advancing robotic manipulation to human level. Genesis AI Blog, May 2026. URL <https://www.genesis.ai/blog/gene-26-5-advancing-robotic-manipulation-to-human-level>. Accessed: 2026-07-09.
- Jiayi Geng and Graham Neubig. Effective strategies for asynchronous software engineering agents, 2026. URL <https://arxiv.org/abs/2603.21489>.
- Alireza Ghafarollahi and Markus J. Buehler. Sparks: Multi-agent artificial intelligence model discovers protein design principles, 2025. URL <https://arxiv.org/abs/2504.19017>.
- Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J Szostkiewicz, Dmytro Shved, Gavin J Gyimesi, Jon M Laurent, Samantha M Wright, Muhammed T Razzak, et al. A multi-agent system for automating scientific discovery. *Nature*, 655:497–505, 2026. doi: 10.1038/s41586-026-10652-y. URL <https://www.nature.com/articles/s41586-026-10652-y>.
- Rohit Girdhar and Kristen Grauman. Anticipative video transformer, 2021. URL <https://arxiv.org/abs/2106.02036>.
- GitHub. GitHub introduces coding agent for GitHub Copilot. <https://github.com/newsroom/press-releases/coding-agent-for-github-copilot>, May 2025. Accessed: 2026-06-22.
- GitHub. Agentic autofix for code scanning alerts in public preview. <https://github.blog/changelog/2026-07-10-agentic-autofix-for-code-scanning-alerts-in-public-preview/>, July 2026a. GitHub Changelog.
- GitHub. Introducing the agents tab in your repository. <https://github.blog/changelog/2026-01-26-introducing-the-agents-tab-in-your-repository/>, January 2026b. GitHub Changelog.
- GitHub. Copilot Memory supports user preferences for Pro, Pro+ users. <https://github.blog/changelog/2026-05-15-copilot-memory-supports-user-preferences-for-pro-pro-users/>, May 2026c. Published: May 15, 2026; Accessed: 2026-10-06.
- GitHub. Trace any Copilot coding agent commit to its session logs. <https://github.blog/changelog/2026-03-20-trace-any-copilot-coding-agent-commit-to-its-session-logs/>, March 2026d. GitHub Changelog.
- GitHub. About github copilot cloud agent. <https://docs.github.com/copilot/concepts/agents/coding-agent/about-coding-agent>, n.d.a. Accessed: 2026-05-09.
- GitHub. Rolling back changes made during a GitHub Copilot CLI session. <https://docs.github.com/en/copilot/how-tos/copilot-cli/use-copilot-cli/roll-back-changes>, n.d.b. GitHub Docs, accessed July 28, 2026.
- Mathyas Giudici, Alessandro Sironi, Ismaele Villa, Samuele Scherini, and Franca Garzotto. Generating homeassistant automations using an llm-based chatbot. *arXiv preprint arXiv:2505.02802*, 2025.
- Thibaud Gloaguen, Niels Mündler, Mark Müller, Veselin Raychev, and Martin Vechev. Coding agents don’t know when to act, 2026. URL <https://arxiv.org/abs/2605.07769>.

- Andrew Goldberg, Kavish Kondap, Tianshuang Qiu, Zehan Ma, Letian Fu, Justin Kerr, Huang Huang, Kaiyuan Chen, Kuan Fang, and Ken Goldberg. Blox-net: Generative design-for-robot-assembly using vlm supervision, physics simulation, and a robot with reset. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15493–15500. IEEE, 2025.
- David Goldberg, David Nichols, Brian M Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.
- Anahita Golrang, Kshitij Sharma, et al. Propact: A proactive ai-driven adaptive collaborative tutor for pair programming. *arXiv preprint arXiv:2605.02703*, 2026.
- Matthew Gombolay, Xi Jessie Yang, Brad Hayes, Nicole Seo, Zixi Liu, Samir Wadhwan, Tania Yu, Neel Shah, Toni Golen, and Julie Shah. Robotic assistance in coordination of patient care. In *Robotics: Science and Systems XII*, RSS2016. Robotics: Science and Systems Foundation, 2016. doi: 10.15607/rss.2016.xii.026. URL <http://dx.doi.org/10.15607/rss.2016.xii.026>.
- David Gomes and Travis McPeak. Governing agent autonomy with Auto-review. <https://cursor.com/blog/agent-autonomy-auto-review>, June 2026. Published June 11, 2026.
- Sitong Gong, Tianyu Yan, Caixin Kang, Bo Zheng, Xiang Ruan, Huchuan Lu, Kaipeng Zhang, Yoichi Sato, and Yifei Huang. Vinci2: Providing proactive assistance in continuous egocentric videos. *arXiv preprint arXiv:2607.11523*, 2026.
- Alvaro Gonzalez, M. H. Hasan Shovo, and Ali Ayub. Personalized and robust proactive robot assistance with uncertainty-guided llm reasoning. *arXiv preprint arXiv:2606.08458*, 2026.
- Tyler J. Goodman, Michael E. Miller, Christina F. Rusnock, and Jason M. Bindewald. Effects of agent timing on the human-agent team. *Cognitive Systems Research*, 46:40–51, December 2017. ISSN 1389-0417. doi: 10.1016/j.cogsys.2017.02.007. URL <http://dx.doi.org/10.1016/j.cogsys.2017.02.007>.
- Google Cloud. Agentic AI for security operations. Product overview, 2026. URL <https://cloud.google.com/security/resources/agentic-soc>. Living product documentation; accessed 2026-09-13.
- Google DeepMind. Gemini Robotics 2 brings whole body intelligence to robots. <https://deepmind.google/blog/gemini-robotics-2-brings-whole-body-intelligence-to-robots/>, July 2026a. Accessed: 2026-10-05.
- Google DeepMind. Introducing Gemini Robotics ER 2. <https://blog.google/innovation-and-ai/models-and-research/google-deepmind/gemini-robotics-er-2/>, July 2026b. Accessed: 2026-10-05.
- Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with co-scientist. *Nature*, 655:487–496, 2026. doi: 10.1038/s41586-026-10644-y. URL <https://www.nature.com/articles/s41586-026-10644-y>.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing, 2024. URL <https://arxiv.org/abs/2305.11738>.
- Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abraham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, June 2022. doi: 10.1109/CVPR52688.2022.01842.

- URL https://openaccess.thecvf.com/content/CVPR2022/html/Grauman_Ego4D_Around_the_World_in_3000_Hours_of_Egocentric_Video_CVPR_2022_paper.html.
- Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16295–16336. PMLR, 2024. URL <https://proceedings.mlr.press/v235/greenblatt24a.html>.
- Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you’ve signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, AISec ’23, pp. 79–90. Association for Computing Machinery, 2023. doi: 10.1145/3605764.3623985. URL <https://doi.org/10.1145/3605764.3623985>.
- Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and Christina Mack. Agentic ai for scientific discovery: A survey of progress, challenges, and future directions, 2025. URL <https://arxiv.org/abs/2503.08979>.
- Nathan Grinsztajn, Johan Ferret, Olivier Pietquin, Philippe Preux, and Matthieu Geist. There is no turning back: A self-supervised approach for reversibility-aware reinforcement learning, 2021. URL <https://arxiv.org/abs/2106.04480>.
- Sanford J. Grossman and Oliver D. Hart. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy*, 94(4):691–719, 1986. doi: 10.1086/261404.
- Chenwei Gu, Qian Wu, Zeqi Lai, Hewu Li, Yuxuan Weng, Weisen Liu, Jihao Li, Jun Liu, and Yuanjie Li. Novaplan: An efficient plan of renting ground stations for emerging leo satellite networks. In *IEEE INFOCOM 2025-IEEE Conference on Computer Communications*, pp. 1–10. IEEE, 2025.
- Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms, 2026. URL <https://arxiv.org/abs/2410.13648>.
- Anmol Gulati, Hariom Gupta, Elias Lumer, Sahil Sen, and Vamse Kumar Subbiah. Ask early, ask late, ask right: When does clarification timing matter for long-horizon agents?, 2026. URL <https://arxiv.org/abs/2605.07937>.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017. URL <https://arxiv.org/abs/1706.04599>.
- Jun Guo, Qiwei Li, Peiyan Li, Zilong Chen, Nan Sun, Yifei Su, Heyun Wang, Yuan Zhang, Xinghang Li, and Huaping Liu. Unified 4d world action modeling from video priors with asynchronous denoising. *arXiv preprint arXiv:2604.26694*, 2026a.
- Qianyun Guo, Yibo Li, Yue Liu, and Bryan Hooi. Towards realistic personalization: Evaluating long-horizon preference following in personalized user–LLM interactions. *arXiv preprint arXiv:2603.04191*, 2026b. doi: 10.48550/arXiv.2603.04191. URL <https://arxiv.org/abs/2603.04191>.
- Shuchen Guo, Ehsan Latif, Yifan Zhou, Xuan Huang, and Xiaoming Zhai. Using generative ai and multi-agents to provide automatic feedback. *arXiv preprint arXiv:2411.07407*, 2024.
- Rohan Gupta, Trevor Asbery, Zain Merchant, Abrar Anwar, and Jesse Thomason. Robotfleet: An open-source framework for centralized multi-robot task planning. *arXiv preprint arXiv:2510.10379*, 2025.
- David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, pp. 2450–2462. Curran Associates, Inc., 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/2de5d16682c3c35007e4e92982f1a2ba-Abstract.html.
- Aaron Haag, Altay Kacan, Bertram Fuchs, and Oliver Lohse. Test-time scaling for CAD generation via verifier-free consensus selection. *arXiv preprint arXiv:2608.09706*, 2026.
- Gillian K. Hadfield. Legal infrastructure for transformative ai governance. *Proceedings of the National Academy of Sciences*, 123(30):e2509742123, July 2026. ISSN 1091-6490. doi: 10.1073/pnas.2509742123. URL <http://dx.doi.org/10.1073/pnas.2509742123>.

- Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 220–227, 2017. doi: 10.24963/ijcai.2017/32. URL <https://doi.org/10.24963/ijcai.2017/32>.
- Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models, 2022. URL <https://arxiv.org/abs/2010.02193>.
- Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, April 2025. doi: 10.1038/s41586-025-08744-2. URL <https://www.nature.com/articles/s41586-025-08744-2>.
- Dongge Han, Camille Couturier, Daniel Madrigal Diaz, Xuchao Zhang, Victor Rühle, and Saravan Rajmohan. LEGOMem: Modular Procedural Memory for Multi-agent LLM Systems for Workflow Automation, 2025a. URL <https://arxiv.org/abs/2510.04851>.
- Jindong Han, Yansong Ning, Zirui Yuan, Hang Ni, Fan Liu, Tengfei Lyu, and Hao Liu. Large language model powered intelligent urban agents: Concepts, capabilities, and applications. *arXiv preprint arXiv:2507.00914*, 2025b.
- Jinyi Han, Xinyi Wang, Haiquan Zhao, Tingyun li, Zishang Jiang, Sihang Jiang, Jiaqing Liang, Xin Lin, Weikang Zhou, Zeye Sun, Fei Yu, and Yanghua Xiao. A stitch in time saves nine: Proactive self-refinement for language models, 2025c. URL <https://arxiv.org/abs/2508.12903>.
- Sophia Simeng Han, Yoshiki Takashima, Shannon Zejiang Shen, Chen Liu, Yixin Liu, Roque K Thuo, Sonia Knowlton, Ruzica Piskac, Scott J Shapiro, and Arman Cohan. Courtreasoner: Can llm agents reason like judges? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35279–35294, 2025d.
- Tingxu Han, Yi Zhang, Wei Song, Chunrong Fang, Zhenyu Chen, Youcheng Sun, and Lijie Hu. SWE-Skills-Bench: Do agent skills actually help in real-world software engineering?, 2026a. URL <https://arxiv.org/abs/2603.15401>.
- Xiao Han, Pinbo Wang, Yuanshao Zhu, Guojiang Shen, and Xiangjie Kong. Mobiwave: Dispatch-oriented graph wavelets and drift-guided selective optimization for autonomous fleet rebalancing. *arXiv preprint arXiv:2607.24365*, 2026b.
- Yu Han and Tianwen Qian. Wm-r1: Training gui agents to reason and leverage world models with reinforcement learning, 2026. URL <https://arxiv.org/abs/2608.27508>.
- Uri Hanani, Bracha Shapira, and Peretz Shoval. Information filtering: Overview of issues, research and systems. *User modeling and user-adapted interaction*, 11(3):203–259, 2001.
- Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model, 2023. URL <https://arxiv.org/abs/2305.14992>.
- Yupu Hao, Pengfei Cao, Zhuoran Jin, Huanxuan Liao, Yubo Chen, Kang Liu, and Jun Zhao. Evaluating personalized tool-augmented llms from the perspectives of personalization and proactivity, 2025. URL <https://arxiv.org/abs/2503.00771>.
- Oliver Hart. *Firms, Contracts, and Financial Structure*. Clarendon Press, Oxford, 1995.
- Oliver Hart and John Moore. Incomplete contracts and renegotiation. *Econometrica*, 56(4):755–785, 1988. doi: 10.2307/1912698.
- Oliver Hart and John Moore. Property rights and the nature of the firm. *Journal of Political Economy*, 98(6):1119–1158, 1990. doi: 10.1086/261729.
- Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati (eds.), *Advances in Psychology*, volume 52, pp. 139–183. North-Holland, 1988. ISBN 9780444703880. doi: 10.1016/S0166-4115(08)62386-9.
- Ahmed E. Hassan, Hao Li, Dayi Lin, Bram Adams, Tse-Hsun Chen, Yutaro Kashiwa, and Dong Qiu. Agentic software engineering: Foundational pillars and a research roadmap, 2025. URL <https://arxiv.org/abs/2509.06216>.
- Wajih Ul Hassan, Shengjian Guo, Ding Li, Zhengzhang Chen, Kangkook Jee, Zhichun Li, and Adam Bates. NoDoze: Combatting threat alert fatigue with automated provenance triage. In *Network and Distributed System Security*

- Symposium (NDSS)*, 2019. URL https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_03B-1-3_UlHassan_paper.pdf.
- Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1), April 2022. ISSN 1573-7454. doi: 10.1007/s10458-022-09552-y. URL <http://dx.doi.org/10.1007/s10458-022-09552-y>.
- Chengbo He, Sheng Li, Chenyang Ma, Bochao Zou, Li Sun, Jiansheng Chen, Junliang Xing, Yuanchun Shi, and Huimin Ma. Pact: Proactive asking for continual task assistance in human-robot collaboration. *arXiv preprint arXiv:2605.24350*, 2026a.
- Tiantian He, Yihang Chen, Keyue Jiang, Ka Yiu Lee, Kaiwen Zhou, Kun Shao, and Shuai Wang. Ee-mcp: Self-evolving mcp-gui agents via automated environment generation and experience learning, 2026b. URL <https://arxiv.org/abs/2604.09815>.
- Yufei He, Ruoyu Li, Alex Chen, Yue Liu, Yulin Chen, Yuan Sui, Cheng Chen, Yi Zhu, Luca Luo, Frank Yang, and Bryan Hooi. Enabling self-improving agents to learn at test time with human-in-the-loop guidance, 2025a. URL <https://arxiv.org/abs/2507.17131>.
- Zhanpeng He, Yifeng Cao, and Matei Ciocarlie. Uncertainty comes for free: Human-in-the-loop policies with diffusion models, 2025b. URL <https://arxiv.org/abs/2503.01876>.
- Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Kang Liu, and Jun Zhao. AgentsCourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9399–9416, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.549. URL <https://aclanthology.org/2024.findings-emnlp.549/>.
- heise online. 1X to deliver humanoid household robot Neo to US customers in 2026. <https://www.heise.de/en/news/1X-to-deliver-humanoid-household-robot-Neo-to-US-customers-in-2026-11287205.html>, May 2026. Accessed: 2026-10-05.
- Fnu Heman, Yixuan Wang, Kolin Xu, Conner Wallace, John Dang, Akhil Joshi, Jun Sheng, Pinhas Ben-Tzvi, and Mingyu Cai. Uncertainty-aware intention prediction for human-to-robot assembly teleoperation. *arXiv preprint arXiv:2606.08341*, 2026.
- Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining collaborative filtering recommendations, 2000. URL <https://doi.org/10.1145/358916.358995>.
- Germund Hesslow. Conscious thought as simulation of behaviour and perception, 2002. URL [https://doi.org/10.1016/s1364-6613\(02\)01913-7](https://doi.org/10.1016/s1364-6613(02)01913-7).
- Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In *International Conference on Learning Representations*, 2016.
- Zoe Hitzig, Maxim Massenkoff, Eva Lyubich, Shaoyi Zhang, Ryan Heller, and Peter McCrory. Agentic coding and persistent returns to expertise. <https://www.anthropic.com/research/claude-code-expertise>, June 2026. Published June 16, 2026.
- Bo-Jhang Ho, Bharathan Balaji, Mehmet Koseoglu, Sandeep Sandha, Siyou Pei, and Mani Srivastava. Quick question: Interrupting users for microtasks with reinforcement learning, 2020. URL <https://arxiv.org/abs/2007.09515>.
- Joyce Ho and Stephen S. Intille. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI05, pp. 909–918. ACM, April 2005. doi: 10.1145/1054972.1055100. URL <http://dx.doi.org/10.1145/1054972.1055100>.

- Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015. doi: 10.1177/0018720814547570.
- G. Hoffman and C. Breazeal. Cost-based anticipatory action selection for human–robot fluency. *Trans. Rob.*, 23(5): 952–961, October 2007. ISSN 1552-3098. doi: 10.1109/TRO.2007.907483. URL <https://doi.org/10.1109/TRO.2007.907483>.
- Guy Hoffman. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems*, 49(3):209–218, 2019.
- Shengxin Hong, Chang Cai, Sixuan Du, Haiyue Feng, Siyuan Liu, and Xiuyi Fan. “my grade is wrong!”: A contestable ai framework for interactive feedback in evaluating student essays. In *International Conference on Artificial Intelligence in Education*, pp. 27–35. Springer, 2025.
- Wenyi Hong, Weihang Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14281–14290, 2024.
- HONOR. MagicOS 11. <https://www.honor.com/cn/magic-os/>, 2026. Official product page. Accessed October 5, 2026.
- Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.
- Andreas Horni, Kai Nagel, and Kay W. Axhausen (eds.). *The Multi-Agent Transport Simulation MATSim*. Ubiquity Press, London, 2016. doi: 10.5334/baw.
- Benedikt Hornig and Reuth Mirsky. The intelligent disobedience game: Formulating disobedience in stackelberg games and markov decision processes, 2026. URL <https://arxiv.org/abs/2603.20994>.
- Adam O. Horvath and Leslie S. Greenberg. Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology*, 36(2):223–233, 1989. doi: 10.1037/0022-0167.36.2.223.
- Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 159–166. Association for Computing Machinery, 1999. doi: 10.1145/302979.303030. URL <https://doi.org/10.1145/302979.303030>.
- Eric Horvitz and Johnson Apacible. Learning and reasoning about interruption. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, pp. 20–27. ACM, 2003. doi: 10.1145/958432.958440. URL <https://doi.org/10.1145/958432.958440>.
- Eric Horvitz, Paul Koch, and Johnson Apacible. BusyBody: Creating and fielding personalized models of the cost of interruption. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work, CSCW04*, pp. 507–510. ACM, November 2004. doi: 10.1145/1031607.1031690. URL <https://doi.org/10.1145/1031607.1031690>.
- Eric J. Horvitz, John S. Breese, David Heckerman, David Hovel, and Koos Rommelse. The lumiere project: Bayesian user modeling for inferring the goals and needs of software users, 2013a. URL <https://arxiv.org/abs/1301.7385>.
- Eric J. Horvitz, Andy Jacobs, and David Hovel. Attention-sensitive alerting, 2013b. URL <https://arxiv.org/abs/1301.6707>.
- Damian Hostettler, Simon Mayer, Jan Liam Albert, Kay Erik Jenss, and Christian Hildebrand. Real-time adaptive industrial robots: Improving safety and comfort in human-robot collaboration, 2024. URL <https://arxiv.org/abs/2409.09429>.
- Chengkai Hou, Kun Wu, Jiaming Liu, Zhengping Che, Di Wu, Fei Liao, Guangrun Li, Jingyang He, Qiuxuan Feng, Zhao Jin, et al. Robomind 2.0: A multimodal, bimanual mobile manipulation dataset for generalizable embodied intelligence. *arXiv preprint arXiv:2512.24653*, 2025.
- Yiwei Hou, Hao Wang, Muxi Lyu, Marius Momeu, Eric Nguyen, Taige Yang, Koushik Sen, Dawn Song, and David Wagner. Revelio: Cost-efficient agentic memory safety vulnerability detection for repository-scale codebases, 2026. URL <https://arxiv.org/abs/2606.22263>.

- Stephanie Houde, Kristina Brimijoin, Michael Muller, Steven I. Ross, Dario Andres Silva Moran, Gabriel Enrique Gonzalez, Siya Kunde, Morgan A. Foreman, and Justin D. Weisz. Controlling ai agent participation in group conversations: A human-centered approach, 2025. URL <https://arxiv.org/abs/2501.17258>.
- Mo Houtti, Moyan Zhou, Loren Terveen, and Stevie Chancellor. Observe, ask, intervene: Designing AI agents for more inclusive meetings. In Naomi Yamashita, Vanessa Evers, Koji Yatani, Sharon Xianghua Ding, Bongshin Lee, Marshini Chetty, and Phoebe O. Toups Dugas (eds.), *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 709:1–709:18. Association for Computing Machinery, April 2025. doi: 10.1145/3706598.3713838. URL <https://doi.org/10.1145/3706598.3713838>.
- Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966. doi: 10.1109/TSSC.1966.300074. URL <https://doi.org/10.1109/TSSC.1966.300074>.
- Lavinia Hriscu, Valerio Bo, Alberto Sanfeliu, and Anaís Garrell. When should robots intervene? balancing engagement and intrusiveness in human-robot interaction. *arXiv preprint arXiv:2609.21734*, 2026.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Krizan, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What’s the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=kIoBbc76Sy>.
- Bihao Hu, Jiayi Zhu, Yiyang Pei, and Xiaoqing Gu. Exploring the potential of llm to enhance teaching plans through teaching simulation. *NPJ Science of Learning*, 10, 2025a. URL <https://api.semanticscholar.org/CorpusID:276158247>.
- Haoyi Hu, Qirong Lyu, Xianghan Kong, Weiwen Liu, Jianghao Lin, Zixuan Guo, Yan Xu, Yasheng Wang, Weinan Zhang, and Yong Yu. Anticipate and learn: Unleashing idle-time compute in proactive agents. *arXiv preprint arXiv:2605.25971*, 2026a. doi: 10.48550/arXiv.2605.25971. URL <https://arxiv.org/abs/2605.25971>.
- Silan Hu and Xiaoning Wang. Foke: A personalized and explainable education framework integrating foundation models, knowledge graphs, and prompt engineering. In *China national conference on big data and social computing*, pp. 399–411. Springer, 2024.
- Tao Hu, Jiabin Ai, Licheng Wen, Xueheng Li, Shu Zou, Siqi Li, Nianchen Deng, Xinyu Cai, Hongbin Zhou, Pinlong Cai, Daocheng Fu, Yu Yang, Hairong Zhang, Botian Shi, and Xuemeng Yang. IterCAD: An iterative multimodal agent for visually-grounded CAD generation and editing. *arXiv preprint arXiv:2606.13368*, 2026b.
- Tianshuai Hu, Xiaolu Liu, Song Wang, Yiyao Zhu, Ao Liang, Lingdong Kong, Guoyang Zhao, Zeyang Gong, Jun Cen, Zhiyu Huang, Xiaoshuai Hao, Linfeng Li, Hang Song, Xiangtai Li, Jun Ma, Shaojie Shen, Jianke Zhu, Dacheng Tao, Ziwei Liu, and Junwei Liang. Vision-language-action models for autonomous driving: Past, present, and future. *arXiv preprint arXiv:2512.16760*, 2025b.
- Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets, 2008. URL <https://doi.org/10.1109/icdm.2008.22>.
- Yuanda Hu, Wenbin Zuo, Yiting Shen, Tianle Chen, Hector Fabio Calero Tobar, Yate Ge, Xiaohua Sun, and Weiwei Guo. Egoask: Egocentric teaching of personalized object knowledge for household robots. *arXiv preprint arXiv:2609.16766*, 2026c.
- Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *The Fourteenth International Conference on Learning Representations*, 2026d. URL <https://openreview.net/forum?id=DT7JyQC3MR>.
- Yulin Hu, Zimo Long, Jiahe Guo, Xingyu Sui, Xing Fu, Weixiang Zhao, Yanyan Zhao, and Bing Qin. OP-Bench: Benchmarking over-personalization for memory-augmented personalized conversational agents. *arXiv preprint arXiv:2601.13722*, 2026e. doi: 10.48550/arXiv.2601.13722. URL <https://arxiv.org/abs/2601.13722>.
- Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, Zhenrong Cheng, Xuanbo Fan, Jiabin Guo, Xinlei Yu, Zhenhong Zhou, Zewen Hu, Jiahao Huo, Junhao Wang, Yuwei Niu, Yu Wang, Zhenfei Yin, Xiaobin Hu, Yue Liao, Qiankun Li, Kun Wang, Wangchunshu Zhou, Yixin Liu, Dawei Cheng, Qi Zhang, Tao Gui, Shirui Pan, Yan Zhang, Philip Torr, Zhicheng Dou, Ji-Rong

- Wen, Xuanjing Huang, Yu-Gang Jiang, and Shuicheng Yan. Memory in the age of ai agents, 2026f. URL <https://arxiv.org/abs/2512.13564>.
- Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models, 2024. URL <https://arxiv.org/abs/2402.03271>.
- Shengyi Hua, Jianfeng Wu, Tianle Shen, Kangzhe Hu, Zhongzhen Huang, Shujuan Ni, Zhihong Zhang, Yuan Li, Zhe Wang, and Xiaofan Zhang. Pathfound: An agentic multimodal model activating evidence-seeking pathological diagnosis, 2025. URL <https://arxiv.org/abs/2512.23545>.
- Wenyue Hua, Mengting Wan, Shashank Vadrevu, Ryan Nadel, Yongfeng Zhang, and Chi Wang. Interactive speculative planning: Enhance agent efficiency through co-design of system and user interface, 2024. URL <https://arxiv.org/abs/2410.00079>.
- Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh V. Chawla. DeepCrime: Attentive hierarchical recurrent networks for crime prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1423–1432, 2018. doi: 10.1145/3269206.3271793.
- Chien-Ming Huang and Bilge Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, pp. 83–90. IEEE, 2016.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet, 2024. URL <https://arxiv.org/abs/2310.01798>.
- Justin Huang and Maya Cakmak. Supporting mental model accuracy in trigger-action programming. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pp. 215–225, 2015. doi: 10.1145/2750858.2805830.
- Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Ryan Li, Yusuf Roohani, Lin Qiu, Shiyi Cao, Gavin Li, Junze Zhang, Di Yin, Rick Wierenga, Deniz Kavi, Sherry Liu, Tianwei She, Shruti Marwaha, Jennefer N. Carter, Xin Zhou, Matthew T. Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael P. Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Autonomous biomedical research with an artificial intelligence agent. *Science*, 0(0):eadz4351, 2026a. doi: 10.1126/science.adz4351. URL <https://www.science.org/doi/abs/10.1126/science.adz4351>.
- Mianqiu Huang, Taofeng Xue, Chong Peng, Jinrui Ding, Sicheng Fan, Jiale Hong, Yufei Gao, Xiaocheng Zhang, Linsen Guo, Xin Yang, Dengchang Zhao, Yuchen Xie, Peng Pei, Xunliang Xie, and Xipeng Qiu. Evocua-1.5: Online reinforcement learning for multi-turn computer-use agents, 2026b. URL <https://arxiv.org/abs/2607.09773>.
- Philip Huang, Ruixuan Liu, Shobhit Aggarwal, Changliu Liu, and Jiaoyang Li. Apex-mr: Multi-robot asynchronous planning and execution for cooperative assembly. *arXiv preprint arXiv:2503.15836*, 2025a.
- Runxi Huang, Liyu Zhang, Shengzhong Liu, and Xiaomin Ouyang. Mobileexplorer: Accelerating on-device inference for mobile gui agents via online exploration, 2026c. URL <https://arxiv.org/abs/2605.26546>.
- Yue Huang, Zhangchen Xu, Yuchen Ma, Wenjie Wang, Zheyuan Liu, Ziwei Xu, Pin-Yu Chen, Michel Galley, Zinan Lin, Stefan Feuerriegel, Radha Poovendran, Misha Sra, Alex Pentland, Xiangliang Zhang, and Zichen Chen. Reward hacking challenges oversight of autonomous research agents, 2026d. URL <https://arxiv.org/abs/2609.28614>.
- Yuzhou Huang, Benjin Zhu, Hengtong Lu, Victor Shea-Jay Huang, Haiming Zhang, Wei Chen, Jifeng Dai, Yan Xie, and Hongsheng Li. Mindvla-ul: Vla beats va with unified streaming architecture for autonomous driving. *arXiv preprint arXiv:2605.12624*, 2026e.
- Zhongzhan Huang, Shanshan Zhong, Pan Zhou, Shanghua Gao, Marinka Zitnik, and Liang Lin. A causality-aware paradigm for evaluating creativity of multimodal large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.
- Scott Hudson, James Fogarty, Christopher Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny Lee, and Jie Yang. Predicting human interruptibility with sensors: a wizard of oz feasibility study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 257–264, 2003.

- Dana Hughes, Huao Li, Max Chis, Ini Oguntola, Simon Stepputtis, Keyang Zheng, Joseph Campbell, Katia Sycara, and Michael Lewis. A framework for intervention based team support in time critical tasks. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 4805–4812. IEEE, October 2023. doi: 10.1109/smc53992.2023.10393881. URL <http://dx.doi.org/10.1109/smc53992.2023.10393881>.
- John Hughes. How we built Claude Code auto mode: A safer way to skip permissions. <https://www.anthropic.com/engineering/claude-code-auto-mode>, March 2026. Anthropic Engineering.
- Humanoids Daily. Sunday Robotics Raises \$165M to Transition from Demos to Real-World Deployment. <https://www.humanoidsdaily.com/news/sunday-robotics-raises-165m-to-transition-from-demos-to-real-world-deployment>, March 2026. Accessed: 2026-10-05.
- Hunton Andrews Kurth LLP. NHTSA Takes Major Steps in Establishing an Autonomous Vehicle Framework. <https://www.hunton.com/insights/legal/nhtsa-takes-major-steps-in-establishing-an-autonomous-vehicle-framework>, August 2026. Summarizes the July 30–31, 2026 actions: Zoox Part 555 exemption (Fed. Reg. 2026-15485), interim final rule on exemptions (2026-15482), ASCEND consortium, updated AV technical guidance; accessed: 2026-10-05.
- i-SCOOP. ISO 25785-1 explained and what it means for humanoid robot safety. <https://www.i-scoop.eu/iso-25785-1-explained-and-what-it-means-for-humanoid-robot-safety/>, September 2026. Accessed: 2026-10-05.
- Ifueko Igbinedion and Sertac Karaman. Learning when to ask for help: Efficient interactive navigation via implicit uncertainty estimation, 2024. URL <https://arxiv.org/abs/2305.16502>.
- Ramon Iglesias, Federico Rossi, Kevin Wang, David Hallac, Jure Leskovec, and Marco Pavone. Data-driven model predictive control of autonomous mobility-on-demand systems. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6019–6025, 2018. doi: 10.1109/ICRA.2018.8460966.
- Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, et al. $\pi_{0.6}^*$: a vla that learns from experience. *arXiv preprint arXiv:2511.14759*, 2025a.
- Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025b.
- International Federation of Robotics. Five Million Robots now Operate in Factories Globally. <https://ifr.org/ifr-press-releases/five-million-robots-now-operate-in-factories-globally>, September 2026. Accessed: 2026-10-05.
- International Maritime Organization. Autonomous shipping. <https://www.imo.org/en/MediaCentre/HotTopics/Pages/Autonomous-shipping.aspx>, n.d. Accessed 2026-06-12.
- International Organization for Standardization. ISO/CD 25785-1: Robotics — Safety requirements for dynamically stable industrial mobile robots (legged, wheeled, or other forms of locomotion) — Part 1: Robots. <https://www.iso.org/standard/91469.html>, July 2026. Committee Draft, ISO/TC 299. Accessed: 2026-10-05.
- Shamsi T Iqbal and Brian P Bailey. Effects of intelligent notification management on users and their tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 93–102, 2008.
- Shamsi T. Iqbal and Brian P. Bailey. OASIS: A framework for linking notification delivery to the perceptual structure of goal-directed tasks. *ACM Transactions on Computer-Human Interaction*, 17(4):15:1–15:28, December 2010. ISSN 1073-0516. doi: 10.1145/1879831.1879833. URL <https://doi.org/10.1145/1879831.1879833>.
- Shamsi T. Iqbal and Eric Horvitz. Disruption and recovery of computing tasks: Field study, analysis, and directions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 677–686. ACM, April 2007. doi: 10.1145/1240624.1240730. URL <https://doi.org/10.1145/1240624.1240730>.
- Shamsi T. Iqbal, Piotr D. Adamczyk, Xianjun Sam Zheng, and Brian P. Bailey. Towards an index of opportunity: understanding changes in mental workload during task execution. In *Proceedings of the SIGCHI Conference on*

- Human Factors in Computing Systems*, CHI05, pp. 311–320. ACM, April 2005. doi: 10.1145/1054972.1055016. URL <http://dx.doi.org/10.1145/1054972.1055016>.
- Chadni Islam, Muhammad Ali Babar, and Surya Nepal. A multi-vocal review of security orchestration. *ACM Computing Surveys*, 52(2), 2019. doi: 10.1145/3305268. URL <https://doi.org/10.1145/3305268>.
- Vitor Hideyo Isume, Takuya Kiyokawa, Natsuki Yamanobe, Yukiyasu Domae, Weiwei Wan, and Kensuke Harada. Prompt2Craft: Generating functional craft assemblies with LLMs. *arXiv preprint arXiv:2512.04568*, 2025.
- Jacob Jackson, Phillip Kravtsov, and Shomil Jain. Improving Cursor Tab with online RL. Cursor Blog (Research), September 2025. URL <https://cursor.com/en-US/blog/tab-rl>. Accessed: 2026-02-05.
- Jay Jacobs, Sasha Romanosky, Idris Adjerid, and Wade Baker. Improving vulnerability remediation through better exploit prediction. *Journal of Cybersecurity*, 6(1):tyaa015, 2020. doi: 10.1093/cybsec/tyaa015. URL <https://doi.org/10.1093/cybsec/tyaa015>.
- Jay Jacobs, Sasha Romanosky, Benjamin Edwards, Idris Adjerid, and Michael Roytman. Exploit prediction scoring system (EPSS). *Digital Threats: Research and Practice*, 2(3), 2021. doi: 10.1145/3436242. URL <https://doi.org/10.1145/3436242>.
- Léo Jacqmin, Lina M. Rojas-Barahona, and Benoit Favre. "do you follow me?": A survey of recent approaches in dialogue state tracking, 2022. URL <https://arxiv.org/abs/2207.14627>.
- Eunchae Jang and S. Shyam Sundar. Less interaction but more explanation: A communications perspective on agentic AI interfaces. In *Proceedings of the CHI 2026 Workshop on Human-Centered Explainable AI (HCXAI)*. Zenodo, April 2026. doi: 10.5281/zenodo.19700293. URL <https://doi.org/10.5281/zenodo.19700293>.
- Jiyong Jang, Abeer Agrawal, and David Brumley. ReDeBug: Finding unpatched code clones in entire OS distributions. In *2012 IEEE Symposium on Security and Privacy*, pp. 48–62, 2012. doi: 10.1109/SP.2012.13. URL <https://doi.org/10.1109/SP.2012.13>.
- Andrew K.S. Jardine, Daming Lin, and Dragan Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance, 2006. URL <https://doi.org/10.1016/j.ymssp.2005.09.012>.
- Charlene Jennett, Anna L Cox, Paul Cairns, Samira Dhoparee, Andrew Epps, Tim Tijs, and Alison Walton. Measuring and defining the experience of immersion in games. *International journal of human-computer studies*, 66(9):641–661, 2008.
- Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity, 2024. URL <https://arxiv.org/abs/2403.14403>.
- Sijie Ji, Xinzhe Zheng, Wei Gao, and Mani Srivastava. Transforming mental health care with autonomous llm agents at the edge. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, SenSys '25, pp. 692–693. ACM, May 2025. doi: 10.1145/3715014.3724073. URL <http://dx.doi.org/10.1145/3715014.3724073>.
- Zimo Ji, Zekai Zhang, Congying Xu, Zongjie Li, Yudong Gao, Shuai Wang, and Shing-Chi Cheung. Coding agents are guessing: Measuring action-boundary violations in underspecified DevOps instructions, 2026. URL <https://arxiv.org/abs/2607.02294>.
- Hongrui Jia, Jitong Liao, Xi Zhang, Haiyang Xu, Tianbao Xie, Chaoya Jiang, Ming Yan, Si Liu, Wei Ye, and Fei Huang. Osworld-mcp: Benchmarking mcp tool invocation in computer-use agents, 2025. URL <https://arxiv.org/abs/2510.24563>.
- Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1):53–71, 2000.
- Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.acl-long.91.

- Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate feedback loops in recommender systems, 2019. URL <https://arxiv.org/abs/1902.10730>.
- Song Jiang, Da JU, Andrew Cohen, Sasha Mitts, Aaron Foss, Justine T Kao, Xian Li, and Yuandong Tian. Towards full delegation: Designing ideal agentic behaviors for travel planning, 2024b. URL <https://arxiv.org/abs/2411.13904>.
- Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. MedAgentBench: A virtual EHR environment to benchmark medical LLM agents. *NEJM AI*, 2(9):AIdbp2500144, 2025. doi: 10.1056/AIdbp2500144. URL <https://doi.org/10.1056/AIdbp2500144>.
- Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation, 2023. URL <https://arxiv.org/abs/2305.06983>.
- Ziyan Jiang, Li An, Yujian Liu, Jiabao Ji, Qiucheng Wu, Jacob Andreas, Yang Zhang, and Shiyu Chang. VISUALSKILL: Multimodal Skills for Computer-Use Agents, 2026. URL <https://arxiv.org/abs/2606.18448>.
- Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/edac78c3e300629acfe6cbe9ca88fb84-Abstract-Conference.html.
- Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. Mmtom-qa: Multimodal theory of mind question answering, 2024a. URL <https://arxiv.org/abs/2401.08743>.
- Hyoungwook Jin, Seonghee Lee, Hyungyu Shin, and Juho Kim. Teach ai how to code: Using large language models as teachable agents for programming education. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–28, 2024b.
- Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–28, 2025.
- Jiarui Jin, Ming Zhou, Weinan Zhang, Minne Li, Zilong Guo, Zhiwei Qin, Yan Jiao, Xiaocheng Tang, Chenxi Wang, Jun Wang, Guobin Wu, and Jieping Ye. CoRide: Joint order dispatching and fleet management for multi-scale ride-hailing platforms. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1983–1992, 2019. doi: 10.1145/3357384.3357978.
- Ruofan Jin, Mingyang Xu, Fei Meng, Guancheng Wan, Qingran Cai, Yize Jiang, Jin Han, Yuanyuan Chen, Wanqing Lu, Mengyang Wang, Zhiqian Lan, Yuxuan Jiang, Junhong Liu, Dongyao Wang, Le Cong, and Zaixi Zhang. Stella: Towards a biomedical world model with self-evolving multimodal agents. *bioRxiv*, 2026. URL <https://api.semanticscholar.org/CorpusID:279973197>.
- Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: Assessing causal reasoning in language models, 2024c. URL <https://arxiv.org/abs/2312.04350>.
- Huihao Jing, Wenbin Hu, Hongyu Luo, Jianhui Yang, Wei Fan, Haoran Li, and Yangqiu Song. Maslegalbench: Benchmarking multi-agent systems in deductive legal reasoning. *arXiv preprint arXiv:2509.24922*, 2025.
- Zhi Jing, Jinbin Qiao, Ouyang Lu, Jicong Ao, Shuang Qiu, Huazhe Xu, Yu-Gang Jiang, and Chenjia Bai. Assemblm: A spatial reasoning multimodal large language model for robotic assembly, 2026. URL <https://arxiv.org/abs/2604.08983>.
- Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD02*, pp. 133–142. ACM, July 2002. doi: 10.1145/775047.775067. URL <http://dx.doi.org/10.1145/775047.775067>.
- Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and*

- development in information retrieval*, SIGIR05, pp. 154–161. ACM, August 2005. doi: 10.1145/1076034.1076063. URL <http://dx.doi.org/10.1145/1076034.1076063>.
- Arnold Johnson, Kelley Dempsey, Ron Ross, Sarbari Gupta, and Dennis Bailey. Guide for security-focused configuration management of information systems. NIST Special Publication 800-128, National Institute of Standards and Technology, August 2011. URL <https://doi.org/10.6028/NIST.SP.800-128>. Includes updates as of October 2019.
- Benjamin Johnson, Michael W. Floyd, Alexandra Coman, Mark A. Wilson, and David W. Aha. *Goal Reasoning and Trusted Autonomy*, pp. 47–66. Springer International Publishing, 2018. ISBN 9783319648163. doi: 10.1007/978-3-319-64816-3_3. URL http://dx.doi.org/10.1007/978-3-319-64816-3_3.
- Joseph L Jones, Newton E Mack, David M Nugent, and Paul E Sandin. Autonomous floor-cleaning robot, April 26 2005. US Patent 6,883,201.
- Michael I. Jordan and David E. Rumelhart. Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16(3):307–354, July 1992. ISSN 1551-6709. doi: 10.1207/s15516709cog1603_1. URL https://doi.org/10.1207/s15516709cog1603_1.
- Ujjwal K. C. and Jacques Chodorowski. A case study of adding proactivity in indoor social robots using belief–desire–intention (bdi) model. *Biomimetics*, 4(4):74, November 2019. ISSN 2313-7673. doi: 10.3390/biomimetics4040074. URL <http://dx.doi.org/10.3390/biomimetics4040074>.
- David B. Kaber and Mica R. Endsley. The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, 5(2):113–153, March 2004. ISSN 1464-536X. doi: 10.1080/1463922021000054335. URL <http://dx.doi.org/10.1080/1463922021000054335>.
- Ravi Kiran Kadaboina. Jagarin: A three-layer architecture for hibernating personal duty agents on mobile, 2026. URL <https://arxiv.org/abs/2603.05069>.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL <https://arxiv.org/abs/2207.05221>.
- Daniel Kahneman and Amos Tversky. The simulation heuristic. In Daniel Kahneman, Paul Slovic, and Amos Tversky (eds.), *Judgment under Uncertainty: Heuristics and Biases*, pp. 201–208. Cambridge University Press, 1982. doi: 10.1017/CBO9780511809477.015. URL <https://doi.org/10.1017/CBO9780511809477.015>.
- Nidhi Kalra and Susan M. Paddock. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice*, 94:182–193, December 2016. ISSN 0965-8564. doi: 10.1016/j.tra.2016.09.010. URL <http://dx.doi.org/10.1016/j.tra.2016.09.010>.
- Disha Kamale and Dmitry Berenson. Pddl-art: Autonomous symbolic abstraction from demonstration for long-horizon robotic manipulation using vision-language models. *arXiv preprint arXiv:2608.17146*, 2026.
- Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective, 2025. URL <https://arxiv.org/abs/2411.02385>.
- Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pp. 197–206. IEEE, 2018.
- Rahul Kapoor and Miray Kas. Breaking the observability tax: Dynamic resolution anomaly detection via topology-aware active llm agents. *IEEE Access*, 14:43552–43568, 2026. ISSN 2169-3536. doi: 10.1109/access.2026.3675074. URL <http://dx.doi.org/10.1109/access.2026.3675074>.
- Sayash Kapoor, Benedikt Stroebel, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Transactions on Machine Learning Research*, 2025. URL <https://arxiv.org/abs/2407.01502>.

- Priyanshu Karmakar, Soumyabrata Chaudhuri, Shubhojit Mallick, Manish Gupta, Abhik Jana, and Shreya Ghosh. TripTide: A benchmark for adaptive travel planning under disruptions. *arXiv preprint arXiv:2510.21329*, 2025. doi: 10.48550/arXiv.2510.21329. URL <https://arxiv.org/abs/2510.21329>.
- Atoosa Kasirzadeh and Iason Gabriel. Characterizing AI agents for alignment and governance. *arXiv preprint arXiv:2504.21848*, 2025.
- Kirandeep Kaur, Xingda Lyu, and Chirag Shah. Knowing isn't understanding: Re-grounding generative proactivity with epistemic and behavioral insight, 2026. URL <https://arxiv.org/abs/2602.15259>.
- Henry A. Kautz and James F. Allen. Generalized plan recognition. In Tom Kehler (ed.), *Proceedings of the 5th National Conference on Artificial Intelligence. Philadelphia, PA, USA, August 11-15, 1986. Volume 1: Science*, pp. 32–37. Morgan Kaufmann, 1986. URL <http://www.aaai.org/Library/AAAI/1986/aaai86-006.php>.
- Dipesh KC and Anjila Budathoki. Handoff debt: The rediscovery cost when coding agents take over interrupted tasks, 2026. URL <https://arxiv.org/abs/2606.02875>.
- Jack Kelly and William Knottenbelt. Neural NILM: Deep neural networks applied to energy disaggregation. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments (BuildSys)*, pp. 55–64, 2015. doi: 10.1145/2821650.2821672.
- Forest Key. Firefly ai assistant now available in public beta. <https://blog.adobe.com/en/publish/2026/04/27/firefly-ai-assistant-public-beta>, 2026.
- Waleed Bin Khalid and Byung-Cheol Min. Containing behavioral cascades from manipulated claims in llm-powered multi-robot systems. *arXiv preprint arXiv:2609.30523*, 2026.
- Ruhul Amin Khalil, Kashif Ahmad, and Hazrat Ali. Redefining elderly care with agentic ai: Challenges and opportunities. *arXiv preprint arXiv:2507.14912*, 2025.
- Arpandeeep Khatua, Hao Zhu, Peter Tran, Arya Prabhudesai, Frederic Sadrieh, Johann K. Lieberwirth, Xinkai Yu, Yicheng Fu, Michael J. Ryan, Jiaxin Pei, and Diyi Yang. CooperBench: Why coding agents cannot be your teammates yet, 2026. URL <https://arxiv.org/abs/2601.13295>.
- Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.
- Mina Kian, Lydia Ignatova, Jiong Wang, Ji Min Lee, Jiancheng Li, Qianwei Guo, Emily Weiss, Amy O'Connell, Kaitlin Zareno, Jiani Li, et al. A deployable architecture for robot-mediated tasks (DART): Evaluation in socially assistive robot-guided cognitive behavioral therapy exercises. *arXiv preprint arXiv:2609.12349*, 2026.
- Auk Kim, Woohyeok Choi, Jungmi Park, Kyeyoon Kim, and Uichin Lee. Interrupting drivers for interactions: Predicting opportune moments for in-vehicle proactive auditory-verbal tasks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–28, December 2018a. ISSN 2474-9567. doi: 10.1145/3287053. URL <http://dx.doi.org/10.1145/3287053>.
- Beomsoo Kim, Byeongju Kim, Dohyun Kim, Dongwon Kim, Eunchong Kim, Hongmin Kim, Hyeojung Im, Hyeonbin Hwang, Hyeonghwan Kim, Hyoseok Seol, et al. Pubg ally: A conversational embodied agent as an ai teammate. *arXiv preprint arXiv:2609.29837*, 2026a.
- Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions, 2023. URL <https://arxiv.org/abs/2310.15421>.
- Jiho Kim, Junseong Choi, Woosog Chay, Daeun Kyung, Yeonsu Kwon, Yohan Jo, and Edward Choi. ProPerSim: Developing proactive and personalized AI assistants through user-assistant simulation. In *International Conference on Learning Representations*, 2026b. URL <https://arxiv.org/abs/2509.21730>.
- Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. *Textual Explanations for Self-Driving Vehicles*, pp. 577–593. Springer International Publishing, 2018b. ISBN 9783030012168. doi: 10.1007/978-3-030-01216-8_35. URL http://dx.doi.org/10.1007/978-3-030-01216-8_35.

- Joonkyung Kim, Wenxi Chen, Davood Soleymanzadeh, Yi Ding, Xiangbo Gao, Zhengzhong Tu, Ruqi Zhang, Fan Fei, Sushant Veer, Yiwei Lyu, et al. Modular safety guardrails are necessary for foundation-model-enabled robots in the real world. *arXiv preprint arXiv:2602.04056*, 2026c.
- Junhyeok Kim, Min Soo Kim, Jiwan Chung, Jungbin Cho, Jisoo Kim, Sungwoong Kim, Gyeongbo Sim, and Youngjae Yu. Egospeak: Learning when to speak for egocentric conversational agents in the wild, 2025a. URL <https://arxiv.org/abs/2502.14892>.
- Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024a.
- Namyoung Kim, Kai Tzu iunn Ong, Yeonjun Hwang, Minseok Kang, Iiseo Jihn, Gayoung Kim, Minju Kim, and Jinyoung Yeo. Principles: Synthetic strategy memory for proactive dialogue agents, 2025b. URL <https://arxiv.org/abs/2509.17459>.
- Seulbae Kim, Seunghoon Woo, Heejo Lee, and Hakjoo Oh. VUDDY: A scalable approach for vulnerable code clone discovery. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 595–614, 2017. doi: 10.1109/SP.2017.62. URL <https://doi.org/10.1109/SP.2017.62>.
- Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. “I’m Not Sure, But...”: Examining the impact of large language models’ uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 822–835. Association for Computing Machinery, June 2024b. doi: 10.1145/3630106.3658941. URL <https://doi.org/10.1145/3630106.3658941>.
- Taehyung Kim and Jongeun Choi. Collision snapshot guided time-reversed safety-critical scenario generation. *arXiv preprint arXiv:2609.06433*, 2026.
- Taesoo Kim, HyungSeok Han, Soyeon Park, Dae R. Jeong, Dohyeok Kim, Dongkwan Kim, Eunsoo Kim, Jiho Kim, Joshua Wang, Kangsu Kim, Sangwoo Ji, Woosun Song, Hanqing Zhao, Andrew Chin, Gyejin Lee, Kevin Stevens, Mansour Alharthi, Yizhuo Zhai, Cen Zhang, Joonun Jang, Yeongjin Jang, Ammar Askar, Dongju Kim, Fabian Fleischer, Jeongin Cho, Junsik Kim, Kyungjoon Ko, Insu Yun, Sangdon Park, Dowoo Baik, Haein Lee, Hyeon Heo, Minjae Gwon, Minjae Lee, Minwoo Baek, Seunggi Min, Wonyoung Kim, Yonghwi Jin, Younggi Park, Yunjae Choi, Jinho Jung, Gwanhyun Lee, Junyoung Jang, Kyuheon Kim, Yeonghyeon Cha, and Youngjoon Kim. ATLANTIS: AI-driven threat localization, analysis, and triage intelligence system, September 2025c. URL <https://arxiv.org/abs/2509.14589>. arXiv preprint.
- Yejin Kim, Wilbert Pumacay, Omar Rayyan, Max Argus, Winson Han, Eli VanderBilt, Jordi Salvador, Abhay Deshpande, Rose Hendrix, Snehal Jauhri, et al. Molmospaces: A large-scale open ecosystem for robot navigation and manipulation. *arXiv preprint arXiv:2602.11337*, 2026d.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems 37*, 2024c. URL <https://api.semanticscholar.org/CorpusID:269303028>.
- Evan King, Haoxiang Yu, Sangsu Lee, and Christine Julien. Sasha: Creative goal-oriented reasoning in smart homes with large language models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1), March 2024. doi: 10.1145/3643505. URL <https://doi.org/10.1145/3643505>.
- Ross D King, Jem Rowland, Stephen G Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N Soldatova, et al. The automation of science. *Science*, 324(5923):85–89, 2009.
- David Kinny and Michael P. Georgeff. Commitment and effectiveness of situated agents. In John Mylopoulos and Raymond Reiter (eds.), *Proceedings of the 12th International Joint Conference on Artificial Intelligence. Sydney, Australia, August 24-30, 1991*, pp. 82–88. Morgan Kaufmann, 1991. URL <http://ijcai.org/Proceedings/91-1/Papers/014.pdf>.
- Peter Kirgis, Sayash Kapoor, Andrew Schwartz, Stephan Rabanser, David Africa, Konstantinos Voudouris, Viet Nguyen, Toby Pilditch, Magda Dubois, Harry Coppock, Cozmin Ududec, Nitya Nadgir, Matilda Orona, Tilman Bayer, Derrick

- Chan-Sew, Yue Ling, Abhishek Shetty, Helen Toner, Gillian Hadfield, Seth Lazar, Steve Newman, Shoshannah Tekofsky, Rishi Bommasani, and Arvind Narayanan. Can ai agents conduct open-ended ai research? early evidence from two case studies, 2026. URL <https://arxiv.org/abs/2607.27191>.
- Matthew Klenk, Matt Molineaux, and David W. Aha. Goal-driven autonomy for responding to unexpected events in strategy simulations. *Computational Intelligence*, 29(2):187–206, May 2013. ISSN 1467-8640. doi: 10.1111/j.1467-8640.2012.00445.x. URL <https://doi.org/10.1111/j.1467-8640.2012.00445.x>.
- Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, and Kyle Kelley. Jupyter notebooks-a publishing format for reproducible computational workflows. In *Positioning and power in academic publishing: players, agents and agendas: proceedings of the 20th International Conference on Electronic Publishing*, pp. 87. IOS press, 2016.
- Kurt Knutsson. Figure AI says its humanoid robots ran 24 hours straight sorting packages. <https://foxnews.com/tech/humanoid-robots-work-nonstop-package-test>, May 2026. Fox News / CyberGuy Report. Accessed: 2026-10-05.
- Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 881–905, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.50. URL <https://aclanthology.org/2024.acl-long.50/>.
- Dezhi Kong, Zhengzhao Feng, Qiliang Liang, Hao Wang, Haoifei Sun, Changpeng Yang, Yang Li, Peng Zhou, Shuai Nie, Hongzhen Wang, Linfeng Zhou, Hao Jia, Jiaming Xu, Runyu Shi, and Ying Huang. ProactiveMobile: A comprehensive benchmark for boosting proactive intelligence on mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27503–27513, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Kong_ProactiveMobile_A_Comprehensive_Benchmark_for_Boosting_Proactive_Intelligence_On_Mobile_CVPR_2026_paper.html.
- Philip Koopman. Anatomy of a robotaxi crash: Lessons from the cruise pedestrian dragging mishap. In *International Conference on Computer Safety, Reliability, and Security*, pp. 119–133. Springer, 2024.
- Mikko Korkiakoski, Saeid Sheikhi, Jesper Nyman, Jussi Saariniemi, Kalle Tapio, and Panos Kostakos. An empirical evaluation of ai-powered non-player characters’ perceived realism and performance in virtual reality environments. *arXiv preprint arXiv:2507.10469*, 2025.
- Kirsten Korosec. Boston Dynamics’ next-gen humanoid robot will have Google DeepMind DNA. <https://techcrunch.com/2026/01/05/boston-dynamicss-next-gen-humanoid-robot-will-have-google-deepmind-dna/>, January 2026. TechCrunch. Accessed: 2026-10-05.
- Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45), October 2024. ISSN 1091-6490. doi: 10.1073/pnas.2405460121. URL <http://dx.doi.org/10.1073/pnas.2405460121>.
- Alex Kotliarskyi, Victor Zhu, and Zach Brock. An open-source spec for Codex orchestration: Symphony. <https://openai.com/index/open-source-codex-orchestration-symphony/>, April 2026. Published April 27, 2026.
- Alexander Kott and Paul Theron. Doers, not watchers: Intelligent autonomous agents are a path to cyber resilience, 2022. URL <https://arxiv.org/abs/2201.11111>.
- Alexander Kott, Paul Théron, Martin Drašar, Edlira Dushku, Benoît LeBlanc, Paul Losiewicz, Alessandro Guarino, Luigi Mancini, Agostino Panico, Mauno Pihelgas, Krzysztof Rzadca, and Fabio De Gaspari. Autonomous intelligent cyber-defense agent (aica) reference architecture. release 2.0, 2023. URL <https://arxiv.org/abs/1803.10664>.
- E. S. Kox, J. H. Kerstholt, T. F. Hueting, and P. W. de Vries. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems*, 35(2), June 2021. ISSN 1573-7454. doi: 10.1007/s10458-021-09515-9. URL <http://dx.doi.org/10.1007/s10458-021-09515-9>.

- Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. Penalizing side effects using stepwise relative reachability, 2019. URL <https://arxiv.org/abs/1806.01186>.
- Victoria Krakovna, Laurent Orseau, Richard Ngo, Miljan Martic, and Shane Legg. Avoiding side effects by considering future tasks, 2020. URL <https://arxiv.org/abs/2010.07877>.
- KrASIA. UBTech’s full-size humanoid robot revenue jumps 1,445% in H1 2026. <https://kr-asia.com/ubtech-s-full-size-humanoid-robot-revenue-jumps-1445-in-h1-2026>, August 2026. Accessed: 2026-10-05.
- Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. Effects of proactive dialogue strategies on human-computer trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP ’20, pp. 107–116, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368612. doi: 10.1145/3340631.3394840. URL <https://doi.org/10.1145/3340631.3394840>.
- Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. Modelling and predicting trust for developing proactive dialogue strategies in mixed-initiative interaction. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI ’21, pp. 131–140. ACM, October 2021. doi: 10.1145/3462244.3479906. URL <http://dx.doi.org/10.1145/3462244.3479906>.
- Matthias Kraus, Nicolas Wagner, Ron Riekenbrauck, and Wolfgang Minker. Improving proactive dialog agents using socially-aware reinforcement learning, 2023. URL <https://arxiv.org/abs/2211.15359>.
- Matthias Kraus, Sebastian Zepf, Rebecca Westhäußer, Isabel Feustel, Nima Zargham, Ilhan Aslan, Justin Edwards, Sven Mayer, Dimosthenis Kontogiorgos, Nicolas Wagner, et al. Behave ai: Best practices and guidelines for human-centric design and evaluation of proactive ai agents. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 175–178, 2025.
- Phillip Kravtsov. A new tab model. *Cursor Blog/Product*, 2025. URL <https://cursor.com/blog/tab-update>.
- Nir Kshetri. Transforming cybersecurity with agentic ai to combat emerging cyber threats. *Telecommunications Policy*, 49(6):102976, 2025. ISSN 0308-5961. doi: <https://doi.org/10.1016/j.telpol.2025.102976>. URL <https://www.sciencedirect.com/science/article/pii/S0308596125000734>.
- Tyler Kugler. 9 ways AI makes Pixel 10 our most helpful phone yet. Google product announcement, August 2025. URL <https://blog.google/products-and-platforms/devices/pixel/google-pixel-10-ai-features-updates/>. Published August 20, 2025.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Clam: Selective clarification for ambiguous questions with generative language models, 2023a. URL <https://arxiv.org/abs/2212.07769>.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023b. URL <https://arxiv.org/abs/2302.09664>.
- Anagha Kulkarni, Siddharth Srivastava, and Subbarao Kambhampati. Planning for proactive assistance in environments with partial observability, 2021. URL <https://arxiv.org/abs/2105.00525>.
- Thomas Kuntz, Agatha Duzan, Hao Zhao, Francesco Croce, Zico Kolter, Nicolas Flammarion, and Maksym Andriushchenko. Os-harm: A benchmark for measuring safety of computer use agents, 2025. URL <https://arxiv.org/abs/2506.14866>.
- A Gilad Kusne, Heshan Yu, Changming Wu, Huairuo Zhang, Jason Hattrick-Simpers, Brian DeCost, Suchismita Sarker, Corey Oses, Cormac Toher, Stefano Curtarolo, et al. On-the-fly closed-loop materials discovery via bayesian active learning. *Nature communications*, 11(1):5966, 2020.
- Alexander Htet Kyaw, Richa Gupta, Dhruv Shah, Anoop Sinha, Kory Mathewson, Stefanie Pender, Sachin Chitta, Yotto Koga, Faez Ahmed, Lawrence Sass, and Randall Davis. Text to robotic assembly of multi component objects using 3D generative AI and vision language models. *arXiv preprint arXiv:2511.02162*, 2025a.
- Alexander Htet Kyaw, Miana Smith, Se Hwan Jeon, and Neil Gershenfeld. Speech to reality: On-demand production using natural language, 3d generative ai, and discrete robotic assembly. In *Proceedings of the ACM Symposium on Computational Fabrication*, pp. 1–12, 2025b.

- Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality, 2024. URL <https://arxiv.org/abs/2305.00050>.
- Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation, 2025. URL <https://arxiv.org/abs/2505.06120>.
- Paraskevas Lagakis and Stavros Demetriadis. Evaai: a multi-agent framework leveraging large language models for enhanced automated grading. In *International conference on intelligent tutoring systems*, pp. 378–385. Springer, 2024.
- Hanyu Lai, Xiao Liu, Yanxiao Zhao, Han Xu, Hanchen Zhang, Bohao Jing, Yanyu Ren, Shuntian Yao, Yuxiao Dong, and Jie Tang. Computerrl: Scaling end-to-end online reinforcement learning for computer use agents, 2025. URL <https://arxiv.org/abs/2508.14040>.
- Siqi Lai, Zhao Xu, Weijia Zhang, Hao Liu, and Hui Xiong. LLMLight: Large language models as traffic signal control agents. *arXiv preprint arXiv:2312.16044*, 2023.
- John E. Laird, Christian Lebiere, and Paul S. Rosenbloom. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, 38(4):13–26, December 2017. ISSN 2371-9621. doi: 10.1609/aimag.v38i4.2744. URL <http://dx.doi.org/10.1609/aimag.v38i4.2744>.
- LangChain Blog. Introducing ambient agents. <https://blog.langchain.com/introducing-ambient-agents/>, January 2025.
- Thomas Langerak, Sammy Christen, Mert Albaba, Christoph Gebhardt, and Otmar Hilliges. Marlui: Multi-agent reinforcement learning for adaptive uis, 2023. URL <https://arxiv.org/abs/2209.12660>.
- Iliia Larchenko, Gleb Zarin, and Akash Karnatak. Task adaptation of vision-language-action model: 1st place solution for the 2025 behavior challenge. *arXiv preprint arXiv:2512.06951*, 2025.
- Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), November 2018. doi: 10.1145/3274371. URL <https://doi.org/10.1145/3274371>.
- Bettina Laugwitz, Theo Held, and Martin Schrepp. Construction and evaluation of a user experience questionnaire. In Andreas Holzinger (ed.), *HCI and Usability for Education and Work*, volume 5298 of *Lecture Notes in Computer Science*, pp. 63–76, Berlin, Heidelberg, November 2008. Springer. ISBN 978-3-540-89349-3. doi: 10.1007/978-3-540-89350-9_6. URL https://doi.org/10.1007/978-3-540-89350-9_6.
- Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnampati, Andrew D. White, and Samuel G. Rodrigues. Lab-bench: Measuring capabilities of language models for biology research, 2024. URL <https://arxiv.org/abs/2407.10362>.
- Van-Duc Le, Tien-Cuong Bui, and Sang-Kyun Cha. Spatiotemporal deep learning model for citywide air pollution interpolation and prediction. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 55–62, 2020. doi: 10.1109/BigComp48618.2020.00-99.
- Yann LeCun. A path towards autonomous machine intelligence, 2022. URL <https://openreview.net/pdf?id=BZ5a1r-kVsf>. Position paper, version 0.9.2, June 27.
- Geonsun Lee, Min Xia, Nels Numan, Xun Qian, David Li, Yanhe Chen, Ace Kulshrestha, Ishan Chatterjee, Yinda Zhang, Dinesh Manocha, David Kim, and Ruofei Du. Sensible agent: A framework for unobtrusive interaction with proactive ar agent. In *Proceedings of the 39th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 22, 2025a. URL https://durofei.com/papers/Lee_SensibleAgent-AFrameworkForUnobtrusiveInteractionWithProactiveARAgent_UIST2025.pdf.
- Geonsun Lee, Min Xia, Nels Numan, Xun Qian, David Li, Yanhe Chen, Achin Kulshrestha, Ishan Chatterjee, Yinda Zhang, Dinesh Manocha, David Kim, and Ruofei Du. Sensible agent: A framework for unobtrusive interaction with proactive ar agents, 2025b. URL <https://arxiv.org/abs/2509.09255>.

- Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–22. Association for Computing Machinery, 2025c. doi: 10.1145/3706598.3713778.
- John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. doi: 10.1518/hfes.46.1.50_30392. URL https://doi.org/10.1518/hfes.46.1.50_30392.
- Meng-Chen Lee, Costas Panay, Javier Hernandez, Sean Andrist, Dan Bohus, Anatoly Churikov, and Andrew D. Wilson. Respond: Responsive engagement strategy for predictive orchestration and dialogue, 2026. URL <https://arxiv.org/abs/2603.21682>.
- Josh Leeman, Yuhan Liu, Joseph Stiles, Scott B Lee, Prajna Bhatt, Leslie M Schoop, and Robert G Palgrave. Challenges in high-throughput inorganic materials prediction and autonomous synthesis. *PRX Energy*, 3(1):011002, 2024.
- Ido Levy, Ben Wiesel, Sami Marreed, Alon Oved, Avi Yaeli, Nir Mashkif, and Segev Shlomov. ST-WebAgentBench: A benchmark for evaluating safety and trustworthiness in web agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=MUCDzH0ctf>.
- Sebastien Levy, Randolph Yao, Youjiang Wu, Yingnong Dang, Peng Huang, Zheng Mu, Pu Zhao, Tarun Ramani, Naga K. Govindaraju, Xukun Li, Qingwei Lin, Gil Lapid Shafirri, and Murali Chintalapati. Predictive and adaptive failure mitigation to avert production cloud VM interruptions. In *14th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2020, Virtual Event, November 4-6, 2020*, pp. 1155–1170. USENIX Association, 2020. URL <https://www.usenix.org/conference/osdi20/presentation/levy>.
- Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning for negotiation dialogues, 2017. URL <https://arxiv.org/abs/1706.05125>.
- Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models, 2023a. URL <https://arxiv.org/abs/2310.11589>.
- Binxu Li, Tian Yan, Yuanting Pan, Zhe Xu, Jie Luo, Ruiyang Ji, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. Mmedagent: Learning to use medical tools with multi-modal agent. *ArXiv*, abs/2407.02483, 2024a. URL <https://api.semanticscholar.org/CorpusID:270877969>.
- Bo Li, Junwei Ma, Kai Yin, Yiming Xiao, Chia-Wei Hsu, and Ali Mostafavi. Disaster management in the era of agentic ai systems: A vision for collective human-machine intelligence for augmented resilience. *arXiv preprint arXiv:2510.16034*, 2025a.
- Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pp. 80–93. PMLR, 2023b.
- Guangrui Li, Yaochen Xie, Yi Liu, Ziwei Dong, Xingyuan Pan, Tianqi Zheng, Jason Choi, Michael J. Morais, Binit Jha, Shaunak Mishra, Bingrou Zhou, Chen Luo, Monica Xiao Cheng, and Dawn Song. The world won’t stay still: Programmable evolution for agent benchmarks. *arXiv preprint arXiv:2603.05910*, 2026a. URL <https://arxiv.org/abs/2603.05910>.
- Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, et al. Legalagentbench: Evaluating llm agents in legal domain. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2322–2344, 2025b.
- Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. Hello again! llm-powered personalized agent for long-term dialogue, 2025c. URL <https://arxiv.org/abs/2406.05925>.
- Jennica Li, Shirley Zhang, Dakota Sullivan, Bengisu Cagiltay, Heather Kirkorian, Bilge Mutlu, and Kassem Fawaz. "it’s like a pet... but my pet doesn’t collect data about me": Multi-person households’ privacy design preferences for household robots. *arXiv preprint arXiv:2602.16975*, 2026b.
- Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents. *ArXiv*, abs/2405.02957, 2024b. URL <https://api.semanticscholar.org/CorpusID:269605152>.

- Junkai Li, Yunghwei Lai, Tianyi Zhu, Zheng Long Lee, Weizhi Ma, and Yang Liu. TheraAgent: Self-improving therapeutic agent for precise and comprehensive treatment planning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29797–29819, San Diego, California, United States, July 2026c. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1490. URL <https://aclanthology.org/2026.findings-acl.1490/>.
- Lehui Li, Ruixuan Jia, Guo-Ye Yang, and Jia Li. An empirical study of proactive coding assistants in real-world software development. *arXiv preprint arXiv:2605.05700*, 2026d. URL <https://arxiv.org/abs/2605.05700>.
- Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026e.
- Minne Li, Zhiwei Qin, Yan Jiao, Yaodong Yang, Zhichen Gong, Jun Wang, Chenxi Wang, Guobin Wu, and Jieping Ye. Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning. In *The World Wide Web Conference (WWW)*, pp. 983–994, 2019. doi: 10.1145/3308558.3313433.
- Qi Li, Bo Yin, Weiqi Huang, Ruhao Liu, Bojun Zou, Runpeng Yu, Jingwen Ye, Weihao Yu, and Xinchao Wang. Vision-language-action safety: Threats, challenges, evaluations, and mechanisms. *arXiv preprint arXiv:2604.23775*, 2026f.
- Ruiying Li, Yunlang Zhou, YuYao Zhu, Kylin Chen, Jingyuan Wang, Sukai Wang, Kongtao Hu, Minhui Yu, Bowen Jiang, Zhan Su, et al. Roboclaw: An agentic framework for scalable long-horizon robotic tasks. *arXiv preprint arXiv:2603.11558*, 2026g.
- Siyuan Li, Taiyan Lang, Aoqi Yan, Jia Yu, Feifan Liu, Yihan Du, Yu Zheng, Xun Wang, and Peng Liu. Safe task planning with long-term graph memory for embodied agents. *arXiv preprint arXiv:2609.08444*, 2026h. CoRL 2026.
- Wanli Li, Bowen Zhou, Yunyao Yu, Zhou Xu, Yifan Yang, Dongsheng Li, and Caihua Shan. WeaveBench: A long-horizon, real-world benchmark for computer-use agents with hybrid interfaces. *arXiv preprint arXiv:2606.09426*, 2026i. URL <https://arxiv.org/abs/2606.09426>.
- Wenliang K. Li, Rui Yan, Xu Zhang, Li Chen, Hongji Zhu, Jing Zhao, Junjun Li, Mengru Li, Wei Cao, Zihang Jiang, Wei Wei, Kun Zhang, and Shaohua Kevin Zhou. Macd: Multi-agent clinical diagnosis with self-learned knowledge for llm. *ArXiv*, abs/2509.20067, 2025d. URL <https://api.semanticscholar.org/CorpusID:281505225>.
- Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models, 2025e. URL <https://arxiv.org/abs/2501.05366>.
- Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations (ICLR)*, 2018.
- Yanzhou Li, Yiran Zhang, Xiaoyu Zhang, Xiaoxia Liu, and Yang Liu. CODESKILL: Learning self-evolving skills for coding agents, 2026j. URL <https://arxiv.org/abs/2605.25430>.
- Yian Li, Wentao Tian, Yang Jiao, Jingjing Chen, Tianwen Qian, Bin Zhu, Na Zhao, and Yu-Gang Jiang. Look before you decide: Prompting active deduction of mllms for assumptive reasoning, 2025f. URL <https://arxiv.org/abs/2404.12966>.
- Yinuo Li, Jiayuan Wang, Zhouli Yuan, and Haiyu Zhang. Agent technology for agricultural intelligence: Methodological framework and applications. *Electronics*, 2026k. URL <https://api.semanticscholar.org/CorpusID:287331235>.
- Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, Rui Kong, Yile Wang, Hanfei Geng, Jian Luan, Xuefeng Jin, Zilong Ye, Guanqing Xiong, Fan Zhang, Xiang Li, Mengwei Xu, Zhijun Li, Peng Li, Yang Liu, Ya-Qin Zhang, and Yunxin Liu. Personal llm agents: Insights and survey about the capability, efficiency and security, 2024c. URL <https://arxiv.org/abs/2401.05459>.
- Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6342–6353. Association for Computational Linguistics, 2023c. doi: 10.18653/v1/2023.emnlp-main.391.

- Yunfan Li, Bingbing Xu, Xueyun Tian, Xiucheng Xu, and Huawei Shen. Beyond entangled planning: Task-decoupled planning for long-horizon agents, 2026l. URL <https://arxiv.org/abs/2601.07577>.
- Ze Li, Qian Cheng, Ken Hsieh, Yingnong Dang, Peng Huang, Pankaj Singh, Xinsheng Yang, Qingwei Lin, Youjiang Wu, Sebastien Levy, and Murali Chintalapati. Gandalf: An intelligent, end-to-end analytics service for safe deployment in large-scale cloud infrastructure. In Ranjita Bhagwan and George Porter (eds.), *17th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2020, Santa Clara, CA, USA, February 25-27, 2020*, pp. 389–402. USENIX Association, 2020. URL <https://www.usenix.org/conference/nsdi20/presentation/li>.
- Zeyi Li, Yushi Yang, Shawn Xie, Kyle Xu, Tianxing Chen, Yuran Wang, Zhenhao Shen, Yan Shen, Yue Chen, Wenjun Li, et al. Lhome: A simulation environment for deformable object manipulation in household scenarios. *arXiv preprint arXiv:2604.22363*, 2026m.
- Zhonghang Li, Lianhao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 5351–5362, 2024d. doi: 10.1145/3637528.3671578.
- Kaiqu Liang, Julia Kruk, Shengyi Qian, Xianjun Yang, Shengjie Bi, Yuanshun Yao, Shaoliang Nie, Mingyang Zhang, Lijuan Liu, Jaime Fernández Fisac, Shuyan Zhou, and Saghar Hosseini. Learning personalized agents from human feedback. *arXiv preprint arXiv:2602.16173*, 2026. doi: 10.48550/arXiv.2602.16173. URL <https://arxiv.org/abs/2602.16173>.
- Henry Lieberman. Letizia: An agent that assists web browsing. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 924–929, 1995. URL <https://www.ijcai.org/Proceedings/95-1/Papers/119.pdf>.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step, 2023. URL <https://arxiv.org/abs/2305.20050>.
- Gabriel Lima, Assem Zhunis, Lev Manovich, and Meeyoung Cha. On the social-relational moral standing of AI: An empirical study using AI-generated art. *Frontiers in Robotics and AI*, 8:719944, 2021. doi: 10.3389/frobt.2021.719944.
- Kaixiang Lin, Renyu Zhao, Zhe Xu, and Jiayu Zhou. Efficient large-scale fleet management via multi-agent deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 1774–1783, 2018. doi: 10.1145/3219819.3219993.
- Kevin Lin, Charlie Snell, Yu Wang, Charles Packer, Sarah Wooders, Ion Stoica, and Joseph E. Gonzalez. Sleep-time compute: Beyond inference scaling at test-time, 2025. URL <https://arxiv.org/abs/2504.13171>.
- Kunyang Lin, Xutao Wen, Jingxi Lin, Lanyong Lin, Jiaming Liu, Tianshuo Yang, Xianchi Chen, Yue Han, Yiduo Li, Zhanpeng Zhang, and Ping Luo. Egowild2dex: Learning dexterous robotic manipulation from in-the-wild human experience. *arXiv preprint arXiv:2609.23755*, 2026a.
- Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words, 2022. URL <https://arxiv.org/abs/2205.14334>.
- Wilson Lin. Scaling long-running autonomous coding. Cursor Blog, January 2026a. URL <https://cursor.com/blog/scaling-agents>. Accessed: 2026-05-11.
- Wilson Lin. Towards self-driving codebases. <https://cursor.com/blog/self-driving-codebases>, February 2026b. Cursor Research.
- Zhiyu Lin, Upol Ehsan, Rohan Agarwal, Samihan Dani, Vidushi Vashishth, and Mark Riedl. Beyond following: Mixing active initiative into computational creativity, 2024. URL <https://arxiv.org/abs/2409.16291>.
- Zichuan Lin, Feiyu Liu, Yijun Yang, Jiafei Lyu, Yiming Gao, Yicheng Liu, Zhicong Lu, Yangbin Yu, Mingyu Yang, Junyou Li, Deheng Ye, and Jie Jiang. UI-Voyager: A self-evolving GUI agent learning via failed experience, 2026b. URL <https://arxiv.org/abs/2603.24533>.
- Dejian Liu, Ronghuai Huang, Ying Chen, Michael Agyemang Adarkwah, Xiangling Zhang, Xin Li, Junjie Zhang, and Ting Da. Personalized tutoring through conversational agents. In *Using Educational Robots to Enhance Learning: An Analysis of 100 Academic Articles*, pp. 59–85. Springer, 2024a.

- Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, and Percy Liang. Reinforcement learning on web interfaces using workflow-guided exploration. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=ryTp3f-0->.
- Fengkai Liu, Hao Su, Haozhuang Chi, Rui Geng, Congzhi Ren, Xuqing Liu, Yucheng Xu, Yuichi Ohsita, and Liyun Zhang. Event-driven proactive assistive manipulation with grounded vision-language planning. *arXiv preprint arXiv:2603.23950*, 2026a.
- Fengrui Liu, Xiao He, and Tieying Zhang. Help without being asked: A deployed proactive agent system for on-call support with continuous self-improvement, 2026b. URL <https://arxiv.org/abs/2604.09579>.
- Haokun Liu, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, and Chenhao Tan. Hypobench: Towards systematic and principled benchmarking for hypothesis generation, 2026c. URL <https://arxiv.org/abs/2504.11524>.
- Jia Liu and Min Chen. Fagel: Fabric llms agent empowered embodied intelligence evolution with autonomous human-machine collaboration, 2024. URL <https://arxiv.org/abs/2412.20297>.
- Jiachen Liu, Jiaxin Pei, Jintao Huang, Chenglei Si, Ao Qu, Xiangru Tang, Runyu Lu, Lichang Chen, Xiaoyan Bai, Haizhong Zheng, Carl Chen, Zhiyang Chen, Haojie Ye, Yujuan Fu, Zexue He, Zijian Jin, Zhenyu Zhang, Shangquan Sun, Maestro Harmon, John Dianzhuo Wang, Jianqiao Zeng, Jiachen Sun, Mingyuan Wu, Baoyu Zhou, Chenyu You, Shijian Lu, Yiming Qiu, Fan Lai, Yuan Yuan, Yao Li, Junyuan Hong, Ruihao Zhu, Beidi Chen, Alex Pentland, Ang Chen, Mosharaf Chowdhury, and Zechen Zhang. The last human-written paper: Agent-native research artifacts, 2026d. URL <https://arxiv.org/abs/2604.24658>.
- Jiateng Liu, Sha Li, Zhenhailong Wang, Manling Li, and Heng Ji. A language-first approach for procedure planning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1941–1954, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.122. URL <https://aclanthology.org/2023.findings-acl.122/>.
- Jiateng Liu, Lin Ai, Zizhou Liu, Payam Karisani, Zheng Hui, Yi Fung, Preslav Nakov, Julia Hirschberg, and Heng Ji. PropaInsight: Toward deeper understanding of propaganda in terms of techniques, appeals, and intent. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5607–5628, Abu Dhabi, UAE, January 2025a. Association for Computational Linguistics. URL <https://aclanthology.org/2025.coling-main.376/>.
- Jiateng Liu, Bingxuan Li, Zhenhailong Wang, Rushi Wang, Kaiwen Hong, Cheng Qian, Jiayu Liu, Denghui Zhang, Katherine Driggs-Campbell, Manling Li, and Heng Ji. Brick-composer: Using mllms for assembly with diverse bricks, 2026e. URL <https://arxiv.org/abs/2606.05445>.
- Jiateng Liu, Rushi Wang, Bingxuan Li, Kunlun Zhu, Yifan Shen, Qingyun Wang, Ahmed Abbasi, Denghui Zhang, and Heng Ji. Augmenting interface usability heuristics for reliable computer-use agents, 2026f. URL <https://arxiv.org/abs/2605.02729>.
- Jiateng Liu, Rushi Wang, Cheng Qian, Xuejun Zhang, Sun Li, Jiayu Liu, Yifan Shen, Xu Cao, Jiarui Yao, Bingxuan Li, Ruhi Sarikaya, and Heng Ji. Lmbuild: Evaluating llm agents for generating buildable and functional structures, 2026g. URL <https://arxiv.org/abs/2610.04292>.
- Jiateng Liu, Zhenhailong Wang, Rushi Wang, Bingxuan Li, Jeonghwan Kim, Aditi Tiwari, Pengfei Yu, Denghui Zhang, and Heng Ji. OSExpert: Computer-use agents learning professional skills via exploration, 2026h. URL <https://arxiv.org/abs/2603.07978>.
- Jiayu Liu, Cheng Qian, Zhenhailong Wang, Bingxuan Li, Jiateng Liu, Heng Wang, Jeonghwan Kim, Yumeng Wang, Xiushi Chen, Yi R. Fung, and Heng Ji. AdaPlanBench: Evaluating adaptive planning in large language model agents under world and user constraints. *arXiv preprint arXiv:2606.05622*, 2026i. doi: 10.48550/arXiv.2606.05622. URL <https://arxiv.org/abs/2606.05622>.
- Kaige Liu, Yang Li, Lijun Zhu, and Weinan Zhang. Pepa: a persistently autonomous embodied agent with personalities, 2026j. URL <https://arxiv.org/abs/2603.00117>.

- Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-conditioned reinforcement learning: Problems and solutions, 2022a. URL <https://arxiv.org/abs/2201.08299>.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b. doi: 10.1162/tacl_a_00638. URL <https://aclanthology.org/2024.tacl-1.9/>.
- Philip R. Liu, Sparsh Bansal, Jimmy Dinh, Aditya Pawar, Ramani Satishkumar, Shail Desai, Neeraj Gupta, Xin Wang, and Shu Hu. Medchat: A multi-agent framework for multimodal diagnosis with large language models. *2025 IEEE 8th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 456–462, 2025b. URL <https://api.semanticscholar.org/CorpusID:279250322>.
- Qi Liu, Can Li, and Wanjing Ma. Gatsim: Urban mobility simulation with generative agents. *arXiv preprint arXiv:2506.23306*, 2025c.
- Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J Malan. Teaching cs50 with ai: leveraging generative artificial intelligence in computer science education. In *Proceedings of the 55th ACM technical symposium on computer science education V. 1*, pp. 750–756, 2024c.
- Ruixuan Liu, Rui Chen, and Changliu Liu. Task-agnostic adaptation for safe human-robot handover. *IFAC-PapersOnLine*, 55(41):175–180, 2022b.
- Ruixuan Liu, Rui Chen, Abulikemu Abuduweili, and Changliu Liu. Proactive human-robot co-assembly: Leveraging human intention prediction and robust safe control. In *2023 IEEE Conference on Control Technology and Applications (CCTA)*, pp. 339–345. IEEE, 2023b.
- Ruixuan Liu, Philip Huang, Ava Pun, Kangle Deng, Shobhit Aggarwal, Zhenran Tang, Michelle Liu, Deva Ramanan, Jun-Yan Zhu, Jiaoyang Li, et al. Prompt-to-product: Generative assembly via bimanual manipulation. *IEEE Robotics & Automation Magazine*, 2026k.
- Shipeng Liu, Boshen Zhang, and Zhehui Huang. Towards real-time adaptation of embodied agent in human-robot collaboration, 2026l. URL <https://arxiv.org/abs/2412.00435>.
- Shuochen Liu, Junyi Zhu, Long Shu, Junda Lin, Yuhao Chen, Haotian Zhang, Chao Zhang, Derong Xu, Jia Li, Bo Tang, Zhiyu Li, Feiyu Xiong, Enhong Chen, and Tong Xu. PERMA: Benchmarking personalized memory agents via event-driven preference and realistic task environments. *arXiv preprint arXiv:2603.23231*, 2026m. doi: 10.48550/arXiv.2603.23231. URL <https://arxiv.org/abs/2603.23231>.
- Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Tianfan Fu, and Yue Zhao. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. *ArXiv*, abs/2411.15692, 2024d. URL <https://api.semanticscholar.org/CorpusID:274234223>.
- Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 29982–30009, 2025d.
- Tengxiao Liu, Zifeng Wang, Jin Miao, I-Hung Hsu, Jun Yan, Jiefeng Chen, Rujun Han, Fangyuan Xu, Yanfei Chen, Ke Jiang, Samira Daruki, Yi Liang, William Yang Wang, Tomas Pfister, and Chen-Yu Lee. Budget-aware tool-use enables effective agent scaling, 2025e. URL <https://arxiv.org/abs/2511.17006>.
- Tianjian Liu, Hongzheng Zhao, Yuheng Liu, Xingbo Wang, and Zhenhui Peng. Compeer: A generative conversational agent for proactive peer support. *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 2024e. URL <https://api.semanticscholar.org/CorpusID:271432028>.
- Tianjian Liu, Fanqi Wan, Jiajian Guo, and Xiaojun Quan. Proactiveeval: A unified evaluation framework for proactive dialogue agents, 2025f. URL <https://arxiv.org/abs/2508.20973>.
- Wei Liu, Chenxi Wang, Yifei Wang, Zihao Xie, Rennai Qiu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, and Chen Qian. Autonomous agents for collaborative task under information asymmetry, 2024f. URL <https://arxiv.org/abs/2406.14928>.

- Xiaoan Liu, DaeHo Lee, Eric J Gonzalez, Mar Gonzalez-Franco, and Ryo Suzuki. Visionclaw: Always-on ai agents through smart glasses, 2026n. URL <https://arxiv.org/abs/2604.03486>.
- Xiaoze Liu, Ruowang Zhang, Amir H. Abdi, Michel Galley, Zhikai Chen, Siheng Xiong, Xiaoqian Wang, and Jing Gao. Do proactive agents really need an LLM to decide when to wake and what to anchor? *arXiv preprint arXiv:2605.30152*, 2026o. doi: 10.48550/arXiv.2605.30152.
- Xingyu "Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, and Ruofei Du. Visual captions: Augmenting verbal communication with on-the-fly visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023c. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581566. URL <https://doi.org/10.1145/3544548.3581566>.
- Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang Anthony Chen. Proactive conversational agents with inner thoughts, 2025g. URL <https://arxiv.org/abs/2501.00383>.
- Xinyuan Liu, Eren Sadikoglu, Riana Chatterjee, and Ransalu Senanayake. Physical agentic ai: An architecture for orchestrating a robot crew with llms. *arXiv preprint arXiv:2608.22657*, 2026p.
- Yafei Liu, Hongjin Qian, Hengpeng Xu, and Jinmao Wei. Speaker or listener? the role of a dialog agent. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4861–4869. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.437. URL <http://dx.doi.org/10.18653/v1/2020.findings-emnlp.437>.
- Yuhan Liu, Michael J. Q. Zhang, and Eunsol Choi. User feedback in human-llm dialogues: A lens to understand users but noisy as a learning signal, 2025h. URL <https://arxiv.org/abs/2507.23158>.
- Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 1831–1847. USENIX Association, 2024g. URL <https://arxiv.org/abs/2310.12815>.
- Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. A dynamic llm-powered agent network for task-oriented agent collaboration. *arXiv preprint arXiv:2310.02170*, 2023d.
- Ziyi Liu, Bahar Sarrafzadeh, Pei Zhou, Longqi Yang, Jieyu Zhao, and Ashish Sharma. Promediate: A socio-cognitive framework for evaluating proactive agents in multi-party negotiation, 2026q. URL <https://arxiv.org/abs/2510.25224>.
- Shih-Yun Lo and Andrea L. Thomaz. Communication strategy for efficient guidance providing : Domain-structure awareness, performance trade-offs, and value of future observations. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3133–3139. IEEE, May 2021. doi: 10.1109/icra48506.2021.9561880. URL <http://dx.doi.org/10.1109/icra48506.2021.9561880>.
- Ryan Lopopolo. Harness engineering: Leveraging Codex in an agent-first company. <https://openai.com/index/harness-engineering/>, February 2026. OpenAI Engineering Blog. Accessed: 2026-06-22.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024. doi: 10.48550/arXiv.2408.06292. URL <https://arxiv.org/abs/2408.06292>.
- Runyu Lu, Yubo Wu, Ethan Kou, Max Fu, Wenli Xiao, Ajay Mandlekar, Yinzhen Xu, Guanya Shi, Ken Goldberg, Ang Chen, Mosharaf Chowdhury, Yuke Zhu, Linxi Fan, and Guanzhi Wang. Aspire: Agentic /skills discovery for robotics. *arXiv preprint*, 2026a.
- Yaxi Lu, Shenzi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, Weiwen Liu, Yasheng Wang, Zhiyuan Liu, Fangming Liu, and Maosong Sun. Proactive agent: Shifting LLM agents from reactive responses to active assistance. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/75c37811e830bf029584b1c6fac17726-Abstract-Conference.html.

- Yuxuan Lu, Jing Huang, Hui Liu, Jiri Gesi, Yan Han, Shihan Fu, Tianqi Zheng, and Dakuo Wang. Webserv: A browser-server environment for efficient training of reinforcement learning-based web agents at scale, 2025b. URL <https://arxiv.org/abs/2510.16252>.
- Zhengxi Lu, Fei Tang, Guangyi Liu, Kaitao Song, Xu Tan, Jin Ma, Wenqi Zhang, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Ui-copilot: Advancing long-horizon gui automation via tool-integrated policy optimization, 2026b. URL <https://arxiv.org/abs/2604.13822>.
- Marin Lujak, Elizabeth Sklar, and Frédéric Semet. On multi-agent coordination of agri-robot fleets. In *ATT@ECAI*, 2020. URL <https://api.semanticscholar.org/CorpusID:225048086>.
- Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiwen Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, Yicheng Feng, and Zongqing Lu. Being-h0.5: Scaling human-centric robot learning for cross-embodiment generalization. *arXiv preprint arXiv:2601.12993*, 2026a.
- Haohao Luo, Zexi Li, Yuexiang Xie, Wenhao Zhang, Yaliang Li, and Ying Shen. Intentrl: Training proactive user-intent agents for open-ended deep research via reinforcement learning, 2026b. URL <https://arxiv.org/abs/2602.03468>.
- Xiang Luo, Zhiwen Tang, Jin Wang, and Xuejie Zhang. Duetsim: Building user simulator with dual large language models for task-oriented dialogues, 2024a. URL <https://arxiv.org/abs/2405.13028>.
- Xusheng Luo, Shaojun Xu, Ruixuan Liu, and Changliu Liu. Decomposition-based hierarchical task allocation and planning for multi-robots under hierarchical temporal logic specifications. *IEEE Robotics and Automation Letters*, 9(8):7182–7189, 2024b.
- Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J. Ritchie, Soren Mindermann, Evan Hubinger, Ethan Perez, and Kevin Troy. Agentic misalignment: How llms could be insider threats, 2025. URL <https://arxiv.org/abs/2510.05179>.
- Jakub Lála, Odhran O’Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodrigues, and Andrew D. White. Paperqa: Retrieval-augmented generative agent for scientific research, 2023. URL <https://arxiv.org/abs/2312.07559>.
- Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature machine intelligence*, 6(5):525–535, 2024.
- Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. AgentBoard: An analytical evaluation board of multi-turn LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, pp. 74325–74362, 2024a. doi: 10.52202/079017-2365. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/877b40688e330a0e2a3fc24084208dfa-Abstract-Datasets_and_Benchmarks_Track.html.
- Yuan Ma and Jürgen Ziegler. The effect of proactive cues on the use of decision aids in conversational recommender systems. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct ’24, pp. 305–315, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704666. doi: 10.1145/3631700.3665186. URL <https://doi.org/10.1145/3631700.3665186>.
- Zhao Ma, Simon Duenser, Christian Schumacher, Romana Rust, Moritz Bäcker, Fabio Gramazio, Matthias Kohler, and Stelian Coros. Robotsculptor: Artist-directed robotic sculpting of clay. In *Proceedings of the 5th annual ACM symposium on computational fabrication*, pp. 1–12, 2020.
- Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. Language model can listen while speaking, 2024b. URL <https://arxiv.org/abs/2408.02622>.
- Machina Labs. Machina Labs Raises \$124 Million to Scale Manufacturing Infrastructure for Defense and Advanced Mobility. <https://machinalabs.ai/resources/machina-labs-raises-124-million-to-scale-manufacturing-infrastructure-for-defense-and-advanced-mobility>, February 2026a. Accessed: 2026-10-05.
- Machina Labs. Intelligent, software-defined factories for complex metal structures. <https://machinalabs.ai/>, 2026b. Accessed: 2026-07-20.

- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL <https://arxiv.org/abs/2303.17651>.
- David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer, 2018. URL <https://arxiv.org/abs/1711.06664>.
- Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL <https://aclanthology.org/2024.acl-long.747/>.
- Vincent Mai, Philippe Maisonneuve, Tianyu Zhang, Hadi Nekoei, Liam Paull, and Antoine Lesage-Landry. Multi-agent reinforcement learning for fast-timescale demand response of residential loads. *arXiv preprint arXiv:2301.02593*, 2023.
- Jijo Malayil. Google and Mercedes-Benz back humanoid robot Apollo in new global funding drive. <https://interestingengineering.com/ai-robotics/us-startup-raises-funds-to-deploy-apollo>, February 2026. Interesting Engineering. Accessed: 2026-10-05.
- Nathan Malkin, Joe Deatrack, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy attitudes of smart speaker users. *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2019(4):250–271, 2019. doi: 10.2478/popets-2019-0068.
- Marios Evangelos Mamalis, Evangelos Kalampokis, Fotios Fitsilis, Georgios Theodorakopoulos, and Konstantinos Tarabanis. A large language model agent based legal assistant for governance applications. In *International Conference on Electronic Government*, pp. 286–301. Springer, 2024.
- Indrajeet Mandal, Jitendra Soni, Mohd Zaki, Morten M Smedskjaer, Katrin Wondraczek, Lothar Wondraczek, Nitya Nand Gosvami, and NM Anoop Krishnan. Evaluating large language model agents for automation of atomic force microscopy. *Nature Communications*, 16(1):9104, 2025.
- Mehak Maniktala, Christa Cody, Amy Isvik, Nicholas Lytle, Min Chi, and Tiffany Barnes. Extending the hint factory for the assistance dilemma: A novel, data-driven helpneed predictor for proactive problem-solving help. *Journal of Educational Data Mining*, 12(4):24–65, 2020. doi: 10.5281/ZENODO.4399683. URL <https://jedm.educationaldatamining.org/index.php/JEDM/article/view/450>.
- Jiayuan Mao, Tomás Lozano-Pérez, Joshua B Tenenbaum, and Leslie Pack Kaelbling. Learning reusable manipulation strategies. In *Conference on Robot Learning*, pp. 1467–1483. PMLR, 2023.
- Kangkun Mao, Fang Xu, Jinru Ding, Yidong Jiang, Yujun Yao, Yirong Chen, Junming Liu, Xiaoqin Wu, Qian Wu, Xiaoyan Huang, and Jie Xu. Epiplanagent: Agentic automated epidemic response planning. *arXiv preprint arXiv:2512.10313*, 2025.
- Gloria Mark, Daniela Gudith, and Ulrich Klocke. The cost of interrupted work: More speed and stress. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 107–110. Association for Computing Machinery, 2008. doi: 10.1145/1357054.1357072.
- Atsushi Masumori and Takashi Ikegami. Do large language model agents exhibit a survival instinct? an empirical study in a sugarscape-style simulation, 2025. URL <https://arxiv.org/abs/2508.12920>.
- Michael Mateas. An experiment in building a fully-realized interactive drama. In *Game Developers Conference, Game Design Track, 2003*, 2003.
- Jordan K Matelsky, Felipe Parodi, Tony Liu, Richard D Lange, and Konrad P Kording. A large language model-assisted education tool to provide feedback on open-ended responses. *arXiv preprint arXiv:2308.02439*, 2023.
- Matic. *Matic Robot Vacuum and Mop: The Privacy-First Floor Cleaning Robot*. Matic, 2026. URL <https://maticrobots.com/>. Technical specifications and product information for the vision-based, localized AI floor cleaning robot.

- Tara Matthews, Anind K. Dey, Jennifer Mankoff, Scott Carter, and Tye Rattenbury. A toolkit for managing user attention in peripheral displays, 2004. URL <https://doi.org/10.1145/1029632.1029676>.
- Andreas Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3):175–183, 2004. doi: 10.1007/s10676-004-3422-1.
- William P. Maxam, III and James C. Davis. An interview study on Third-Party cyber threat hunting processes in the U.S. department of homeland security. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 2333–2350, Philadelphia, PA, August 2024. USENIX Association. ISBN 978-1-939133-44-1. URL <https://www.usenix.org/conference/usenixsecurity24/presentation/maxam>.
- Daniel C. McFarlane. Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human-Computer Interaction*, 17(1):63–139, March 2002. ISSN 1532-7051. doi: 10.1207/S15327051HCI1701_2. URL https://doi.org/10.1207/S15327051HCI1701_2.
- Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert J. Hendley, and Mirco Musolesi. My phone and me: Understanding people’s receptivity to mobile notifications. In Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan Pablo Hourcade (eds.), *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, pp. 1021–1032. ACM, 2016. doi: 10.1145/2858036.2858566. URL <https://doi.org/10.1145/2858036.2858566>.
- Kai Mei, Xi Zhu, Wujiang Xu, Wenyue Hua, Mingyu Jin, Zelong Li, Shuyuan Xu, Ruosong Ye, Yingqiang Ge, and Yongfeng Zhang. Aios: Llm agent operating system, 2025. URL <https://arxiv.org/abs/2403.16971>.
- Rui Meng, Bhavana Dalvi Mishra, Jiefeng Chen, Chun-Liang Li, Palash Goyal, Mihir Parmar, Yiwen Song, Yale Song, Rajarishi Sinha, Parthasarathy Ranganathan, Burak Gokturk, Jinsung Yoon, and Tomas Pfister. Scientistone: Towards human-level autonomous research via chain-of-evidence, 2026. URL <https://arxiv.org/abs/2605.26340>.
- Meta. Introducing Muse: The world’s first personal AI agent built for everyone. <https://about.fb.com/news/2026/09/introducing-muse-personal-ai-agent/>, September 2026. Published September 8, 2026. Accessed October 5, 2026.
- Xianghang Mi, Feng Qian, Ying Zhang, and XiaoFeng Wang. An empirical characterization of IFTTT: Ecosystem, usage, and performance. In *Proceedings of the 2017 Internet Measurement Conference (IMC)*, pp. 398–404, 2017. doi: 10.1145/3131365.3131369.
- Microsoft. Human-centered AI. <https://www.microsoft.com/en-us/microsoft-copilot/blog/2025/10/23/human-centered-ai/>, October 2025a. Published October 23, 2025.
- Microsoft. Recall overview. Microsoft Learn, 2025b. URL <https://learn.microsoft.com/en-us/windows/apps/develop/windows-integration/recall/>. Accessed: 2026-07-01.
- Tim Miller. Explanation in artificial intelligence: Insights from the social sciences, 2018. URL <https://arxiv.org/abs/1706.07269>.
- Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D Dragan. Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS nexus*, 4(3):pgaf062, 2025.
- Margaret Mitchell, Avijit Ghosh, Alexandra Sasha Luccioni, and Giada Pistilli. Fully autonomous ai agents should not be developed, 2025. URL <https://arxiv.org/abs/2502.02649>.
- Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C. Landsness, Daniel L. Barabasi, Siddharth Narayanan, Nicky Evans, Shriya Reddy, Martha Foiani, Aizad Kamal, Leah P. Shriver, Fang Cao, Asmamaw T. Wassie, Jon M. Laurent, Edwin Melville-Green, Mayk Caldas, Albert Bou, Kaleigh F. Roberts, Sladjana Zagorac, Timothy C. Orr, Miranda E. Orr, Kevin J. Zvezdaryk, Ali E. Ghareeb, Laurie McCoy, Bruna Gomes, Euan A. Ashley, Karen E. Duff, Tonio Buonassisi, Tom Rainforth, Randall J. Bateman, Michael Skarlinski, Samuel G. Rodrigues, Michaela M. Hinks, and Andrew D. White. Kosmos: An ai scientist for autonomous discovery, 2025. URL <https://arxiv.org/abs/2511.02824>.
- Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Munoz, Xinjie Yao, René Zurbrugg, Nikita Rudin, et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025.

- Thomas M. Moerland, Joost Broekens, Aske Plaat, and Catholijn M. Jonker. Model-based reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, 16(1):1–118, January 2023. doi: 10.1561/22000000086. URL <https://doi.org/10.1561/22000000086>.
- Matthew Molineaux, Matthew Klenk, and David W. Aha. Goal-driven autonomy in a navy strategy simulation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1):1548–1554, July 2010. ISSN 2159-5399. doi: 10.1609/aaai.v24i1.7576. URL <https://doi.org/10.1609/aaai.v24i1.7576>.
- Nina Moorman, Julianna Schalkwyk, Vriksha Srihari, Qingyu Xiao, Kamel Alrashedy, Hongseok Jeong, Kiersten Lange, Matthew B. Luebbers, and Matthew Gombolay. Longitudinal robot learning from demonstration with care providers in a home environment. *arXiv preprint arXiv:2608.25196*, 2026.
- Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020. URL <https://proceedings.mlr.press/v119/mozannar20b.html>.
- Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. Reading between the lines: Modeling user behavior and costs in ai-assisted programming, 2024a. URL <https://arxiv.org/abs/2210.14306>.
- Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. When to show a suggestion? integrating human feedback in ai-assisted programming, 2024b. URL <https://arxiv.org/abs/2306.04930>.
- Hussein Mozannar, Gagan Bansal, Cheng Tan, Adam Fourney, Victor Dibia, Jingya Chen, Jack Gerrits, Tyler Payne, Matheus Kunzler Maldaner, Madeleine Grunde-McLaughlin, Eric Zhu, Griffin Bassman, Jacob Alber, Peter Chang, Ricky Loynd, Friederike Niedtner, Ece Kamar, Maya Murad, Rafah Hosn, and Saleema Amershi. Magentic-ui: Towards human-in-the-loop agentic systems, 2025. URL <https://arxiv.org/abs/2507.22358>.
- Muhammad Murad, Muhammad Ahmed, Nizam ul din, Muhammad Farrukh Shahid, Shahbaz Siddiqui, Daniel Byers, Muhammad Hassan Tanveer, and Razvan Cristian Voicu. Agentic ai framework to automate traditional farming for smart agriculture. *AgriEngineering*, 2026. URL <https://api.semanticscholar.org/CorpusID:284507438>.
- Chanti Raju Mylay, Bobin Deng, Zhipeng Cai, and Honghui Xu. Agrisentinel: Privacy-enhanced embedded-llm crop disease alerting system. *ArXiv*, abs/2509.09103, 2025. URL <https://api.semanticscholar.org/CorpusID:281252749>.
- Asen Nachkov, Xi Wang, and Luc Van Gool. Llm agents beyond utility: An open-ended perspective, 2025. URL <https://arxiv.org/abs/2510.14548>.
- Inderjeet Jayakumar Nair, Jiaye Tan, Xiaotian Su, Anne Gere, Xu Wang, and Lu Wang. Closing the loop: Learning to generate writing feedback via language model simulated student revisions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16636–16657, 2024.
- Ryo Nakahashi and Seiji Yamada. Balancing performance and human autonomy with implicit guidance agent, 2021. URL <https://arxiv.org/abs/2109.00414>.
- Jaehyun Nam, Jinsung Yoon, Yanzhou Pan, Yubo Wang, Rui Meng, Parthasarathy Ranganathan, and Tomas Pfister. ScientistTwo: Pioneering the human knowledge frontier with autonomous ai, 2026. URL <https://arxiv.org/abs/2609.19644>.
- Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. Robocasa365: A large-scale simulation framework for training and benchmarking generalist robots. *arXiv preprint arXiv:2603.04356*, 2026. ICLR 2026.
- Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI94, pp. 72–78. ACM, April 1994. doi: 10.1145/191666.191703. URL <http://dx.doi.org/10.1145/191666.191703>.
- Manisha Natarajan, Chunyue Xue, Sanne van Waveren, Karen Feigh, and Matthew Gombolay. Mixed-initiative human-robot teaming under suboptimality with online bayesian adaptation, 2024. URL <https://arxiv.org/abs/2403.16178>.
- Abhijnan Nath, Carine Graff, and Nikhil Krishnaswamy. Collaborate, deliberate, evaluate: How llm alignment affects coordinated multi-agent outcomes, 2026. URL <https://arxiv.org/abs/2509.05882>.

- Deepak Nathani, Cheng Zhang, Chang Huan, Jiaming Shan, Yinfei Yang, Alkesh Patel, Zhe Gan, William Yang Wang, Michael Saxon, and Xin Eric Wang. Proactive agent research environment: Simulating active users to evaluate proactive assistants, 2026. URL <https://arxiv.org/abs/2604.00842>.
- National Highway Traffic Safety Administration. Automated Driving Systems. <https://www.nhtsa.gov/vehicle-manufacturers/automated-driving-systems>, 2026. Includes guidance on operational design domains, object and event detection and response, and fallback conditions; accessed: 2026-07-20.
- National Highway Traffic Safety Administration. Automated Vehicle Safety. <https://www.nhtsa.gov/vehicle-safety/automated-vehicle-safety>, n.d. Accessed: 2026-10-06.
- Grey Nearing, Deborah Cohen, Vusumuzi Dube, Martin Gauch, Oren Gilon, Shaun Harrigan, Avinatan Hassidim, Daniel Klotz, Frederik Kratzert, Asher Metzger, Sella Nevo, Florian Pappenberger, Christel Prudhomme, Guy Shalev, Shlomo Shenzis, Tadele Yednkachw Tekalign, Dana Weitzner, and Yossi Matias. Global prediction of extreme floods in ungauged watersheds. *Nature*, 627:559–563, 2024. doi: 10.1038/s41586-024-07145-1.
- Alex Nelson, Sanjay Rekhi, Murugiah Souppaya, and Karen Scarfone. Incident response recommendations and considerations for cybersecurity risk management: A CSF 2.0 community profile. NIST Special Publication 800-61r3, National Institute of Standards and Technology, April 2025. URL <https://csrc.nist.gov/pubs/sp/800/61/r3/final>.
- Allen Newell and Herbert A Simon. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3):113–126, March 1976. doi: 10.1145/360018.360022. URL <https://doi.org/10.1145/360018.360022>.
- Dang Nguyen, Jian Chen, Yu Wang, Gang Wu, Namyong Park, Zhengmian Hu, Hanjia Lyu, Junda Wu, Ryan Aponte, Yu Xia, Xintong Li, Jing Shi, Hongjie Chen, Viet Dac Lai, Zhouhang Xie, Sungchul Kim, Ruiyi Zhang, Tong Yu, Mehrab Tanjim, Nesreen K. Ahmed, Puneet Mathur, Seunghyun Yoon, Lina Yao, Branislav Kveton, Jihyung Kil, Thien Huu Nguyen, Trung Bui, Tianyi Zhou, Ryan A. Rossi, and Franck Dernoncourt. GUI Agents: A Survey, 2024. URL <https://arxiv.org/abs/2412.13501>.
- Ha-Thanh Nguyen, Wachara Fungwacharakorn, and Ken Satoh. Multi-agent legal verifier systems for data transfer planning. *arXiv preprint arXiv:2511.10925*, 2025.
- Loc X. Nguyen, Avi Deb Raha, Huy Q. Le, Eui-Nam Huh, Dusit Niyato, and Choong Seon Hong. Self-evolving AI for humanoids: Mechanisms, safety, and evaluation of post-deployment self-improvement. *arXiv preprint arXiv:2609.13236*, 2026.
- Jianyuan Ni, Hao Tang, Syed Tousiful Haque, Yan Yan, and Anne H. H. Ngu. A survey on multimodal wearable sensor-based human action recognition. *arXiv preprint arXiv:2404.15349*, 2024.
- Neil Nie, Wenlong Huang, Jiayuan Mao, Li Fei-Fei, Weiyu Liu, and Jiajun Wu. Learning composable skills by discovering spatial and temporal structure with foundation models. *learning*, 43:47, 2026.
- David Nield. Figure ai streamed humanoid robots sorting packages for 8 hours straight — and not everyone is convinced it was fully real. TechRadar, May 2026. URL <https://www.techradar.com/ai-platforms-assistants/figure-ai-streamed-humanoid-robots-sorting-packages-for-8-hours-straight-and-not-everyone-is-convinced-it-was-fully-real>. Accessed: 2026-07-09.
- Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research*, 36(5-7):618–634, February 2017. ISSN 1741-3176. doi: 10.1177/0278364917690593. URL <http://dx.doi.org/10.1177/0278364917690593>.
- Jingjie Ning, Shanshan Zhong, Xiaochuan Li, and Ji Zeng. Scores alone do not prove discovery: The discovery certification protocol for auditing ai research agents, 2026a. URL <https://arxiv.org/abs/2609.09219>.
- Jingjie Ning, Shanshan Zhong, Xiaochuan Li, Ji Zeng, and Chenyan Xiong. One run is not an idea: The implementation lottery in automated research, 2026b. URL <https://arxiv.org/abs/2607.26587>.
- Xuying Ning, Dongqi Fu, Tianxin Wei, Hanqing Zeng, Yuanchen Bei, Bingxuan Li, Zihao Li, Qifan Wang, Xiang Shen, Yifan Wu, et al. Evoharness-rl: Learning self-evolving runtime harness for long-horizon llm agents. *arXiv preprint arXiv:2608.05446*, 2026c.

- Xuying Ning, Katherine Tieu, Dongqi Fu, Tianxin Wei, Zihao Li, Yuanchen Bei, Jiaru Zou, Mengting Ai, Zhining Liu, Ting-Wei Li, Lingjie Chen, Yanjun Zhao, Ke Yang, Bingxuan Li, Cheng Qian, Gaotang Li, Xiao Lin, Zhichen Zeng, Ruizhong Qiu, Sirui Chen, Yifan Sun, Xiyuan Yang, Ruida Wang, Rui Pan, Chenyuan Yang, Dylan Zhang, Liri Fang, Zikun Cui, Yang Cao, Pan Chen, Dorothy Sun, Ren Chen, Mahesh Srinivasan, Nipun Mathur, Yinglong Xia, Hong Li, Hong Yan, Pan Lu, Lingming Zhang, Tong Zhang, Hanghang Tong, and Jingrui He. Code as agent harness. *arXiv preprint arXiv:2605.18747*, May 2026d. URL <https://arxiv.org/abs/2605.18747>.
- Helen Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–157, 2004.
- S.Y. Nof. *Handbook of Industrial Robotics*. Electrical and electronic engineering. Wiley, 1999. ISBN 9780471177838. URL <https://books.google.com/books?id=K271EAAAQBAJ>.
- NPR. Waymo will recall software after its self-driving cars passed stopped school buses. <https://www.npr.org/2025/12/06/nx-s1-5635614/waymo-school-buses-recall>, December 2025. Accessed: 2026-10-05.
- Andrew Nuxoll and John E. Laird. Extending cognitive architecture with episodic memory. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*, pp. 1560–1564. AAAI Press, 2007. URL <http://www.aaai.org/Library/AAAI/2007/aaai07-247.php>.
- NVIDIA. Nvidia ace for games. <https://developer.nvidia.com/ace-for-games>, 2026a.
- NVIDIA. NVIDIA Isaac GR00T N1.7: Open Reasoning VLA Model for Humanoid Robots. <https://huggingface.co/blog/nvidia/gr00t-n1-7>, April 2026b. Accessed: 2026-10-05.
- NVIDIA. Pretrained to Imagine, Fine-Tuned to Act: The Rise of World-Action Models. <https://developer.nvidia.com/blog/pretrained-to-imagine-fine-tuned-to-act-the-rise-of-world-action-models/>, June 2026c. Accessed: 2026-10-05.
- Linus Nwankwo, Bjoern Ellensohn, Christian Rauch, and Elmar Rueckert. Sil: Symbiotic interactive learning for language-conditioned human-agent co-adaptation, 2026. URL <https://arxiv.org/abs/2511.05203>.
- Janghoon Ock, Radheesh Sharma Meda, Srivathsan Badrinarayanan, Neha S. Aluru, Achuth Chandrasekhar, and Amir Barati Farimani. Large language model agent for modular task execution in drug discovery, 2025. URL <https://arxiv.org/abs/2507.02925>.
- Martin Odersky, Yaoyu Zhao, Yichen Xu, Oliver Bračevac, and Cao Nguyen Pham. Securing agents with tracked capabilities. In *Proceedings of the ACM Conference on AI and Agent Systems, CAIS '26*, pp. 812–838. Association for Computing Machinery, May 2026. doi: 10.1145/3786335.3813127. URL <https://www.yichenxu.me/files/publications/securing-agents/paper.pdf>.
- Dimitri Ognibene, Sabrina Patania, Luca Annesse, Cansu Koyuturk, Franca Garzotto, Giuseppe Vizzari, Azzurra Ruggeri, and Simone Colombani. Scoop: A framework for proactive collaboration and social continual learning through natural language interaction and causal reasoning, 2025. URL <https://arxiv.org/abs/2503.10241>.
- Jeesun Oh, Wooseok Kim, Sungbae Kim, Hyeonjeong Im, and Sangsu Lee. Better to ask than assume: Proactive voice assistants’ communication strategies that respect user agency in a smart home environment. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI)*, 2024. doi: 10.1145/3613904.3642193.
- Juhyun Oh, Eunsu Kim, and Alice Oh. Flex-TravelPlanner: A benchmark for flexible planning with language agents. In *Workshop on Reasoning and Planning for Large Language Models at ICLR*, 2025. URL <https://openreview.net/forum?id=a7unQ5jMx7>.
- Kazuo Okamura and Seiji Yamada. Adaptive trust calibration for human-ai collaboration. *PLOS ONE*, 15(2):e0229132, February 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0229132. URL <http://dx.doi.org/10.1371/journal.pone.0229132>.
- Stephen M. Omohundro. The basic AI drives. In Pei Wang, Ben Goertzel, and Stan Franklin (eds.), *Artificial General Intelligence 2008, Proceedings of the First AGI Conference, AGI 2008, March 1-3, 2008, University of Memphis, Memphis, TN, USA*, volume 171 of *Frontiers in Artificial Intelligence and Applications*, pp. 483–492. IOS Press, 2008. URL <http://www.booksonline.iospress.nl/Content/View.aspx?piid=8341>.

- OpenAI. Video generation models as world simulators. Technical report, OpenAI, February 2024. URL <https://openai.com/index/video-generation-models-as-world-simulators/>.
- OpenAI. Introducing Codex. <https://openai.com/index/introducing-codex/>, May 2025a. Accessed: 2026-06-22.
- OpenAI. Introducing ChatGPT pulse. <https://openai.com/index/introducing-chatgpt-pulse/>, 2025b.
- OpenAI. Introducing the Codex app. <https://openai.com/index/introducing-the-codex-app/>, February 2026a.
- OpenAI. Running Codex safely at OpenAI. <https://openai.com/index/running-codex-safely/>, May 2026b.
- OpenAI. Introducing dots. <https://openai.com/index/introducing-dots/>, September 2026c. Published September 29, 2026.
- OpenAI. On the Navier–Stokes millennium prize problem, September 2026d. URL <https://openai.com/index/navier-stokes-solution/>. Published September 8, 2026. Accessed: 2026-10-03.
- OpenAI. Ten advances in mathematics and theoretical computer science, August 2026e. URL <https://openai.com/index/ten-advances-in-mathematics/>. Published August 1, 2026. Accessed: 2026-10-03.
- OpenAI. Codex documentation. <https://developers.openai.com/codex>, n.d. Accessed: 2026-06-14.
- Christiane Oppermann, Karola Pitsch, Ramin Yaghoubzadeh, and Stefan Kopp. The communicative activity of “making suggestions” as an interactional process: Towards a dialog model for hai. In *Proceedings of the 5th International Conference on Human Agent Interaction*, HAI ’17, pp. 161–170. ACM, October 2017. doi: 10.1145/3125739.3125752. URL <http://dx.doi.org/10.1145/3125739.3125752>.
- Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016. doi: 10.3390/s16010115.
- Yuki Orimo, Iori Kurata, Hodaka Mori, Ryuhei Okuno, Ryohto Sawada, and Daisuke Okanohara. Parc: An autonomous self-reflective coding agent for robust execution of long-horizon tasks, 2025. URL <https://arxiv.org/abs/2512.03549>.
- Shochi Otogi, Hung-Hsuan Huang, Ryo Hotta, and Kyoji Kawagoe. Finding the timings for a guide agent to intervene in user conversation in considering their gaze behaviors. In *Proceedings of the 6th workshop on Eye gaze in intelligent human machine interaction: gaze in multimodal interaction*, ICMI ’13, pp. 19–24. ACM, December 2013. doi: 10.1145/2535948.2535957. URL <http://dx.doi.org/10.1145/2535948.2535957>.
- Pierre-Yves Oudeyer, Frédéric Kaplan, and Verena V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286, April 2007. ISSN 1089-778X. doi: 10.1109/TEVC.2006.890271. URL <https://doi.org/10.1109/TEVC.2006.890271>.
- Siru Ouyang, Jun Yan, Yanfei Chen, Rujun Han, Zifeng Wang, Bhavana Dalvi Mishra, Rui Meng, Chun-Liang Li, Yizhu Jiao, Kaiwen Zha, Maohao Shen, Vishy Tirumalashetty, George Lee, Jiawei Han, Tomas Pfister, and Chen-Yu Lee. SkillIOS: Learning skill curation for self-evolving agents, 2026. URL <https://arxiv.org/abs/2605.06614>.
- William Overman and Mohsen Bayati. The oversight game: Learning to cooperatively balance an ai agent’s safety and autonomy, 2026. URL <https://arxiv.org/abs/2510.26752>.
- Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. doi: 10.48550/arXiv.2310.08560. URL <https://arxiv.org/abs/2310.08560>.
- E. S. Page. Continuous inspection schemes, 1954. URL <https://doi.org/10.2307/2333009>.
- Matthew Paisner, Michael Cox, Michael Maynard, and Don Perlis. Goal-driven autonomy for cognitive systems. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36(36):2085–2090, 2014. URL <https://escholarship.org/uc/item/5vq1h9jc>.
- Peter Pak, Victor Alvarado, and Amir Barati Farimani. AI agentic selective laser sintering process optimization. *arXiv preprint arXiv:2608.25928*, 2026.

- Yiyuan Pan, Xusheng Luo, Hanjiang Hu, Peiqi Yu, and Changliu Liu. Emergent neural automaton policies: Learning symbolic structure from visuomotor trajectories. *arXiv preprint arXiv:2603.25903*, 2026.
- Athanasios Aris Panagopoulos, Muddasser Alam, Alex Rogers, and Nicholas R. Jennings. Adaheat: A general adaptive intelligent agent for domestic heating control. In Gerhard Weiss, Pinar Yolum, Rafael H. Bordini, and Edith Elkind (eds.), *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2015, Istanbul, Turkey, May 4-8, 2015*, pp. 1295–1303. ACM, 2015. URL <http://dl.acm.org/citation.cfm?id=2773317>.
- Dimitris Panagopoulos, Giannis Petousakis, Aniketh Ramesh, Tianshu Ruan, Grigoris Nikolaou, Rustam Stolkin, and Manolis Chiou. A hierarchical variable autonomy mixed-initiative framework for human-robot teaming in mobile robotics, 2022. URL <https://arxiv.org/abs/2211.14095>.
- Raja Parasuraman and Dietrich H. Manzey. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3):381–410, June 2010. ISSN 1547-8181. doi: 10.1177/0018720810376055. URL <http://dx.doi.org/10.1177/0018720810376055>.
- Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2): 230–253, June 1997. ISSN 1547-8181. doi: 10.1518/001872097778543886. URL <http://dx.doi.org/10.1518/001872097778543886>.
- Raja Parasuraman, Robert Molloy, and Indramani L. Singh. Performance consequences of automation-induced “complacency”. *The International Journal of Aviation Psychology*, 3(1):1–23, 1993. doi: 10.1207/S15327108IJAP0301_1.
- Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3):286–297, May 2000. ISSN 1083-4427. doi: 10.1109/3468.844354. URL <https://doi.org/10.1109/3468.844354>.
- Pere Pardo and Christian Strasser. The goal after tomorrow: Offline goal reasoning with norms. *Journal of Artificial Intelligence Research*, 80:1703–1759, August 2024. ISSN 1076-9757. doi: 10.1613/jair.1.15566. URL <http://dx.doi.org/10.1613/jair.1.15566>.
- Soffa Pardo-Pina, Jörn Germer, R. Suay-Cortes, Manuel Platero-Horcajadas, and Francisco-Javier Ferrández-Pastor. An agent-based service architecture for smart greenhouses: Telemetry analytics and decision support with rag-grounded llm agents. *Smart Agricultural Technology*, 2026. URL <https://api.semanticscholar.org/CorpusID:285991943>.
- Zachary A Pardos, Matthew Tang, Ioannis Anastasopoulos, Shreya K Sheel, and Ethan Zhang. Oatutor: An open-source adaptive tutoring system and curated content library for learning sciences research. In *Proceedings of the 2023 chi conference on human factors in computing systems*, pp. 1–17, 2023.
- Eli Pariser. *The filter bubble: What the Internet is hiding from you*. penguin UK, 2011.
- Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Social simulacra: Creating populated prototypes for social computing systems, 2022. URL <https://arxiv.org/abs/2208.04024>.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 2:1–2:22. ACM, 2023. doi: 10.1145/3586183.3606763. URL <https://doi.org/10.1145/3586183.3606763>.
- Keonwoo Park and Ilkyeong Moon. Multi-agent deep reinforcement learning approach for ev charging scheduling in a smart grid. *Applied Energy*, 328:120111, 2022. ISSN 0306-2619. doi: <https://doi.org/10.1016/j.apenergy.2022.120111>. URL <https://www.sciencedirect.com/science/article/pii/S030626192201368X>.
- Minju Park, Sojung Kim, Seunghyun Lee, Soonwoo Kwon, and Kyuseok Kim. Empowering personalized learning through a conversation-based tutoring system with student modeling. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–10, 2024.

- Ivan Pashchenko, Henrik Plate, Serena Elisa Ponta, Antonino Sabetta, and Fabio Massacci. Vulnerable open source dependencies: Counting those that matter. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM '18*. Association for Computing Machinery, 2018. doi: 10.1145/3239235.3268920. URL <https://arxiv.org/abs/1808.09753>.
- Gil Pasternak, Dheeraj Rajagopal, Julia White, Dhruv Atreja, Matthew Thomas, George Hurn-Maloney, and Ash Lewis. Beyond reactivity: Measuring proactive problem solving in llm agents, 2026. URL <https://arxiv.org/abs/2510.19771>.
- Deep Anil Patel, Iain Melvin, Christopher Malon, and Martin Renqiang Min. Discussllm: Teaching large language models when to speak, 2026. URL <https://arxiv.org/abs/2508.18167>.
- Maithili Patel and Sonia Chernova. Proactive robot assistance via spatio-temporal object modeling. *arXiv preprint arXiv:2211.15501*, 2022. Conference on Robot Learning (CoRL).
- Maithili Patel and Sonia Chernova. Robots that take initiative: A framework for building and evaluating proactive robots. *arXiv preprint arXiv:2609.28910*, 2026.
- Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2778–2787. PMLR, 2017. URL <https://proceedings.mlr.press/v70/pathak17a.html>.
- Eshika Pathak and Leela Krishna. When should a failing robot ask? initiating corrective human-robot dialogue from audited sensor evidence. *arXiv preprint arXiv:2609.21942*, 2026.
- Shishir G. Patil, Tianjun Zhang, Vivian Fang, Noppapon C., Roy Huang, Aaron Hao, Martin Casado, Joseph E. Gonzalez, Raluca Ada Popa, and Ion Stoica. Goex: Perspectives and designs towards a runtime for autonomous llm applications, 2024. URL <https://arxiv.org/abs/2404.06921>.
- Norman W. Paton and Oscar Díaz. Active database systems. *ACM Computing Surveys*, 31(1):63–103, March 1999. ISSN 1557-7341. doi: 10.1145/311531.311623. URL <http://dx.doi.org/10.1145/311531.311623>.
- Hammond Pearce, Benjamin Tan, Baleegh Ahmad, Ramesh Karri, and Brendan Dolan-Gavitt. Examining zero-shot vulnerability repair with large language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 2339–2356, 2023. doi: 10.1109/SP46215.2023.10179420. URL <https://doi.org/10.1109/SP46215.2023.10179420>.
- Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, 2 edition, 2009. doi: 10.1017/CBO9780511803161.
- Judea Pearl. *The book of why: The new science of cause and effect*. Basic Books, 2018.
- Veljko Pejovic and Mirco Musolesi. InterruptMe: Designing intelligent prompting mechanisms for pervasive applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 897–908. Association for Computing Machinery, 2014. doi: 10.1145/2632048.2632062. URL <https://doi.org/10.1145/2632048.2632062>.
- Jacopo Pellegrino, Marco Maggiora, and Walter Allasia. A multi-agent approach for autonomous digital preservation. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6. IEEE, June 2015. doi: 10.1109/icmew.2015.7169866. URL <http://dx.doi.org/10.1109/icmew.2015.7169866>.
- Qiyao Peng, Hongtao Liu, Hua Huang, Qing Yang, and Minglai Shao. A survey on llm-powered agents for recommender systems. *arXiv preprint arXiv:2502.10050*, 2025.
- Zhenhui Peng, Yunhwan Kwon, Jiaan Lu, Ziming Wu, and Xiaojuan Ma. Design and evaluation of service robot’s proactivity in decision-making support process. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, pp. 1–13, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300328. URL <https://doi.org/10.1145/3290605.3300328>.
- Juan C. Perdomo, Tijana Zrnica, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction, 2021. URL <https://arxiv.org/abs/2002.06673>.

- Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models, 2022. URL <https://arxiv.org/abs/2211.09527>. NeurIPS 2022 ML Safety Workshop.
- Physical Intelligence, Sergey Levine, Chelsea Finn, Karl Pertsch, et al. $\pi_{0.7}$: a steerable generalist robotic foundation model with emergent capabilities. *arXiv preprint arXiv:2604.15483*, 2026.
- Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025. doi: 10.48550/arXiv.2502.08691. URL <https://arxiv.org/abs/2502.08691>.
- Martin Pielot, Karen Church, and Rodrigo de Oliveira. An in-situ study of mobile phone notifications. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, MobileHCI '14, pp. 233–242. ACM, September 2014. doi: 10.1145/2628363.2628364. URL <http://dx.doi.org/10.1145/2628363.2628364>.
- Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. Beyond interruptibility: Predicting opportune moments to engage mobile phone users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–25, 2017. doi: 10.1145/3130956.
- PLAUD. PLAUD NotePin: Your wearable ai note taker and memory capsule. Official product page, 2026. URL <https://www.plaud.ai/products/plaud-notepin>. Accessed: 2026-07-01.
- Pony AI Inc. Pony AI Inc. Among the First to Receive Permit for Fully Driverless Commercial Robotaxi Services in Shanghai’s Pudong New Area. <https://ir.pony.ai/news-releases/news-release-details/pony-ai-inc-among-first-receive-permit-fully-driverless>, July 2025a. Accessed: 2026-07-20.
- Pony AI Inc. PONY AI Inc. Realized Gen-7 Robotaxi city-wide UE Breakeven; Set to Surpass 2025 Fleet Target and Expand to 3,000+ Vehicles by End of Next Year. <https://www.globenewswire.com/news-release/2025/11/25/3194073/0/en/PONY-AI-Inc-Realized-Gen-7-Robotaxi-city-wide-UE-Breakeven-Set-to-Surpass-2025-Fleet-Target-and-Expand-to-3-000-Vehicles-by-End-of-Next-Year.html>, November 2025b. Accessed: 2026-10-05.
- Raluca Ada Popa and Four Flynn. Introducing CodeMender: An AI agent for code security. Google DeepMind research blog, October 2025. URL <https://deepmind.google/blog/introducing-codemender-an-ai-agent-for-code-security/>. Published 2025-10-06; developer report; accessed 2026-09-13.
- Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. Adapt: As-needed decomposition and planning with language models, 2024. URL <https://arxiv.org/abs/2311.05772>.
- David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4): 515–526, December 1978. ISSN 1469-1825. doi: 10.1017/s0140525x00076512. URL <http://dx.doi.org/10.1017/s0140525x00076512>.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models, 2023. URL <https://arxiv.org/abs/2210.03350>.
- Loizos Psarakis, Dimitris Nathanael, and Nicolas Marmaras. Communicating robots’ intent through visual cues enhances human anticipatory behavior in human–dual robot collaboration. *Robotics and Computer-Integrated Manufacturing*, 92:102886, 2025.
- Kevin Pu, Daniel Lazaro, Ian Arawjo, Haijun Xia, Ziang Xiao, Tovi Grossman, and Yan Chen. Assistance or disruption? exploring and evaluating the design and trade-offs of proactive AI programming support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pp. 152:1–152:21. ACM, April 2025a. doi: 10.1145/3706598.3713357. URL <https://doi.org/10.1145/3706598.3713357>.
- Kevin Pu, Ting Zhang, Naveen Sendhilnathan, Sebastian Freitag, Raj Sodhi, and Tanya R. Jonker. ProMemAssist: Exploring timely proactive assistance through working memory modeling in multi-modal wearable devices. In Andrea Bianchi, Elena L. Glassman, Wendy E. Mackay, Shengdong Zhao, Jeeun Kim, and Ian Oakley (eds.), *Proceedings of*

- the 38th Annual ACM Symposium on User Interface Software and Technology*, pp. 56:1–56:19. ACM, September 2025b. doi: 10.1145/3746059.3747770. URL <https://doi.org/10.1145/3746059.3747770>.
- Ava Pun, Kangle Deng, Ruixuan Liu, Deva Ramanan, Changliu Liu, and Jun-Yan Zhu. Generating physically stable and buildable brick structures from text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14798–14809, 2025.
- Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y Zhu, et al. Egoverse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026.
- Ethan Qi and Shaochen Wang. Global Humanoid Robot Shipments Soar Nearly 300% YoY in H1 2026, Driven by Commercial Deployments. <https://counterpointresearch.com/en/insights/global-humanoid-robot-shipments-soar-nearly-300-percent-yoy-in-h1-2026>, August 2026. Counterpoint Research. Accessed: 2026-10-05.
- Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, Jie Tang, and Yuxiao Dong. Webrl: Training llm web agents via self-evolving online curriculum reinforcement learning. In *International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=oVKEAFjEqv>.
- Cheng Qian, Bingxiang He, Zhong Zhuang, Jia Deng, Yujia Qin, Xin Cong, Zhong Zhang, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Tell me more! towards implicit user intention understanding of language model driven agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1088–1113, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.61. URL <https://aclanthology.org/2024.acl-long.61/>.
- Cheng Qian, Zuxin Liu, Akshara Prabhakar, Zhiwei Liu, Jianguo Zhang, Haolin Chen, Heng Ji, Weiran Yao, Shelby Heinecke, Silvio Savarese, Caiming Xiong, and Huan Wang. UserBench: An interactive gym environment for user-centric agents. *arXiv preprint arXiv:2507.22034*, 2025a. doi: 10.48550/arXiv.2507.22034. URL <https://arxiv.org/abs/2507.22034>. Presented at the NeurIPS 2025 Workshop on Scaling Environments for Agents.
- Yaoyao Qian, Jindan Huang, Yuanli Wang, Simon Yu, Kyrie Zhixuan Zhou, Jiayuan Mao, Mingfu Liang, and Hanhan Zhou. When to act, when to wait: Modeling the intent-action alignment problem in dialogue, 2025b. URL <https://arxiv.org/abs/2506.01881>.
- Shuofei Qiao, Runnan Fang, Zhisong Qiu, Xiaobin Wang, Ningyu Zhang, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Benchmarking agentic workflow generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=vunPX0Fmoi>.
- Dantong Qin, Yike Guo, Qinlin Liu, Alessandro Bozzon, and Pan Wang. Cobrush: A hierarchical planning framework for human-robot co-painting. *arXiv preprint arXiv:2609.35476*, 2026.
- Lufu Qin, Xingcai Wu, Xinyu Dong, Huan Wang, Tingwei Yang, and Qi Wang. Pdd-agent: Multimodal large language model-driven ai agent for enhanced plant disease diagnosis. *2025 IEEE International Conference on Image Processing (ICIP)*, 2025a. URL <https://api.semanticscholar.org/CorpusID:280732396>.
- Tian Qin, Felix Bai, Ting-Yao Hu, Raviteja Vemulapalli, Hema Swetha Koppula, Zhiyang Xu, Bowen Jin, Mert Cemri, Jiarui Lu, Zirui Wang, and Meng Cao. COMPASS: A multi-turn benchmark for tool-mediated planning and preference optimization. *arXiv preprint arXiv:2510.07043*, 2025b. doi: 10.48550/arXiv.2510.07043. URL <https://arxiv.org/abs/2510.07043>.
- Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, Xiaojun Xiao, Kai Cai, Chuang Li, Yaowei Zheng, Chaolin Jin, Chen Li, Xiao Zhou, Minchao Wang, Haoli Chen, Zhaojian Li, Haihua Yang, Haifeng Liu, Feng Lin, Tao Peng, Xin Liu, and Guang Shi. Ui-tars: Pioneering automated gui interaction with native agents, 2025c. URL <https://arxiv.org/abs/2501.12326>.
- Yansheng Qiu, Haoquan Zhang, Zhaopan Xu, Ming Li, Diping Song, Zheng Wang, and Kaipeng Zhang. Ai idea bench 2025: Ai research idea generation benchmark, 2025. URL <https://arxiv.org/abs/2504.14191>.

- Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling, 2024. URL <https://arxiv.org/abs/2306.10193>.
- Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind, 2018. URL <https://arxiv.org/abs/1802.07740>.
- Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex ‘Sandy’ Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum, and Michael Wellman. Machine behaviour. *Nature*, 568(7753):477–486, Apr 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1138-y. URL <https://doi.org/10.1038/s41586-019-1138-y>.
- Prithvi Rajasekaran. Harness design for long-running application development. <https://www.anthropic.com/engineering/harness-design-long-running-apps>, March 2026. Anthropic Engineering.
- Prithvi Rajasekaran, Ethan Dixon, Carly Ryan, and Jeremy Hadfield. Effective context engineering for AI agents. <https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents>, September 2025. Anthropic Engineering.
- Dongchuan Ran, Linyu Ou, Xueheng Li, Wenwen Tong, Chenxu Guo, Hewei Guo, Kaibin Wang, and Lewei Lu. EgoPro-Bench: Benchmarking personalized proactive interaction in egocentric video streams, 2026. URL <https://arxiv.org/abs/2605.07299>.
- Anand S. Rao and Michael P. Georgeff. BDI agents: From theory to practice. In Victor R. Lesser and Les Gasser (eds.), *Proceedings of the First International Conference on Multiagent Systems*, pp. 312–319. MIT Press, June 1995. URL <https://cdn.aaai.org/ICMAS/1995/ICMAS95-042.pdf>.
- Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information, 2018. URL <https://arxiv.org/abs/1805.04655>.
- Petros Raptopoulos, Giorgos Filandrianos, Maria Lymperaiou, and Giorgos Stamou. PAKTON: A multi-agent framework for question answering in long legal agreements. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7948–7984, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.403. URL <https://aclanthology.org/2025.emnlp-main.403/>.
- Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory, 2025. URL <https://arxiv.org/abs/2501.13956>.
- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696, 2020. doi: 10.1609/aaai.v34i05.6394. URL <https://ojs.aaai.org/index.php/AAAI/article/view/6394>.
- Zachary Ravichandran, Alexander Robey, Vijay Kumar, George J. Pappas, and Hamed Hassani. Safety guardrails for LLM-enabled robots. *arXiv preprint arXiv:2503.07885*, 2026. IEEE Robotics and Automation Letters.
- Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Toyama, Robert Berry, Divya Tyamagundlu, Timothy Lillicrap, and Oriana Riva. AndroidWorld: A Dynamic Benchmarking Environment for Autonomous Agents, 2024. URL <https://arxiv.org/abs/2405.14573>.
- Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka. May I interrupt? diverging opinions on proactive smart speakers. In *Proceedings of the 3rd Conference on Conversational User Interfaces (CUI)*, pp. 34:1–34:10, 2021. doi: 10.1145/3469595.3469629.
- Gary B Reid and Thomas E Nygren. The subjective workload assessment technique: A scaling procedure for measuring mental workload. In *Advances in psychology*, volume 52, pp. 185–218. Elsevier, 1988.

- Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners, 2023. URL <https://arxiv.org/abs/2307.01928>.
- Shuo Ren, Can Xie, Pu Jian, Zhenjiang Ren, Chunlin Leng, and Jiajun Zhang. Towards scientific intelligence: A survey of llm-based scientific agents, 2026. URL <https://arxiv.org/abs/2503.24047>.
- Zhiyao Ren, Yibing Zhan, Baosheng Yu, Liang Ding, Pingbo Xu, and Dacheng Tao. Healthcare agent: eliciting the power of large language models for medical consultation. *npj Artificial Intelligence*, 1, 2025. URL <https://api.semanticscholar.org/CorpusID:281064731>.
- Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pp. 175–186, 1994.
- Bradley J. Rhodes and Pattie Maes. Just-in-time information retrieval agents. *IBM Systems Journal*, 39(3-4):685–704, 2000.
- Bradley J Rhodes, Thad Starner, et al. Remembrance agent: A continuously running automated information retrieval system. In *PAAM*, volume 96, pp. 487–495, 1996.
- Mark Owen Riedl and Vadim Bulitko. Interactive narrative: An intelligent systems approach. *Ai Magazine*, 34(1):67–67, 2013.
- Dmitriy Rivkin, Francois Hogan, Amal Feriani, Abhisek Konar, Adam Sigal, Steve Liu, and Greg Dudek. Sage: Smart home agent with grounded execution, 2024. URL <https://arxiv.org/abs/2311.00772>.
- Yara Rizk, Vatche Isahagian, Merve Unuvar, and Yasaman Khazaeni. A snooze-less user-aware notification system for proactive conversational agents, 2020. URL <https://arxiv.org/abs/2003.02097>.
- Roborock. Saros Z70 - Grasp the Future with Intelligence. <https://us.roborock.com/pages/roborock-saros-z70>, January 2025. Accessed: 2026-10-05.
- Robotics 24/7. VIDEO: Sunday launches Memo personal robot that ‘actually learns your home’. https://www.robotics247.com/article/video_sunday_launches_memo_personal_robot_that_actually_learns_your_home, November 2025. Accessed: 2026-10-05.
- Neal J. Roesse. Counterfactual thinking. *Psychological Bulletin*, 121(1):133–148, 1997. ISSN 0033-2909. doi: 10.1037/0033-2909.121.1.133. URL <http://dx.doi.org/10.1037/0033-2909.121.1.133>.
- Diederik Marijn Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making, 2014. URL <https://arxiv.org/abs/1402.0590>.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Arturo Rosenblueth, Norbert Wiener, and Julian Bigelow. Behavior, purpose and teleology. *Philosophy of Science*, 10(1):18–24, 1943. doi: 10.1086/286788. URL <https://doi.org/10.1086/286788>.
- M. S. Roulston and L. A. Smith. The boy who cried wolf revisited: The impact of false alarm intolerance on cost-loss scenarios, 2004. URL [https://doi.org/10.1175/1520-0434\(2004\)019<0391:tbwcr>2.0.co;2](https://doi.org/10.1175/1520-0434(2004)019<0391:tbwcr>2.0.co;2).
- Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=GEcwtMk1uA>.
- Andy Rubin. Android @ i/o: The playground is open. <https://blog.google/products-and-platforms/platforms/android/android-io-playground-is-open/>, 2012.

- Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving, 2025. URL <https://arxiv.org/abs/2503.20523>.
- Stuart Russell and Eric Wefald. Principles of metareasoning. *Artificial Intelligence*, 49(1-3):361–395, May 1991. ISSN 0004-3702. doi: 10.1016/0004-3702(91)90015-c. URL [http://dx.doi.org/10.1016/0004-3702\(91\)90015-c](http://dx.doi.org/10.1016/0004-3702(91)90015-c).
- Richard M. Ryan and Edward L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78, 2000. ISSN 0003-066X. doi: 10.1037/0003-066x.55.1.68. URL <http://dx.doi.org/10.1037/0003-066X.55.1.68>.
- Krzysztof Rządca, Paweł Findeisen, Jacek Swiderski, Przemysław Zych, Przemysław Broniek, Jarek Kusmirek, Paweł Nowak, Beata Strack, Piotr Witusowski, Steven Hand, and John Wilkes. Autopilot: workload autoscaling at google. In *Proceedings of the Fifteenth European Conference on Computer Systems, EuroSys '20*, pp. 1–16. ACM, April 2020. doi: 10.1145/3342195.3387524. URL <http://dx.doi.org/10.1145/3342195.3387524>.
- SAE Media Group. How Machina Labs is Reshaping Defense Manufacturing with AI-Driven 7-Axis Robotics. <https://www.mobilityengineeringtech.com/component/content/article/55129-how-machina-labs-is-reshaping-defense-manufacturing-with-ai-driven-7-axis-robotics>, May 2026. Accessed: 2026-10-05.
- Vildan Salikutluk, Janik Schöpfer, Franziska Herbert, Katrin Scheuermann, Eric Frodl, Dirk Balfanz, Frank Jäkel, and Dorothea Koert. An evaluation of situational autonomy for human-ai collaboration in a shared workspace setting. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24*, pp. 1–17. ACM, May 2024. doi: 10.1145/3613904.3642564. URL <http://dx.doi.org/10.1145/3613904.3642564>.
- Chris Samarinas and Hamed Zamani. ProCIS: A benchmark for proactive retrieval in conversations. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (eds.), *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pp. 830–840. ACM, July 2024. doi: 10.1145/3626772.3657869. URL <https://doi.org/10.1145/3626772.3657869>.
- Samsung. How to use now nudge on Samsung Galaxy phone. <https://www.samsung.com/us/support/answer/ANS10010355/>, 2026. Official support documentation. Accessed October 5, 2026.
- Abraham Sanders, Tomek Strzalkowski, Mei Si, Albert Chang, Deepanshu Dey, Jonas Braasch, and Dakuo Wang. Towards a progression-aware autonomous dialogue agent. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1194–1212. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.87. URL <http://dx.doi.org/10.18653/v1/2022.naacl-main.87>.
- Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research. *PLoS computational biology*, 9(10):e1003285, 2013.
- Akshat Sanghvi, Naren Akash, Raza Imam, Amit Sharma, and Mohit Jain. MeDxAgent: Multi-agent consultation for interactive medical diagnosis. arXiv preprint arXiv:2606.03416, 2026. URL <https://arxiv.org/abs/2606.03416>.
- Filippo Santoni de Sio and Jeroen van den Hoven. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5, 2018. doi: 10.3389/frobt.2018.00015.
- Matthew Saponaro, Haoran Wei, and Keith Decker. Towards learning efficient intervention policies for wearable devices. In *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 298–299. IEEE, July 2017. doi: 10.1109/chase.2017.111. URL <http://dx.doi.org/10.1109/chase.2017.111>.
- Alexander Scarlatos, Ryan S Baker, and Andrew Lan. Exploring knowledge tracing in tutor-student dialogues using llms. In *Proceedings of the 15th international learning analytics and knowledge conference*, pp. 249–259, 2025a.

- Alexander Scarlatos, Yusong Wu, Ian Simon, Adam Roberts, Tim Cooijmans, Natasha Jaques, Cassie Tarakajian, and Cheng-Zhi Anna Huang. Realjam: Real-time human-ai music jamming with reinforcement learning-tuned transformers, 2025b. URL <https://arxiv.org/abs/2502.21267>.
- Paul Scerri, David V. Pynadath, and Milind Tambe. Towards adjustable autonomy for the real world. *Journal of Artificial Intelligence Research*, 17:171–228, September 2002. ISSN 1076-9757. doi: 10.1613/jair.1037. URL <https://doi.org/10.1613/jair.1037>.
- Daniel L. Schacter, Donna Rose Addis, and Randy L. Buckner. Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience*, 8(9):657–661, September 2007. ISSN 1471-0048. doi: 10.1038/nrn2213. URL <http://dx.doi.org/10.1038/nrn2213>.
- Peter Schaldenbrand and Jean Oh. Visual sculpting: Visually-aligned planning representations for long-horizon robot clay sculpting. *arXiv preprint arXiv:2605.17556*, 2026.
- Peter Schaldenbrand, James McCann, and Jean Oh. Frida: A collaborative robot painter with a differentiable, real2sim2real planning environment. *arXiv preprint arXiv:2210.00664*, 2022.
- Peter Schaldenbrand, Gaurav Parmar, Jun-Yan Zhu, James McCann, and Jean Oh. Cofrida: Self-supervised fine-tuning for human-robot co-painting. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2296–2302. IEEE, 2024.
- Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. Should i follow ai-based advice? measuring appropriate reliance in human-ai decision-making. *arXiv preprint arXiv:2204.06916*, 2022.
- Bill N. Schilit, Norman Adams, and Roy Want. Context-aware computing applications. In *Proceedings of the 1994 First Workshop on Mobile Computing Systems and Applications*, pp. 85–90. IEEE, December 1994. doi: 10.1109/WMCSA.1994.16. URL <https://doi.org/10.1109/WMCSA.1994.16>.
- Jeremy Schlatter, Benjamin Weinstein-Raun, and Jeffrey Ladish. Incomplete tasks induce shutdown resistance in some frontier llms, 2026. URL <https://arxiv.org/abs/2509.14260>.
- Samuel Schmidgall and Michael Moor. Agentrxiv: Towards collaborative autonomous research, 2025. URL <https://arxiv.org/abs/2503.18102>.
- Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants, 2025. URL <https://arxiv.org/abs/2501.04227>.
- Samuel Schmidgall, Xiaokai Zhu, Marian Shaw, Lin Yang, Valentin Liévin, Jingyun Yang, Yuchen Zhuang, Tim Strother, Alex Bijamov, Min Woo Sun, Anil Palepu, Justin Chen, David Steiner, Jacqueline Shreibati, Wei-Hung Weng, Yilin Zhao, Xingjian Hu, Nicholas Zahn, Sadhya Garg, Julia Kirby, Yuxiang Gan, Jiaoli Li, Divy Thakkar, Shekoofeh Azizi, David Racz, Juraj Gottweis, Vivek Natarajan, Chenglin Wu, Tal Danino, Keran Rong, Haozhe Wang, Benoit Schillings, Yong Cheng, Quoc V. Le, and Tao Tu. Accelerating scientific research with Gemini in the real-world, 2026. URL <https://arxiv.org/abs/2608.26701>.
- Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, September 2010. ISSN 1943-0604. doi: 10.1109/TAMD.2010.2056368. URL <https://doi.org/10.1109/TAMD.2010.2056368>.
- Chris Schmitz, Jonathan Rystrom, and Jan Batzner. Oversight structures for agentic AI in public-sector organizations. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM)*, 2025. doi: 10.18653/v1/2025.realm-1.21.
- Tim Schopf and Michael Färber. Is this idea novel? an automated benchmark for judgment of research ideas, 2026. URL <https://arxiv.org/abs/2603.10303>.
- Martijn Schut, Michael Wooldridge, and Simon Parsons. The theory and practice of intention reconsideration. *Journal of Experimental & Theoretical Artificial Intelligence*, 16(4):261–293, October 2004. ISSN 1362-3079. doi: 10.1080/09528130412331309277. URL <http://dx.doi.org/10.1080/09528130412331309277>.

- Matthew Schwall, Tom Daniel, Trent Victor, Francesca Favaro, and Henning Hohnhold. Waymo public road safety performance data, 2020. URL <https://arxiv.org/abs/2011.00038>.
- Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models’ (lack of) theory of mind: A plug-and-play multi-character belief tracker, 2023. URL <https://arxiv.org/abs/2306.00924>.
- James Scott, A. J. Bernheim Brush, John Krumm, Brian Meyers, Michael Hazas, Stephen Hodges, and Nicolas Villar. PreHeat: Controlling home heating using occupancy prediction. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp)*, pp. 281–290, 2011. doi: 10.1145/2030112.2030151.
- Screenpipe. screenpipe: Ai that knows what you’ve seen, said, or heard. GitHub repository, 2026. URL <https://github.com/screenpipe/screenpipe>. Accessed: 2026-07-01.
- Segway Navimow. *Segway Navimow i110: Smart, Wire-Free Robotic Lawn Mower*. Segway, 2026. URL <https://navimow.com/products/navimow-i110>. Commercial specifications and capabilities for the RTK and vision-assisted autonomous robotic lawn mower designed for residential deployment.
- Martin Seifrid, Robert Pollice, Andres Aguilar-Granda, Zamyra Morgan Chan, Kazuhiro Hotta, Cher Tian Ser, Jenya Vestfrid, Tony C Wu, and Alan Aspuru-Guzik. Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. *Accounts of Chemical Research*, 55(17):2454–2466, 2022.
- B. Sellner, F.W. Heger, L.M. Hiatt, R. Simmons, and S. Singh. Coordinated multiagent teams and sliding autonomy for large-scale assembly. *Proceedings of the IEEE*, 94(7):1425–1444, July 2006. ISSN 1558-2256. doi: 10.1109/jproc.2006.876966. URL <http://dx.doi.org/10.1109/jproc.2006.876966>.
- Procheta Sen, Debasis Ganguly, and Gareth Jones. Procrastination is the thief of time: Evaluating the effectiveness of proactive search systems. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR ’18*, pp. 1157–1160, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210114. URL <https://doi.org/10.1145/3209978.3210114>.
- Gyuhyeon Seo, Jungwoo Yang, Junseong Pyo, Nalim Kim, Jonggeun Lee, and Yohan Jo. Simuhome: A temporal- and environment-aware benchmark for smart home llm agents. *arXiv preprint arXiv:2509.24282*, 2025.
- Pedro Sequeira, Vidyasagar Sadhu, and Melinda Gervasio. Tomcat: Theory-of-mind for cooperative agents in teams via multiagent diffusion policies, 2025. URL <https://arxiv.org/abs/2502.18438>.
- Esmail Seraj, Jerry Xiong, Mariah Schrum, and Matthew C. Gombolay. Mixed-initiative multiagent apprenticeship learning for human training of robot teams. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6f5288d7059cbe3f5a19dad1b3bf17e1-Abstract-Conference.html.
- Preethi Seshadri, Samuel Cahyawijaya, Ayomide Odumakinde, Sameer Singh, and Seraphina Goldfarb-Tarrant. Lost in simulation: Llm-simulated users are unreliable proxies for human users in agentic evaluations, 2026. URL <https://arxiv.org/abs/2601.17087>.
- Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. URL https://burrsettles.com/pub/settles.activelearning_20090109.pdf.
- Zahraa Khais Shahid, Saguna Saguna, and Christer Åhlund. Detecting anomalies in daily activity routines of older persons in single resident smart homes: Proof-of-concept study. *JMIR Aging*, 5(2):e28260, 2022. doi: 10.2196/28260.
- Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.
- Ye Shang, Quanjun Zhang, Haichuan Hu, Chunrong Fang, Liang Xiao, and Zhenyu Chen. Breaking, stale, or missing? benchmarking coding agents on project-level test evolution, 2026. URL <https://arxiv.org/abs/2605.06125>.
- Minghao Shao, Sofija Jancheska, Meet Udeshi, Brendan Dolan-Gavitt, Haoran Xi, Kimberly Milner, Boyuan Chen, Max Yin, Siddharth Garg, et al. NYU CTF Bench: A scalable open-source benchmark dataset for evaluating LLMs

- in offensive security. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 37, 2024. URL <https://arxiv.org/abs/2406.05590>.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2025. URL <https://arxiv.org/abs/2310.13548>.
- Yonadav Shavit, Sandhini Agarwal, et al. Practices for governing agentic AI systems. OpenAI white paper, December 2023. URL <https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf>.
- Ryan Shea and Zhou Yu. AutoSpec: An agentic framework for automatically drafting patent specification. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12829–12840, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.687. URL <https://aclanthology.org/2025.findings-emnlp.687/>.
- Wei Shen, Xiaonan He, Chuheng Zhang, Xuyun Zhang, Xiaolong Xu, and Wanchun Dou. Reward-driven interaction: Enhancing proactive dialogue agents through user satisfaction prediction, 2025. URL <https://arxiv.org/abs/2505.18731>.
- Yifan Shen, Jiateng Liu, Xinzhuo Li, Yuanzhe Liu, Bingxuan Li, Houze Yang, Wenqi Jia, Yijiang Li, Tianjiao Yu, James Matthew Rehg, Xu Cao, and Ismini Lourentzou. Egoforge: Goal-directed egocentric world simulator, 2026a. URL <https://arxiv.org/abs/2603.20169>.
- Zhuokang Shen, Yifan Wang, Hanyu Chen, Yunhang Shen, Wenxuan Huang, Gaoqi He, Jiao Xie, Rongrong Ji, and Shaohui Lin. Groupgpt: A token-efficient and privacy-preserving agentic framework for multi-user chat assistant, 2026b. URL <https://arxiv.org/abs/2603.01059>.
- Thomas B. Sheridan and William L. Verplank. Human and computer control of undersea teleoperators. Technical Report ADA057655, Man-Machine Systems Laboratory, Massachusetts Institute of Technology, Cambridge, MA, July 1978. URL <https://doi.org/10.21236/ada057655>.
- Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. Muma-tom: Multi-modal multi-agent theory of mind, 2025a. URL <https://arxiv.org/abs/2408.12574>.
- Haoting Shi, Wenhao Wang, Weicheng Fang, Yaozhong Liang, Tian Jin, Pengxiang Zhao, Guangyi Liu, Siheng Chen, and Yanfeng Wang. Cua-universe: A scalable and dynamic environment for hybrid gui+cli agents. *arXiv preprint arXiv:2609.05374*, 2026a. URL <https://arxiv.org/abs/2609.05374>.
- Hongrui Shi, Shunbao Li, Zhipeng Yuan, and Po Yang. Pestma: Llm-based multi-agent system for informed pest management. *ArXiv*, abs/2504.09855, 2025b. URL <https://api.semanticscholar.org/CorpusID:277781521>.
- Qinglong Shi, Donghai Wang, Hantao Zhou, Jiguo Li, Jun Xu, Jiuchong Gao, Jinghua Hao, and Renqing He. Long-term task-oriented agent: Proactive long-term intent maintenance in dynamic environments, 2026b. URL <https://arxiv.org/abs/2601.09382>.
- Tianneng Shi, Jingxuan He, Zhun Wang, Hongwei Li, Linyu Wu, Wenbo Guo, and Dawn Song. Progent: Securing ai agents with privilege control, 2026c. URL <https://arxiv.org/abs/2504.11703>.
- Yucheng Shi, Wenhao Yu, Zaitang Li, Yonglin Wang, Hongming Zhang, Ninghao Liu, Haitao Mi, and Dong Yu. Mobilegui-rl: Advancing mobile gui agent through reinforcement learning in online environment, 2025c. URL <https://arxiv.org/abs/2507.05720>.
- Victor Shih, David C Jangraw, Paul Sajda, and Sameer Saproo. Towards personalized human ai interaction - adapting the behavior of ai agents using neural signatures of subjective interest, 2017. URL <https://arxiv.org/abs/1709.04574>.
- Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL <https://arxiv.org/abs/2303.11366>.

- A. N. Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8(1): 22–46, January 1963. ISSN 1095-7219. doi: 10.1137/1108002. URL <http://dx.doi.org/10.1137/1108002>.
- Ben Shneiderman. Creativity support tools: accelerating discovery and innovation. *Communications of the ACM*, 50(12):20–32, 2007.
- Amin Shojaeighadikolaie, Arman Ghasemi, Kailani Jones, Yousif Dafalla, Alexandru G. Bardas, Reza Ahmadi, and Morteza Hashemi. Distributed energy management and demand response in smart grids: A multi-agent deep reinforcement learning framework. *arXiv preprint arXiv:2211.15858*, 2022.
- Yuan Shui, Yandong Guan, Zhanwei Zhang, Juncheng Hu, Jing Zhang, Dong Xu, and Qian Yu. ArtiCAD: Articulated CAD assembly design via multi-agent code generation. *arXiv preprint arXiv:2604.10992*, 2026.
- Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. In *International Conference on Learning Representations*, volume 2025, pp. 94003–94092, 2025.
- Pao Siangliulue, Jonathan Bragg, Doug Downey, Joseph Chee Chang, and Daniel S. Weld. Omakase: proactive assistance with actionable suggestions for evolving scientific research projects, 2026. URL <https://arxiv.org/abs/2604.08898>.
- David Silver and Richard S. Sutton. Welcome to the era of experience, 2025. URL <https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf>. Chapter preprint for Designing an Intelligence, MIT Press.
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017. URL <https://arxiv.org/abs/1712.01815>.
- Herbert A. Simon. A formal theory of the employment relationship. *Econometrica*, 19(3):293–305, 1951. doi: 10.2307/1906815.
- Siva Raja Sindiramutty. Autonomous threat hunting: A future paradigm for ai-driven threat intelligence, 2024. URL <https://arxiv.org/abs/2401.00286>.
- Kunal Pratap Singh, Luca Weihs, Alvaro Herrasti, Jonghyun Choi, Aniruddha Kemhavi, and Roozbeh Mottaghi. Ask4help: Learning to leverage an expert for embodied tasks, 2022. URL <https://arxiv.org/abs/2211.09960>.
- Soumitra Sinhaajari, Navonil Majumder, and Soujanya Poria. On the limits of llm-as-judge for scientific novelty assessment, 2026. URL <https://arxiv.org/abs/2606.12071>.
- Chapa Sirithunge, A. G. Buddhika P. Jayasekara, and D. P. Chandima. Proactive robots with the perception of nonverbal human behavior: A review. *IEEE Access*, 7:77308–77327, 2019. ISSN 2169-3536. doi: 10.1109/access.2019.2921986. URL <http://dx.doi.org/10.1109/access.2019.2921986>.
- Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking, 2025. URL <https://arxiv.org/abs/2209.13085>.
- Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178, May 2021. ISSN 0885-2308. doi: 10.1016/j.csl.2020.101178. URL <http://dx.doi.org/10.1016/j.csl.2020.101178>.
- Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnampati, Samuel G. Rodrigues, and Andrew D. White. Language agents achieve superhuman synthesis of scientific knowledge, 2024. URL <https://arxiv.org/abs/2409.13740>.
- Skild AI. Introducing S1: In-Context Learning for Robotics. <https://www.skild.ai/blogs/s1>, August 2026. Accessed: 2026-10-05.
- Smart Analytics Global. SAG: Global Humanoid Robot Shipments Surged 272% YoY to 19.1K Units in 1H 2026; AGIBOT Overtook Unitree for No. 1 Position. <https://smartanalyticsglobal.com/global-humanoid-robot-shipments-2026-agibot-unitree/>, August 2026. Accessed: 2026-10-05.

- David E. Smith. Choosing objectives in over-subscription planning. In Shlomo Zilberstein, Jana Koehler, and Sven Koenig (eds.), *Proceedings of the Fourteenth International Conference on Automated Planning and Scheduling (ICAPS 2004)*, June 3-7 2004, Whistler, British Columbia, Canada, pp. 393–401. AAAI, 2004. URL <http://www.aaai.org/Library/ICAPS/2004/icaps04-046.php>.
- Nate Soares, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. Corrigibility. In Toby Walsh (ed.), *AAAI Workshops: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, volume WS-15-02 of *AAAI Technical Report*. AAAI Press, 2015. URL <https://intelligence.org/files/Corrigibility.pdf>.
- Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. MileBench: Benchmarking MLLMs in long context. In *First Conference on Language Modeling*, 2024a. URL <https://openreview.net/forum?id=Uhwze2LEwq>.
- Kevin Song, Andrew Trotter, and Jake Y. Chen. Llm agent swarm for hypothesis-driven drug discovery. *ArXiv*, abs/2504.17967, 2025a. URL <https://api.semanticscholar.org/CorpusID:278129440>.
- Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error: Exploration-based trajectory optimization for llm agents, 2024b. URL <https://arxiv.org/abs/2403.02502>.
- Zijian Song, Xiaoxin Lin, Tao Pu, Zhenlong Yuan, Guangrun Wang, and Liang Lin. Human-centric open-future task discovery: Formulation, benchmark, and scalable tree-based search, 2025b. URL <https://arxiv.org/abs/2511.18929>.
- Spencer Soper and Cailley LaPara. UPS Purchases 400 Robots to Unload Trucks in Automation Push. <https://www.ttnews.com/article/ups-robots-unload-trucks>, December 2025. Transport Topics (Bloomberg News). Accessed: 2026-10-05.
- Jonathan Spencer, Sanjiban Choudhury, Matthew Barnes, Matthew Schmittle, Mung Chiang, Peter Ramadge, and Sidd Srinivasa. Expert intervention learning: An online framework for robot learning from explicit and implicit human feedback. *Autonomous Robots*, 46(1):99–113, October 2021. ISSN 1573-7527. doi: 10.1007/s10514-021-10006-9. URL <http://dx.doi.org/10.1007/s10514-021-10006-9>.
- Spotify. Spotify debuts a new ai dj, right in your pocket. <https://newsroom.spotify.com/2023-02-22/spotify-debuts-a-new-ai-dj-right-in-your-pocket/>, 2023.
- Spotify. Dj. <https://support.spotify.com/us/article/dj/>, 2026a.
- Spotify. Prompted playlist in beta coming to premium listeners in more markets. <https://newsroom.spotify.com/2026-01-22/prompted-playlists-expansion/>, 2026b.
- Maxwell Standen, Martin Lucas, David Bowman, Toby J. Richer, Junae Kim, and Damian Marriott. CybORG: A gym for the development of autonomous cyber agents, 2021. URL <https://arxiv.org/abs/2108.09118>.
- Stanford Vision and Learning Lab. 2026 BEHAVIOR Challenge. <https://behavior.stanford.edu/challenge/>, July 2026. Accessed: 2026-10-05.
- Daniel Stenberg. The I in LLM stands for intelligence. Personal blog, January 2024. URL <https://daniel.haxx.se/blog/2024/01/02/the-i-in-llm-stands-for-intelligence/>. Published 2024-01-02; accessed 2026-10-03.
- Daniel Stenberg. Death by a thousand slops. Personal blog, July 2025. URL <https://daniel.haxx.se/blog/2025/07/14/death-by-a-thousand-slops/>. Published 2025-07-14; accessed 2026-10-03.
- Ian A. Stewart. Humanoid Robots In The Home. <https://www.mondaq.com/unitedstates/privacy-protection/1819630/humanoid-robots-in-the-home>, July 2026. Wilson Elser / Mondaq. Accessed: 2026-10-05.
- Victoria Stodden, Marcia McNutt, David H Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A Heroux, John PA Ioannidis, and Michela Taufer. Enhancing reproducibility for computational methods. *Science*, 354(6317): 1240–1241, 2016.
- Maddie Stone. 2022 0-day in-the-wild exploitation... so far. Google Project Zero blog, June 2022. URL <https://googleprojectzero.blogspot.com/2022/06/2022-0-day-in-wild-exploitationso-far.html>. Published 2022-06-30; accessed 2026-10-03.

- Hong Su. Actively obtaining environmental feedback for autonomous action evaluation without predefined measurements, 2026a. URL <https://arxiv.org/abs/2601.04235>.
- Hong Su. Simulating human cognition: Heartbeat-driven autonomous thinking activity scheduling for llm-based ai systems, 2026b. URL <https://arxiv.org/abs/2604.14178>.
- Gita Sukthankar, Christopher Geib, Hung Hai Bui, David Pynadath, and Robert P Goldman. *Plan, activity, and intent recognition: Theory and practice*. Newnes, 2014.
- Theodore R. Summers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, February 2024. URL <https://openreview.net/forum?id=1i6ZCvflQJ>.
- Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 1441–1450, 2019.
- Haotian Sun, Yuchen Zhuang, Ling kai Kong, Bo Dai, and Chao Zhang. Adaplaner: Adaptive planning from feedback with language models, 2023. URL <https://arxiv.org/abs/2305.16653>.
- Haoyu Sun, Wenxuan Wang, Mingyang Song, Jujie He, Weinan Zhang, Yang Liu, Yang Yang, and Yu Cheng. Agent planning benchmark: A diagnostic framework for planning capabilities in LLM agents. *arXiv preprint arXiv:2606.04874*, 2026a. doi: 10.48550/arXiv.2606.04874. URL <https://arxiv.org/abs/2606.04874>.
- Jingyun Sun, Chengxiao Dai, Zhongze Luo, Yangbo Chang, and Yang Li. Lawluo: A multi-agent collaborative framework for multi-round chinese legal consultation. *arXiv preprint arXiv:2407.16252*, 2024a.
- Nan Sun, Bo Mao, Yongchang Li, Lumeng Ma, Di Guo, and Huaping Liu. Assistantx: An llm-powered proactive assistant in collaborative human-populated environments. *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3352–3359, 2024b. URL <https://api.semanticscholar.org/CorpusID:272911445>.
- Nan Sun, Bo Mao, Yongchang Li, Di Guo, and Huaping Liu. Assistantx: An llm-powered proactive assistant in collaborative human-populated environment, 2025a. URL <https://arxiv.org/abs/2409.17655>.
- Pei Sun, Henrik Kretschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.
- Weiwei Sun, Xuhui Zhou, Weihua Du, Xingyao Wang, Sean Welleck, Graham Neubig, Maarten Sap, and Yiming Yang. Training proactive and personalized llm agents, 2025b. URL <https://arxiv.org/abs/2511.02208>.
- Zeyi Sun, Ziyu Liu, Yuhang Zang, Yuhang Cao, Xiaoyi Dong, Tong Wu, Dahua Lin, and Jiaqi Wang. SEAgent: Self-evolving computer use agent with autonomous learning from experience. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pp. 116867–116893. PMLR, 2026b. URL <https://proceedings.mlr.press/v306/sun26w.html>.
- Sunday Robotics. Memo: The Helpful Home Robot. <https://www.sunday.ai/>, 2026. Accessed: 2026-07-20.
- Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4): 160–163, July 1991. ISSN 0163-5719. doi: 10.1145/122344.122377. URL <https://doi.org/10.1145/122344.122377>.
- Richard S. Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, August 1999. ISSN 0004-3702. doi: 10.1016/s0004-3702(99)00052-1. URL [http://dx.doi.org/10.1016/s0004-3702\(99\)00052-1](http://dx.doi.org/10.1016/s0004-3702(99)00052-1).
- Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, JMLR Workshop and Conference Proceedings*, pp. 814–823. JMLR.org, 2015. URL <http://proceedings.mlr.press/v37/swaminathan15.html>.
- Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723, 2025.

- Stefan Szeider. What do llm agents do when left alone? evidence of spontaneous meta-cognitive patterns, 2025. URL <https://arxiv.org/abs/2509.21224>.
- Gadiel Sznaier Camps, Chengyang He, Guillaume Sartoretti, Eduardo Montijano, and Mac Schwager. ScaffoldM3C: A multimodal sequential Monte Carlo framework for generative stable construction planning. *arXiv preprint arXiv:2610.00487*, 2026.
- Karl K. Szpunar. Episodic future thought: An emerging concept. *Perspectives on Psychological Science*, 5(2):142–162, March 2010. ISSN 1745-6924. doi: 10.1177/1745691610362350. URL <http://dx.doi.org/10.1177/1745691610362350>.
- Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624(7990):86, 2023.
- Francesco Taioli, Edoardo Zorzi, Gianni Franchi, Alberto Castellini, Alessandro Farinelli, Marco Cristani, and Yiming Wang. Collaborative instance object navigation: Leveraging uncertainty-awareness to minimize human-agent dialogues, 2025. URL <https://arxiv.org/abs/2412.01250>.
- Weihaio Tan, David Koleczek, Siddhant Pradhan, Nicholas Perello, Vivek Chettiar, Vishal Rohra, Aaslesha Rajaram, Soundararajan Srinivasan, H M Sajjad Hossain, and Yash Chandak. On optimizing interventions in shared autonomy, 2022. URL <https://arxiv.org/abs/2112.09169>.
- Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long T. Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439, 2025. doi: 10.18653/v1/2025.acl-long.413. URL <https://aclanthology.org/2025.acl-long.413/>.
- Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark B. Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. *ArXiv*, abs/2311.10537, 2023. URL <https://api.semanticscholar.org/CorpusID:265281260>.
- Yan Tang, Tingyu Cao, Yuanbo Tang, Huaze Tang, and Keer Hu. Proactive service agents: A unified decision framework, methods, and evaluation. *arXiv preprint arXiv:2609.03727*, 2026a.
- Yuanbo Tang, Huaze Tang, Tingyu Cao, Lam Nguyen, Anping Zhang, Xinwen Cao, Chunkang Liu, Wenbo Ding, and Yang Li. ProAgentBench: Evaluating LLM agents for proactive assistance with real-world data. *arXiv preprint arXiv:2602.04482*, 2026b. doi: 10.48550/arXiv.2602.04482. URL <https://arxiv.org/abs/2602.04482>.
- Manan Tayal and Akshay Nambi. ShieldVLA: Feasibility-aware safety alignment for vision-language-action models. *arXiv preprint arXiv:2609.13231*, 2026.
- TCT Magazine. Divergent unveils large-format metal 3D printer with twelve 2kW lasers. <https://www.tctmagazine.com/divergent-unveils-large-format-metal-3d-printer-with-twelve-2kw-lasers/>, June 2026. Accessed: 2026-10-05.
- Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.
- GigaBrain Team, Angen Ye, Axiang Sun, Can Jin, Chenxi Cheng, Chong Shi, Dengke Shang, Dingqian Zhang, Guan Huang, Guangqiang Wang, et al. Gigabrain-0.7: Scaling embodied foundation models to emergent capabilities with a three-system architecture. *arXiv preprint arXiv:2608.15875*, 2026.
- Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.
- TechCrunch. Robotaxi operators will face fines for blocking first responders. <https://techcrunch.com/2026/10/01/robotaxi-operators-will-face-fines-for-blocking-first-responders/>, October 2026a. California Senate Bill 1246; accessed: 2026-10-05.

- TechCrunch. Waymo accelerates robotaxi expansion with launches in Denver, San Diego, and Tampa. <https://techcrunch.com/2026/09/01/waymo-accelerates-robotaxi-expansion-with-launches-in-denver-san-diego-and-tampa/>, September 2026b. Accessed: 2026-10-05.
- TechCrunch. Waymo is scaling fast: Here’s what the fleet data shows. <https://techcrunch.com/2026/09/24/waymo-is-scaling-fast-heres-what-the-fleet-data-shows/>, September 2026c. Accessed: 2026-10-05.
- TechCrunch. Waymo issues recall to deal with a flooding problem. <https://techcrunch.com/2026/05/12/waymo-issues-recall-to-deal-with-a-flooding-problem/>, May 2026d. Recall of 3,791 vehicles; accessed: 2026-10-05.
- TechCrunch. Waymo robotaxis are headed to Munich. <https://techcrunch.com/2026/08/25/waymo-robotaxis-are-headed-to-munich/>, August 2026e. Accessed: 2026-10-05.
- TechCrunch. Zoox clears final federal hurdle to launch paid robotaxi service. <https://techcrunch.com/2026/07/30/zoox-clears-final-federal-hurdle-to-launch-paid-robotaxi-service/>, July 2026f. Accessed: 2026-10-05.
- TechCrunch. Zoox to start charging for robotaxi rides in Las Vegas. <https://techcrunch.com/2026/08/05/zoox-to-start-charging-for-robotaxi-rides-in-las-vegas/>, August 2026g. Paid service began August 10, 2026; accessed: 2026-10-05.
- TechTimes. Pony.ai Proves Robotaxi Profit, Then Deploys 2,000 Robotaxis Across Europe. <https://www.techtimes.com/articles/324445/20260814/ponyai-proves-robotaxi-profit-then-deploys-2000-robotaxis-across-europe.htm>, August 2026. Accessed: 2026-10-05.
- Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. Asking for help using inverse semantics, 2014. URL <https://doi.org/10.15607/rss.2014.x.024>.
- David L. Tennenhouse. Proactive computing. *Communications of the ACM*, 43(5):43–50, 2000. doi: 10.1145/332833.332837. URL <https://doi.org/10.1145/332833.332837>.
- Tesla, Inc. Q2 2026 Update. <https://www.sec.gov/Archives/edgar/data/0001318605/000162828026049213/exhibit991.htm>, July 2026. Form 8-K, Exhibit 99.1. Accessed: 2026-10-05.
- Minh V. T. Thai, Tue Le, Dung Nguyen Manh, Huy Phan Nhat, and Nghi D. Q. Bui. SWE-EVO: Benchmarking coding agents in long-horizon software evolution scenarios. *arXiv preprint arXiv:2512.18470*, 2025. URL <https://arxiv.org/abs/2512.18470>.
- John Thangarajah, Lin Padgham, and Michael Winikoff. Detecting & avoiding interference between goals in intelligent agents. In Georg Gottlob and Toby Walsh (eds.), *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, pp. 721–726. Morgan Kaufmann, 2003. URL <http://ijcai.org/Proceedings/03/Papers/105.pdf>.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback, 2023. URL <https://arxiv.org/abs/2305.14975>.
- TikTok. How tiktok recommends videos #foryou. <https://newsroom.tiktok.com/how-tiktok-recommends-videos-for-you>, 2020.
- Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.
- Gary Tom, Stefan P Schmid, Sterling G Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M Rajaonson, Marta Skreta, et al. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732, 2024.
- Josue Torres-Fonseca, Naihao Deng, Yinpei Dai, Shane Storks, Yichi Zhang, Rada Mihalcea, Casey Kennington, and Joyce Chai. Safetyalfred: Evaluating safety-conscious planning of multimodal large language models, 2026. URL <https://arxiv.org/abs/2604.19638>.

- Tram Thi Minh Tran, Debargha Dey, and Martin Tomitsch. Rethinking external communication of autonomous vehicles: Is the field converging, diverging, or stalling? *arXiv preprint arXiv:2602.06278*, 2026.
- Tu Trinh, Mohamed Elfeki, Guangze Luo, Kelvin Luu, Nathan Hunt, Ernesto Hernandez, Nandan Marwaha, Yanis Yiming He, Charles Wang, Fernando Carabedo, et al. Hil-bench (human-in-loop benchmark): Do agents know when to ask for help? *arXiv preprint arXiv:2604.09408*, 2026.
- Pamela S Tsang and Velma L Velazquez. Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, 39(3):358–381, 1996.
- Endel Tulving et al. Episodic and semantic memory. *Organization of memory*, 1(381-403):1, 1972.
- Ada Defne Tur, Nicholas Meade, Xing Han Lü, Alejandra Zambrano, Arkil Patel, Esin Durmus, Spandana Gella, Karolina Stańczak, and Siva Reddy. Safearena: Evaluating the safety of autonomous web agents, 2025. URL <https://arxiv.org/abs/2503.04957>.
- Melissa Turcotte, François Labrèche, and Serge-Olivier Paquette. Automated alert classification and triage (aact): An intelligent system for the prioritisation of cybersecurity alerts, 2025. URL <https://arxiv.org/abs/2505.09843>.
- Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation, 2020a. URL <https://arxiv.org/abs/1902.09725>.
- Alexander Matt Turner, Neale Ratzlaff, and Prasad Tadepalli. Avoiding side effects in complex environments, 2020b. URL <https://arxiv.org/abs/2006.06547>.
- Iis P. Tussyadiah and Dan Wang. Tourists’ attitudes toward proactive smartphone systems. *Journal of Travel Research*, 55(4):493–508, December 2014. ISSN 1552-6763. doi: 10.1177/0047287514563168. URL <http://dx.doi.org/10.1177/0047287514563168>.
- Uber Technologies, Inc. Wayve and Uber Launch First-Ever Autonomous Rides in the UK. <https://investor.uber.com/news-events/news/press-release-details/2026/Wayve-and-Uber-Launch-First-Ever-Autonomous-Rides-in-the-UK-2026-VoFQI1WbQi/default.aspx>, September 2026. Accessed: 2026-10-05.
- Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Gengyu Wang. From recall to forgetting: Benchmarking long-term memory for personalized agents. *arXiv preprint arXiv:2604.20006*, 2026. doi: 10.48550/arXiv.2604.20006. URL <https://arxiv.org/abs/2604.20006>.
- UK AI Security Institute. Incident report: Unsolicited agent behaviour during cyber testing. Official incident report, August 2026. URL <https://www.aisi.gov.uk/blog/incident-report-unsolicited-agent-behaviour-during-cyber-testing>. Published August 2026; accessed 2026-09-12.
- Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks, 2023. URL <https://arxiv.org/abs/2302.08399>.
- Blase Ur, Elyse McManus, Melwyn Pak Yong Ho, and Michael L. Littman. Practical trigger-action programming in the smart home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’14, pp. 803–812. ACM, April 2014. doi: 10.1145/2556288.2557420. URL <http://dx.doi.org/10.1145/2556288.2557420>.
- Simon Urli, Zhongxing Yu, Lionel Seinturier, and Martin Monperrus. How to design a program repair bot? insights from the repairator project, 2018. URL <https://arxiv.org/abs/1811.09852>.
- Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. PlanBench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In *Advances in Neural Information Processing Systems*, volume 36, pp. 38975–38987, 2023. doi: 10.52202/075280-1693. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/7a92bcdede88c7afd108072faf5485c8-Abstract-Datasets_and_Benchmarks.html.
- Tim van Kasteren, Athanasios Noulas, Gwenn Engleblenne, and Ben Kröse. Accurate activity recognition in a home setting. In *Proceedings of the 10th International Conference on Ubiquitous Computing (UbiComp)*, pp. 1–9, 2008. doi: 10.1145/1409635.1409637.

- Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q. Vera Liao, and Jennifer Wortman Vaughan. Generation probabilities are not enough: Uncertainty highlighting in AI code completions. *ACM Transactions on Computer-Human Interaction*, 32(1):4:1–4:30, April 2025. doi: 10.1145/3702320. URL <https://doi.org/10.1145/3702320>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- Swaroop Vattam, Matthew Klenk, Matthew Molineaux, and David W. Aha. Breadth of approaches to goal reasoning: A research survey. Technical Report CS-TR-5029, Department of Computer Science, University of Maryland, College Park, MD, 2013. URL <https://drum.lib.umd.edu/items/a8fba90e-0485-4524-b0fd-05538c256822>. In *Goal Reasoning: Papers from the ACS Workshop*, edited by David W. Aha, Michael T. Cox, and Héctor Muñoz-Avila, pp. 111–126.
- José R. Vázquez-Canteli and Zoltán Nagy. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied Energy*, 235:1072–1089, 2019. doi: 10.1016/j.apenergy.2018.11.002.
- José R. Vázquez-Canteli, Sourav Dey, Gregor Henze, and Zoltán Nagy. CityLearn: Standardizing research in multi-agent reinforcement learning for demand response and urban energy management. *arXiv preprint arXiv:2012.10504*, 2020.
- Kalyan Veeramachaneni, Ignacio Araldo, Vamsi Korrapati, Constantinos Bassias, and Ke Li. AI²: Training a big data machine to defend. In *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, pp. 49–54, 2016. doi: 10.1109/BigDataSecurity-HPSC-IDS.2016.79. URL <https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.79>.
- Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22184–22202. PMLR, 2022.
- Adela Nedisan Videsjorden, Nikolay Nikolov, Carl-Henrik Lien, Arda Goknil, Sagar Sen, Hui Song, Ahmet Soylu, and Dumitru Roman. Positioning llm-enabled agents as legal compliance aides for data pipelines. In *International Joint Conference on Rules and Reasoning*, pp. 227–236. Springer, 2025.
- Laura Villa, Ramón Hervás, Luis Cabañero, Jesús Fontecha, Gustavo López Herrera, and Jesús Favela. Evaluating proactivity levels in socially assistive robots for elderly care: a user adoption assessment. *Behaviour & Information Technology*, pp. 1–26, 2025. doi: 10.1080/0144929X.2025.2546976. URL <https://doi.org/10.1080/0144929X.2025.2546976>.
- Gabriel Villarrubia, Juan Francisco de Paz, Daniel Hernández de La Iglesia, and Javier Bajo. Combining multi-agent systems and wireless sensor networks for monitoring crop irrigation. *Sensors (Basel, Switzerland)*, 17, 2017. URL <https://api.semanticscholar.org/CorpusID:3508808>.
- Vaishali Vinay. The evolution of agentic ai in cybersecurity: From single llm reasoners to multi-agent systems and autonomous pipelines, 2025. URL <https://arxiv.org/abs/2512.06659>.
- Thomas Vitry, Vanessa Maeder, Kieran von Valeburg, Asihati Hazaiti, Doga Deniz Ates, Connor Gäde, Jan-Gerrit Habekost, Dennis Becker, and Stefan Wermter. When may i help you? on the effect of proactivity on group human-robot collaboration. *arXiv preprint arXiv:2606.28469*, 2026.
- Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *SSRN Electronic Journal*, 2017. ISSN 1556-5068. doi: 10.2139/ssrn.3063289. URL <http://dx.doi.org/10.2139/ssrn.3063289>.
- Marilyn Walker and Steve Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 1990. URL <https://arxiv.org/abs/cmp-lg/9504007>.
- Marilyn Walker, Diane Litman, Candace A Kamm, and Alicia Abella. Paradise: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 271–280, 1997.

- Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training llms to prioritize privileged instructions, 2024. URL <https://arxiv.org/abs/2404.13208>.
- Boyuan Wang, Bohan Li, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Jie Li, Jindi Lv, Jingyu Liu, Lv Feng, et al. Gigabrain-0.5m*: a vla that learns from world model-based reinforcement learning. *arXiv preprint arXiv:2602.12099*, 2026a.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, March 2024a. URL <https://openreview.net/forum?id=ehfRiF0R3a>.
- Hanyu Wang, Yuanpu Cao, Lu Lin, and Jinghui Chen. Preflect: From retrospective to prospective reflection in large language model agents, 2026b. URL <https://arxiv.org/abs/2602.07187>.
- Haoyu Wang, Tao Li, Zhiwei Deng, Dan Roth, and Yang Li. Devil’s advocate: Anticipatory reflection for llm agents, 2024b. URL <https://arxiv.org/abs/2405.16334>.
- Haoyu Wang, Christopher M. Poskitt, and Jun Sun. Agentspec: Customizable runtime enforcement for safe and reliable llm agents, 2025a. URL <https://arxiv.org/abs/2503.18666>.
- Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3221–3241, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.166. URL <https://aclanthology.org/2025.naacl-long.166/>.
- Jian Wang, Dongding Lin, and Wenjie Li. Follow me: Conversation planning for target-driven recommendation dialogue systems. *arXiv preprint arXiv:2208.03516*, 2022.
- Jinyang Wang, Shiwei Li, Junjian Wang, Zhiqiang Deng, Jianbin Gao, Yihang Zhao, Liu Liu, Yongjia Zhao, Jinlong Chen, Huirui Xu, Yifeng Pan, Kangwei Liu, Fan Ren, Ji Tao, and Minghao Yang. Sv-wam: An efficient surround-view world-action model for end-to-end autonomous driving. *arXiv preprint arXiv:2609.03602*, 2026c.
- Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models, 2023a. URL <https://arxiv.org/abs/2305.04091>.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), March 2024c. ISSN 2095-2236. doi: 10.1007/s11704-024-40231-1. URL <http://dx.doi.org/10.1007/s11704-024-40231-1>.
- Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. User behavior simulation with large language model based agents, 2024d. URL <https://arxiv.org/abs/2306.02552>.
- Liang Wang, Heng Meng, Zekai Xiang, Jin Liu, Pingyi Zhou, Litao Chen, and Yongqiang Tang. Text2CAD-Bench: A benchmark for LLM-based text-to-parametric CAD generation. *arXiv preprint arXiv:2605.18430*, 2026d.
- Peiran Wang, Ying Li, and Yuan Tian. Reframing llm agent security as an agent-human interaction problem, 2026e. URL <https://arxiv.org/abs/2605.24309>.
- Peng Wang, Xiang Wei, Fangxu Hu, and Wenjuan Han. TransGPT: Multi-modal generative pre-trained transformer for transportation. *arXiv preprint arXiv:2402.07233*, 2024e.
- Pengyu Wang, Shuchang Ye, Usman Naseem, and Jinman Kim. Mrgagents: A multi-agent framework for improved medical report generation with med-llms. *2025 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–7, 2025c. URL <https://api.semanticscholar.org/CorpusID:278905046>.
- Prince Zizhuang Wang and Shuli Jiang. Prime: Training free proactive reasoning via iterative memory evolution for user-centric agent, 2026. URL <https://arxiv.org/abs/2604.07645>.

- Qian Wang, Jiaying Wu, Zichen Jiang, Zhenheng Tang, Bingqiao Luo, Nuo Chen, Wei Chen, and Bingsheng He. Llm-based human simulations have not yet been reliable, 2025d. URL <https://arxiv.org/abs/2501.08579>.
- Qiuyue Wang, Mingsheng Li, Jian Guan, Jinhui Ye, Sicheng Xie, Yitao Liu, Junhao Chen, Zhixuan Liang, Jie Zhang, Xintong Hu, et al. Qwen-vla: Unifying vision-language-action modeling across tasks, environments, and robot embodiments. *arXiv preprint arXiv:2605.30280*, 2026f.
- Qiyao Wang, Shiwen Ni, Huaren Liu, Shule Lu, Guhong Chen, Xi Feng, Chi Wei, Qiang Qu, Hamid Alinejad-Rokny, Yuan Lin, et al. Autopatent: a multi-agent framework for automatic patent generation. *arXiv preprint arXiv:2412.09796*, 2024f.
- Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. Sotopia- π : Interactive learning of socially intelligent language agents, 2024g. URL <https://arxiv.org/abs/2403.08715>.
- Ruotong Wang, Xinyi Zhou, Lin Qiu, Joseph Chee Chang, Jonathan Bragg, and Amy X. Zhang. Social-rag: Retrieving from group interactions to socially ground ai generation, 2025e. URL <https://arxiv.org/abs/2411.02353>.
- Ruoyao Wang, Graham Todd, Ziang Xiao, Xingdi Yuan, Marc-Alexandre Côté, Peter Clark, and Peter Jansen. Can language models serve as text-based world simulators?, 2024h. URL <https://arxiv.org/abs/2406.06485>.
- Rushi Wang, Jiateng Liu, Cheng Qian, Yifan Shen, Yanzhou Pan, Zhaozhuo Xu, Ahmed Abbasi, Heng Ji, and Denghui Zhang. Context engineering for trustworthiness: Rescorla wagner steering under mixed and inappropriate contexts, 2025f. URL <https://arxiv.org/abs/2509.04500>.
- Suli Wang, Yiqun Duan, Yu Deng, Rundong Zhao, Dai Shi, and Xinliang Zhou. Cognifold: Always-on proactive memory via cognitive folding, 2026g. URL <https://arxiv.org/abs/2605.13438>.
- Tianfu Wang, Zhezheng Hao, Xilin Xia, Lixin Liu, Mengkang Hu, Hongzhang Liu, Xi Chen, Ziyang Liu, Xiankun Lin, Weijia Zhang, Nicholas Jing Yuan, and Hui Xiong. Agentic artifact creation: Systems, evaluation, principles, and opportunities. *arXiv preprint arXiv:2608.28122*, 2026h.
- Wei Wang, Wenqiao Zhang, Yutong Lin, Yuqian Yuan, Tianwei Lin, Jinhao Mao, Zhenxuan Fan, Mingjian Gao, Yang Dai, Wentong Li, et al. Embodiedskills: A unified framework for orchestrating, training, and deploying vla agents. *arXiv preprint arXiv:2609.01281*, 2026i.
- Weixuan Wang, Dongge Han, Daniel Madrigal Diaz, Jin Xu, Victor Rühle, and Saravan Rajmohan. Odysseybench: Evaluating llm agents on long-horizon complex office application workflows. *arXiv preprint arXiv:2508.09124*, 2025g. URL <https://arxiv.org/abs/2508.09124>.
- Weiyun Wang, Shuibao Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, Xizhou Zhu, Ping Luo, Yu Qiao, Jifeng Dai, Wenqi Shao, and Wenhao Wang. Needle in a multimodal haystack. In *Advances in Neural Information Processing Systems*, volume 37, pp. 20540–20565, 2024i. doi: 10.52202/079017-0649. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/24a8968affe71ffe4067d022b9d16566-Abstract-Datasets_and_Benchmarks_Track.html.
- Wenkai Wang, Tao Xiong, Jingchen Ni, Yunpeng Bao, Xiyun Li, Tianqi Liu, Hongcan Guo, Zilong Huang, and Shengyu Zhang. DeskCraft: Benchmarking Desktop Agents on Professional Workflows and Human-in-the-Loop Collaboration, 2026j. URL <https://arxiv.org/abs/2606.03103>.
- Wenxuan Wang, Juluan Shi, Zixuan Ling, Yuk-Kit Chan, Chaozheng Wang, Cheryl Lee, Youliang Yuan, Jen tse Huang, Wenxiang Jiao, and Michael R. Lyu. Learning to ask: When llm agents meet unclear instruction, 2026k. URL <https://arxiv.org/abs/2409.00557>.
- Xiaoyang Wang and Christopher C. Yang. Medihive: A decentralized agent collective for medical reasoning. *ArXiv*, abs/2603.27150, 2026. URL <https://api.semanticscholar.org/CorpusID:286961612>.
- Xin Wang, Yifan Zhang, Xiaojing Zhang, Longhui Yu, Xinna Lin, Jindong Jiang, Bin Ma, and Kaicheng Yu. Patentagent: Intelligent agent for automated pharmaceutical patent analysis. *arXiv preprint arXiv:2410.21312*, 2024j.
- Xingyao Wang, Boxuan Li, Yufan Song, Frank F. Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, Hoang H. Tran, Fuqiang Li, Ren Ma, Mingzhang Zheng, Bill Qian, Yanjun Shao, Niklas Muennighoff,

- Yizhe Zhang, Binyuan Hui, Junyang Lin, Robert Brennan, Hao Peng, Heng Ji, and Graham Neubig. OpenHands: An open platform for AI software developers as generalist agents. In *The Thirteenth International Conference on Learning Representations*, 2025h. URL <https://openreview.net/forum?id=0Jd3ayDDoF>.
- Xinyue Wang, Yuanhe Zhang, Zhengshuo Gong, Haoran Gao, Fanyu Meng, Zhenhong Zhou, Li Sun, Yang Liu, and Sen Su. From helpfulness to toxic proactivity: Diagnosing behavioral misalignment in llm agents, 2026l. URL <https://arxiv.org/abs/2602.04197>.
- Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025i.
- Yiding Wang, Yuxuan Chen, Fangwei Zhong, Long Ma, and Yizhou Wang. Simulating human-like daily activities with desire-driven autonomy, 2025j. URL <https://arxiv.org/abs/2412.06435>.
- Yu Wang and Xi Chen. Mirix: Multi-agent memory system for llm-based agents, 2025. URL <https://arxiv.org/abs/2507.07957>.
- Yueqian Wang, Xiaojun Meng, Yifan Wang, Huishuai Zhang, and Dongyan Zhao. Proactivevideoqa: A comprehensive benchmark evaluating proactive interactions in video large language models. *arXiv preprint arXiv:2507.09313*, 2025k.
- Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. *arXiv preprint arXiv:2311.01455*, 2023b.
- Yufeng Wang, Jinwu Hu, Ziteng Huang, Kunyang Lin, Zitian Zhang, Peihao Chen, Yu Hu, Qianye Wang, Zhuliang Yu, Bin Sun, Xiaofen Xing, Qingfang Zheng, and Mingkui Tan. Enhancing user-oriented proactivity in open-domain dialogues with critic guidance, 2025l. URL <https://arxiv.org/abs/2505.12334>.
- Yuran Wang, Siqiao Huang, Mingleyang Li, Chenhao Zhang, Jiaqi Liang, Weiyang Jin, et al. Openwam: An open, modular exploration towards systematic world-action model pretraining. *arXiv preprint arXiv:2609.07398*, 2026m.
- Zhimin Wang, Duo Wu, Shaokang He, Jinghe Wang, Linjia Kang, Jing Yu, Kai Zhu, Jiawei Li, and Zhi Wang. Collaborative belief reasoning with llms for efficient multi-agent collaboration, 2026n. URL <https://arxiv.org/abs/2509.21981>.
- Zhun Wang, Tianneng Shi, Jingxuan He, Matthew Cai, Jialin Zhang, and Dawn Song. CyberGym: Evaluating AI agents' cybersecurity capabilities with real-world vulnerabilities at scale, 2025m. URL <https://arxiv.org/abs/2506.02548>.
- Zifeng Wang, Chun-Liang Li, Vincent Perot, Long T. Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. CodecLM: Aligning language models with tailored synthetic data. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024k.
- Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents, 2024l. URL <https://arxiv.org/abs/2302.01560>.
- Zilong Wang, Yuedong Cui, Li Zhong, Zimin Zhang, Da Yin, Bill Yuchen Lin, and Jingbo Shang. OfficeBench: Benchmarking Language Agents across Multiple Applications for Office Automation, 2024m. URL <https://arxiv.org/abs/2407.19056>.
- Zilong Wang, Zifeng Wang, Long T. Le, Huaixiu Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, Anush Mattapalli, Ankur Taly, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. Speculative RAG: Enhancing retrieval augmented generation through drafting. In *International Conference on Learning Representations (ICLR)*, 2025n. URL <https://openreview.net/forum?id=xgQfWbV6Ey>.
- Ziqi Wang and Boqin Yuan. L-mars: Legal multi-agent workflow with orchestrated reasoning and agentic search. *arXiv preprint arXiv:2509.00761*, 2025.

- Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiakuan Li, and Yueming Jin. MedAgent-Pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. In *International Conference on Learning Representations*, 2026o. URL <https://openreview.net/forum?id=ZOuUOudyA4>.
- Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63897–63911. PMLR, July 2025o. URL <https://proceedings.mlr.press/v267/wang25bx.html>.
- William Watson, Nicole Cho, Nishan Srishankar, Zhen Zeng, Lucas Cecchi, Daniel Scott, Suchetha Siddagangappa, Rachneet Kaur, Tucker Balch, and Manuela Veloso. Law: Legal agentic workflows for custody and fund services contracts. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 583–594, 2025.
- Waymo. Scaling Waymo One Safely Across Four Cities This Year. <https://waymo.com/blog/2024/03/scaling-waymo-one-safely-across-four-cities-this-year/>, March 2024. Accessed: 2026-07-20.
- Waymo. Building Waymo’s Risk and Insurance Foundation in Europe with Allianz Partners. <https://waymo.com/blog/2026/09/allianzpartnership>, September 2026a. States “more than half a million paid trips weekly”; accessed: 2026-10-05.
- Waymo. From the road — September 24, 2026: Across Waymo’s 270 million autonomous miles. <https://waymo.com/blog/shorts/safetydata-september26/>, September 2026b. Reports 841 fewer injury-causing crashes (82%) and 95% fewer serious-injury-or-worse crashes vs. human benchmark, data through June 2026; accessed: 2026-10-05.
- Waymo. Singapore, Next Stop: Bringing Scalable, Safe Autonomous Mobility to the Lion City. <https://waymo.com/blog/2026/09/waymo-in-singapore>, September 2026c. Accessed: 2026-10-05.
- Waymo. The Waymo World Model: A New Frontier For Autonomous Driving Simulation. <https://waymo.com/blog/2026/02/the-waymo-world-model-a-new-frontier-for-autonomous-driving-simulation/>, February 2026d. Built on Google DeepMind’s Genie 3; accessed: 2026-10-05.
- Adam Waytz, Joy Heafner, and Nicholas Epley. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52:113–117, 2014. doi: 10.1016/j.jesp.2014.01.005.
- Wayve. GAIA-4: Multimodal World Models Powering Closed-Loop Simulation for Safe and Scalable Autonomy. <https://wayve.ai/thinking/gaia-4/>, August 2026a. Accessed: 2026-10-05.
- Wayve. Wayve, Uber and Nissan Announce Collaboration on Robotaxis. <https://wayve.ai/press/wayve-nissan-uber-robotaxi-collaboration/>, March 2026b. Accessed: 2026-10-05.
- Wayve. Wayve Secures \$1.5B to Deploy Its Global Autonomy Platform. <https://wayve.ai/press/series-d/>, February 2026c. Accessed: 2026-07-20.
- WBAP. Volvo, Waabi Launch Autonomous Freight Runs From Dallas To Houston. <https://www.wbap.com/2026/10/05/volvo-waabi-launch-autonomous-freight-runs-from-dallas-to-houston/>, October 2026. Accessed: 2026-10-05.
- Weave Robotics. Isaac 1. <https://www.weaverobotics.com/isaac-1>, July 2026. Accessed: 2026-10-05.
- Gabrielle Wehr, Reuben Rideaux, Amaya J. Fox, David R. Lightfoot, Jason Tangen, Jason B. Mattingley, and Shane E. Ehrhardt. Virtuous machines: Towards artificial general science, 2026. URL <https://arxiv.org/abs/2508.13421>.
- Fei Wei, Daoyuan Chen, Ce Wang, Yilun Huang, Yushuo Chen, Xuchen Pan, Yaliang Li, and Bolin Ding. Grounded in reality: Learning and deploying proactive llm from offline logs, 2025a. URL <https://arxiv.org/abs/2510.25441>.
- Hua Wei, Guanjie Zheng, Huaxiu Yao, and Zhenhui Li. IntelliLight: A reinforcement learning approach for intelligent traffic light control. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 2496–2505, 2018. doi: 10.1145/3219819.3220096.

- Hua Wei, Chacha Chen, Guanjie Zheng, Kan Wu, Vikash Gayah, Kai Xu, and Zhenhui Li. PressLight: Learning max pressure control to coordinate traffic signals in arterial network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 1290–1298, 2019a. doi: 10.1145/3292500.3330949.
- Hua Wei, Nan Xu, Huichu Zhang, Guanjie Zheng, Xinshi Zang, Chacha Chen, Weinan Zhang, Yanmin Zhu, Kai Xu, and Zhenhui Li. Colight: Learning network-level cooperation for traffic signal control. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19*, pp. 1913–1922, New York, NY, USA, 2019b. Association for Computing Machinery. ISBN 9781450369763. doi: 10.1145/3357384.3357902. URL <https://doi.org/10.1145/3357384.3357902>.
- Jinbiao Wei, Qianran Ma, Yilun Zhao, Xiao Zhou, Kangqi Ni, Guo Gan, and Arman Cohan. OpenComputer: Verifiable Software Worlds for Computer-Use Agents, 2026a. URL <https://arxiv.org/abs/2605.19769>.
- Jing Wei, Tilman Dingler, and Vassilis Kostakos. Understanding user perceptions of proactive smart speakers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(4), December 2022a. doi: 10.1145/3494965. URL <https://doi.org/10.1145/3494965>.
- Jing Wei, Benjamin Tag, Johanne R. Trippas, Tilman Dingler, and Vassilis Kostakos. What could possibly go wrong when interacting with proactive smart speakers? a case study using an esm application. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI)*, 2022b. doi: 10.1145/3491102.3517432.
- Jinjie Wei, Dingkan Yang, Yanshu Li, Qingyao Xu, Zhaoyu Chen, Mingcheng Li, Yue Jiang, Xiaolu Hou, and Lihua Zhang. Medaide: Towards an omni medical aide via specialized llm-based multi-agent collaboration. *ArXiv*, abs/2410.12532, 2024. URL <https://api.semanticscholar.org/CorpusID:281616098>.
- Xinming Wei, Jiahao Zhang, Haoran Li, Jiayu Chen, Haoning Guan, Rui Qu, Maoliang Li, Xiang Chen, and Guojie Luo. Agent.xpu: Efficient scheduling of agentic llm workloads on heterogeneous soc, 2026b. URL <https://arxiv.org/abs/2506.24045>.
- Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, Hyokun Yun, and Lihong Li. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7909–7928, Suzhou, China, 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.401. URL <https://aclanthology.org/2025.emnlp-main.401/>.
- Markus Weinmann, Christoph Schneider, and Jan vom Brocke. Digital nudging. *Business & Information Systems Engineering*, 58(6):433–436, October 2016. ISSN 1867-0202. doi: 10.1007/s12599-016-0453-1. URL <http://dx.doi.org/10.1007/s12599-016-0453-1>.
- Mark Weiser. The computer for the 21st century. *Scientific American*, 265(3):94–104, 1991. doi: 10.1038/scientificamerican0991-94.
- Mark Weiser and John Seely Brown. *The Coming Age of Calm Technology*, pp. 75–85. Springer New York, 1997. ISBN 9781461206859. doi: 10.1007/978-1-4612-0685-9_6. URL http://dx.doi.org/10.1007/978-1-4612-0685-9_6.
- Justin D. Weisz, Michael Muller, Arielle Goldberg, and Dario Andres Silva Moran. Expedient assistance and consequential misunderstanding: Envisioning an operationalized mutual theory of mind, 2024. URL <https://arxiv.org/abs/2406.11946>.
- Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models, 2025a. URL <https://arxiv.org/abs/2407.18418>.
- Haowei Wen, Ruixuan Liu, Weiyi Piao, Siyu Li, and Changliu Liu. Bricksim: A physics-based simulator for manipulating interlocking brick assemblies. *arXiv preprint arXiv:2603.16853*, 2026.
- Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models, 2024. URL <https://arxiv.org/abs/2309.16292>.

- Zichen Wen, Yiyu Wang, Chenfei Liao, Boxue Yang, Junxian Li, Weifeng Liu, Haocong He, Bolong Feng, Xuyang Liu, Yuanhuiyi Lyu, Xu Zheng, Xuming Hu, and Linfeng Zhang. Ai for service: Proactive assistance with ai glasses. *ArXiv*, abs/2510.14359, 2025b. URL <https://api.semanticscholar.org/CorpusID:282139278>.
- Brianna Wessling. Agility’s Digit 5 humanoid has new legs, batteries, and safety upgrades. <https://www.therobotreport.com/agilitys-digit-5-humanoid-has-new-legs-batteries-safety-upgrades/>, September 2026. The Robot Report. Accessed: 2026-10-05.
- David Weston. Helping our customers through the CrowdStrike outage. Official Microsoft Blog, July 2024. URL <https://blogs.microsoft.com/blog/2024/07/20/helping-our-customers-through-the-crowdstrike-outage/>. Published 2024-07-20; accessed 2026-10-03.
- Alan White, Austin Tate, and Michael Rovatsos. *CAMP-BDI: A Pre-emptive Approach for Plan Execution Robustness in Multiagent Systems*, pp. 65–84. Springer International Publishing, 2015. ISBN 9783319255248. doi: 10.1007/978-3-319-25524-8_5. URL http://dx.doi.org/10.1007/978-3-319-25524-8_5.
- Norbert Wiener. *Cybernetics: Or Control and Communication in the Animal and the Machine*. The Technology Press, Cambridge, MA, 1948.
- Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. Think twice: Perspective-taking improves large language models’ theory-of-mind capabilities, 2023. URL <https://arxiv.org/abs/2311.10227>.
- Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.
- Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. The dialog state tracking challenge. In Maxine Eskenazi, Michael Strube, Barbara Di Eugenio, and Jason D. Williams (eds.), *Proceedings of the SIGDIAL 2013 Conference*, pp. 404–413, Metz, France, August 2013. Association for Computational Linguistics. URL <https://aclanthology.org/W13-4065/>.
- Kevin Williams, Elizabeth Bilsland, Andrew Sparkes, Wayne Aubrey, Michael Young, Larisa N Soldatova, Kurt De Grave, Jan Ramon, Michaela De Clare, Worachart Sirawaraporn, et al. Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *Journal of the Royal society Interface*, 12(104): 20141289, 2015.
- Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986*, 2023.
- Oliver E. Williamson. *Markets and Hierarchies: Analysis and Antitrust Implications*. Free Press, New York, 1975.
- Oliver E. Williamson. *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. Free Press, New York, 1985.
- Ethan Wilson, Naveen Sendhilnathan, Charlie S. Burlingham, Yusuf Mansour, Robert Cavin, Sai Deep Tetali, Ajoy Savio Fernandes, and Michael J. Proulx. Eye gaze as a signal for conveying user attention in contextual ai systems, 2025. URL <https://arxiv.org/abs/2501.13878>.
- H Wimmer. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception, 1983. URL [https://doi.org/10.1016/0010-0277\(83\)90004-5](https://doi.org/10.1016/0010-0277(83)90004-5).
- Craig Wisneski, Hiroshi Ishii, Andrew Dahley, Matt Gorbet, Scott Brave, Brygg Ullmer, and Paul Yarin. *Ambient Displays: Turning Architectural Space into an Interface between People and Digital Information*, pp. 22–32. Springer Berlin Heidelberg, 1998. ISBN 9783540697060. doi: 10.1007/3-540-69706-3_4. URL http://dx.doi.org/10.1007/3-540-69706-3_4.
- Daniel M. Wolpert, Zoubin Ghahramani, and Michael I. Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, 1995. doi: 10.1126/science.7569931. URL <https://doi.org/10.1126/science.7569931>.
- Michael Wooldridge and Nicholas R. Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995. doi: 10.1017/S0269888900008122.

- Chenming Wu, Xiaofan Li, and Chengkai Dai. From prompts to printable models: Support-effective 3D generation via offset direct preference optimization. *arXiv preprint arXiv:2511.16434*, 2025a.
- Di Wu, Wasi Uddin Ahmad, Dejiao Zhang, Murali Krishna Ramanathan, and Xiaofei Ma. Repoformer: Selective retrieval for repository-level code completion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53270–53290. PMLR, 2024a. URL <https://proceedings.mlr.press/v235/wu24a.html>.
- Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL <https://openreview.net/forum?id=pZiyCaVuti>.
- Di Wu, Zixiang Ji, Asmi Kawatkar, Bryan Kwan, Jia-Chen Gu, Nanyun Peng, and Kai-Wei Chang. LongMemEval-V2: Evaluating long-term agent memory toward experienced colleagues. *arXiv preprint arXiv:2605.12493*, 2026a. doi: 10.48550/arXiv.2605.12493. URL <https://arxiv.org/abs/2605.12493>. Work in progress.
- Jing Wu, Daphne Barretto, Yiye Chen, Nicholas Gydé, Yanan Jian, Yuhang He, and Vibhav Vineet. Os-marathon: Benchmarking computer-use agents on long-horizon repetitive tasks, 2026b. URL <https://arxiv.org/abs/2601.20650>.
- Keyu Wu, Qianjin Yu, Manlin Mei, Ruiting Liu, Jun Wang, Kailai Zhang, and Yelun Bao. Tn-autorca: Benchmark construction and agentic framework for self-improving alarm-based root cause analysis in telecommunication networks, 2025c. URL <https://arxiv.org/abs/2507.18190>.
- Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12156–12163. IEEE, 2024b.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023. URL <https://arxiv.org/abs/2308.08155>.
- Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, et al. Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025d.
- Tong Wu, Jinhua Zhu, Wengang Zhou, and Houqiang Li. Resist: Rationale-enhanced and reward model-based end-to-end social influence dialogue system. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(6):1–23, 2025e.
- Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, Xiaobing Yu, Yu Zhong, Shangqi Deng, Ufaq Khan, Jianghao Wu, Xiaofeng Liu, Imran Razzak, Xiaojun Chang, and Yutong Xie. Selfai: A self-directed framework for long-horizon scientific discovery, 2026c. URL <https://arxiv.org/abs/2512.00403>.
- Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. From human memory to ai memory: A survey on memory mechanisms in the era of llms, 2025f. URL <https://arxiv.org/abs/2504.15965>.
- Yu Yvonne Wu, Yuwei Zhang, Hyungjun Yoon, Ting Dang, Dimitris Spathis, Tong Xia, Qiang Yang, Jing Han, Dong Ma, Sung-Ju Lee, and Cecilia Mascolo. Wearable foundation models should go beyond static encoders, 2026d. URL <https://arxiv.org/abs/2603.19564>.
- Yujiang Wu, Shanshan Zhong, Yubin Kim, and Chenyan Xiong. What generative search engines like and how to optimize web content cooperatively. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026e. URL <https://openreview.net/forum?id=K8EinVWtUB>.
- Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks, 2024c. URL <https://arxiv.org/abs/2307.02477>.
- Zhiyong Wu, Chengcheng Han, Zichen Ding, Zhenmin Weng, Zhoumianze Liu, Shunyu Yao, Tao Yu, and Lingpeng Kong. OS-Copilot: Towards Generalist Computer Agents with Self-Improvement, 2024d. URL <https://arxiv.org/abs/2402.07456>.

- Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph WaveNet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1907–1913, 2019. doi: 10.24963/ijcai.2019/264.
- Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Dingwen Yang, Chenyang Liao, Xin Guo, Wei He, Songyang Gao, Lu Chen, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. Agentgym: Evolving large language model-based agents across diverse environments. *arXiv preprint arXiv:2406.04151*, 2024. URL <https://arxiv.org/abs/2406.04151>.
- Lianghao Xia, Chao Huang, Yong Xu, Peng Dai, Liefeng Bo, Xiyue Zhang, and Tianyi Chen. Spatial-temporal sequential hypergraph network for crime prediction with dynamic multiplex relation learning. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.
- Peng Xia, Jinglu Wang, Yi Peng, Kaide Zeng, Xiangnan Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Shujie Liu, Yan Lu, and Huaxiu Yao. Mmedagent-rl: Optimizing multi-agent collaboration for multimodal medical reasoning. *ArXiv*, abs/2506.00555, 2025. URL <https://api.semanticscholar.org/CorpusID:279075781>.
- Rong Xiang. Structural enforcement of goal integrity in ai agents via separation-of-powers architecture, 2026. URL <https://arxiv.org/abs/2604.23646>.
- Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Carl Yang, Dawn Song, and Bo Li. Guardagent: Safeguard llm agents by a guard agent via knowledge-enabled reasoning, 2025. URL <https://arxiv.org/abs/2406.09187>.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024. URL <https://arxiv.org/abs/2309.17453>.
- Wenli Xiao, Jia Xie, Tonghe Zhang, Haotian Lin, Letian Fu, Haoru Xue, Jalen Lu, Yi Yang, Cunxi Dai, Zi Wang, et al. Empire: Agentic robot policy self-improvement in the real world. *arXiv preprint arXiv:2606.19980*, 2026.
- Annie Xie, Fahim Tajwar, Archit Sharma, and Chelsea Finn. When to ask for help: Proactive interventions in autonomous reinforcement learning, 2022. URL <https://arxiv.org/abs/2210.10765>.
- Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. TravelPlanner: A benchmark for real-world planning with language agents. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54590–54613. PMLR, 2024a. URL <https://proceedings.mlr.press/v235/xie24j.html>.
- Tengyang Xie, John Langford, Paul Mineiro, and Ida Momennejad. Interaction-grounded learning, 2021. URL <https://arxiv.org/abs/2106.04887>.
- Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, volume 37, pp. 52040–52094, 2024b. doi: 10.52202/079017-1650. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.
- Zhifei Xie, Zongzheng Hu, Fangda Ye, Xin Zhang, Haobo Chai, Zihang Liu, Pengcheng Wu, Guibin Zhang, Yue Liao, Xiaobin Hu, Deheng Ye, Chunyan Miao, and Shuicheng Yan. Pask: Toward intent-aware proactive agents with long-term memory, 2026. URL <https://arxiv.org/abs/2604.08000>.
- Eric Xing, Mingkai Deng, and Jinyu Hou. Critique of agent model, 2026. URL <https://arxiv.org/abs/2606.23991>.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms, 2024. URL <https://arxiv.org/abs/2306.13063>.
- Fengli Xu, Jun Zhang, Chen Gao, Jie Feng, and Yong Li. Urban generative intelligence (ugi): A foundational platform for agents in embodied city environment. *arXiv preprint arXiv:2312.11813*, 2023.

- Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models, 2024a. URL <https://arxiv.org/abs/2402.06044>.
- Haiyang Xu, Xi Zhang, Haowei Liu, Junyang Wang, Zhaozai Zhu, Shengjie Zhou, Xuhao Hu, Feiyu Gao, Junjie Cao, Zihua Wang, Zhiyuan Chen, Jitong Liao, Qi Zheng, Jiahui Zeng, Ze Xu, Shuai Bai, Junyang Lin, Jingren Zhou, and Ming Yan. Mobile-agent-v3.5: Multi-platform fundamental gui agents, 2026a. URL <https://arxiv.org/abs/2602.16855>.
- Haowen Xu, Femi Omitaomu, Soheil Sabri, Sisi Zlatanova, Xiao Li, and Yongze Song. Leveraging generative AI for urban digital twins: A scoping review on the autonomous generation of urban data, scenarios, designs, and 3D city models for smart city advancement. *arXiv preprint arXiv:2405.19464*, 2024b.
- Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. Penetrative AI: Making LLMs comprehend the physical world. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024c. doi: 10.18653/v1/2024.findings-acl.437.
- Jiexi Xu. Agentic metacognition: Designing a "self-aware" low-code agent for failure prediction and human handoff, 2025. URL <https://arxiv.org/abs/2509.19783>.
- Jundong Xu, Qingchuan Li, Jiaying Wu, Yihuai Lan, Shuyue Stella Li, Huichi Zhou, Bowen Jiang, Lei Wang, Jun Wang, Anh Tuan Luu, Caiming Xiong, Hae Won Park, Bryan Hooi, and Zhiyuan Hu. Evoarena: Tracking memory evolution for robust llm agents in dynamic environments. *arXiv preprint arXiv:2606.13681*, 2026b. URL <https://arxiv.org/abs/2606.13681>.
- Ke Xu, Yuhao Wang, and Yu Wang. From reactive to proactive: Assessing the proactivity of voice agents via provoice-bench, 2026c. URL <https://arxiv.org/abs/2604.15037>.
- Leo Xu, Letian Li, Alex Cuellar, and Michael Hagenow. Learning to assist: Collaborative vlas for implicit human-robot collaboration. *arXiv preprint arXiv:2606.12475*, 2026d.
- Lily Xu, Shahrzad Gholami, Sara McCarthy, Bistra Dilkina, Andrew Plumptre, Milind Tambe, Rohit Singh, Mustapha Nsubuga, Joshua Mabonga, Margaret Driciru, Fred Wanyama, Aggrey Rwetsiba, Tom Okello, and Eric Enyel. Stay ahead of poachers: Illegal wildlife poaching prediction and patrol planning under uncertainty with field test evaluations (short version). In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1898–1901. IEEE, April 2020. doi: 10.1109/icde48307.2020.00198. URL <http://dx.doi.org/10.1109/icde48307.2020.00198>.
- Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation. *arXiv preprint arXiv:2505.21864*, 2025a.
- Michael F. Xu, Enhui Zhao, Yawen Zhang, Joseph E. Michaelis, Sarah Sebo, and Bilge Mutlu. Designing robots for families: In-situ prototyping for contextual reminders on family routines. *arXiv preprint arXiv:2602.22628*, 2026e.
- Songlin Xu, Hao-Ning Wen, Hongyi Pan, Dallas Dominguez, Dongyin Hu, and Xinyu Zhang. Classroom simulacra: Building contextual student generative agents in online education for learning behavioral simulation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–26, 2025b.
- Tianlong Xu, YiFan Zhang, Zhendong Chu, Shen Wang, and Qingsong Wen. AI-driven virtual teacher for enhanced educational efficiency: Leveraging large pretrain models for autonomous error analysis and correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(28):28801–28809, 2025c. doi: 10.1609/aaai.v39i28.35144. URL <https://doi.org/10.1609/aaai.v39i28.35144>.
- Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents, 2025d. URL <https://arxiv.org/abs/2502.12110>.
- Yifan Xu, Xiao Liu, Xinghan Liu, Jiaqi Fu, Hanchen Zhang, Bohao Jing, Shudan Zhang, Yuting Wang, Wenyi Zhao, and Yuxiao Dong. Mobilerl: Online agentic reinforcement learning for mobile gui agents, 2025e. URL <https://arxiv.org/abs/2509.18119>.
- Yiqing Xu, Linfeng Li, Cunjun Yu, and David Hsu. "stack it up!": 3d stable structure generation from 2d hand-drawn sketch. *arXiv preprint arXiv:2508.02093*, 2025f.

- Yuqiao Xu and Erman Ayday. Seeing less is not seeing safely: Privacy leakage from task-scoped robot perception exports. *arXiv preprint arXiv:2609.03055*, 2026.
- Zhe Xu, Zhixin Li, Qingwen Guan, Dingshui Zhang, Qiang Li, Junxiao Nan, Chunyang Liu, Wei Bian, and Jieping Ye. Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 905–913, 2018. doi: 10.1145/3219819.3219824.
- Zhengsen Xu, Jonathan Li, Sibao Cheng, Xue Rui, Yu Zhao, Hongjie He, Haiyan Guan, Aryan Sharma, Matthew Erxleben, Ryan Chang, and Linlin Xu. Deep learning for wildfire risk prediction: Integrating remote sensing and environmental data. *arXiv preprint arXiv:2405.01607*, 2024d.
- Ziyi Xuan, Yiwen Wu, Zhaoyang Yan, Vinod Nambodiri, and Yu Yang. After talking with 1,000 personas: Learning preference-aligned proactive assistants from large-scale persona interactions, 2026. URL <https://arxiv.org/abs/2602.04000>.
- Jintao Xue, Xiao Li, and Nianmin Zhang. Safe reinforcement learning with online filtering for fatigue-predictive human-robot task planning and allocation in production. *arXiv preprint arXiv:2604.12667*, 2026.
- Adina Yakefu, Bin Xie, Chongyang Xu, Enwen Zhang, Erjin Zhou, Fan Jia, Haitao Yang, Haoqiang Fan, Haowei Zhang, Hongyang Peng, et al. Robochallenge: Large-scale real-robot evaluation of embodied policies. *arXiv preprint arXiv:2510.17950*, 2025.
- Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.
- Weicai Yan, Yuhong Dai, Qi Ran, Haodong Li, Wang Lin, Tao Jin, Xing Xie, Hao Liao, and Jianxun Lian. Proact-vl: A proactive videollm for real-time ai companions, 2026a. URL <https://arxiv.org/abs/2603.03447>.
- Xiaokai Yan, Jingtao Ding, Yong Li, and Zhiwen Yu. Act2Intention: A Benchmark For Developing Active Mobile Agents Through Inferring User Intention from GUI Actions, 2026b. URL <https://arxiv.org/abs/2608.14132>.
- Xuerun Yan, Zhexi Lian, Nuoheng Zhang, Shiyu Fang, Haoran Wang, Chen Lv, Jia Hu, and Binyang Song. Wcog-vla: A dual-level world-cognitive vision-language-action model for end-to-end autonomous driving. *arXiv preprint arXiv:2607.08375*, 2026c.
- Yibo Yan, Shen Wang, Jiahao Huo, Philip S Yu, Xuming Hu, and Qingsong Wen. Mathagent: Leveraging a mixture-of-math-agent framework for real-world multimodal mathematical error detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 69–82, 2025a.
- Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yifan Zhang, Tianlong Xu, Zhendong Chu, et al. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24297–24334, 2026d.
- Yunhe Yan, Shihe Wang, Jiajun Du, Yexuan Yang, Yuxuan Shan, Qichen Qiu, Xianqing Jia, Xinge Wang, Xin Yuan, Xu Han, Mao Qin, Yinxiao Chen, Chen Peng, Shangguang Wang, and Mengwei Xu. Mcpworld: A unified benchmarking testbed for api, gui, and hybrid computer use agents, 2025b. URL <https://arxiv.org/abs/2506.07672>.
- Yuwei Yan, Qingbin Zeng, Zhiheng Zheng, Jingzhe Yuan, Jie Feng, Jun Zhang, Fengli Xu, and Yong Li. Opencity: A scalable platform to simulate urban activities with massive llm agents. *arXiv preprint arXiv:2410.21286*, 2024.
- Kevin P. Yancey and Burr Settles. A sleeping, recovering bandit algorithm for optimizing recurring notifications. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20*, pp. 3008–3016. ACM, August 2020. doi: 10.1145/3394486.3403351. URL <http://dx.doi.org/10.1145/3394486.3403351>.
- Bufang Yang, Lilin Xu, Liekang Zeng, Yunqi Guo, Siyang Jiang, Wenrui Lu, Kaiwei Liu, Hancheng Xiang, Xiaofan Jiang, Guoliang Xing, and Zhenyu Yan. Proagent: Harnessing on-demand sensory contexts for proactive llm agent systems, 2025a. URL <https://arxiv.org/abs/2512.06721>.

- Bufang Yang, Lilin Xu, Liekang Zeng, Kaiwei Liu, Siyang Jiang, Wenrui Lu, Hongkai Chen, Xiaofan Jiang, Guoliang Xing, and Zhenyu Yan. Contextagent: Context-aware proactive llm agents with open-world sensory perceptions, 2025b. URL <https://arxiv.org/abs/2505.14668>.
- Bufang Yang, Lilin Xu, Liekang Zeng, Yunqi Guo, Siyang Jiang, Wenrui Lu, Kaiwei Liu, Yixuan Li, Xiaofan Jiang, Guoliang Xing, and Zhenyu Yan. Proagent: Harnessing on-demand sensory contexts for proactive llm agent systems in the wild, 2026a. URL <https://arxiv.org/abs/2512.06721>.
- Guang-Zhong Yang, James Cambias, Kevin Cleary, Eric Daimler, James Drake, Pierre E. Dupont, Nobuhiko Hata, Peter Kazanzides, Sylvain Martel, Rajni V. Patel, Veronica J. Santos, and Russell H. Taylor. Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy. *Science Robotics*, 2(4), March 2017. ISSN 2470-9476. doi: 10.1126/scirobotics.aam8638. URL <http://dx.doi.org/10.1126/scirobotics.aam8638>.
- Jiarui Yang, Jiajin Zhang, Bin Zhu, Jingjing Chen, and Yu-Gang Jiang. Stable and efficient real-world online vla post-training via asynchronous replay-anchored policy improvement. *arXiv preprint arXiv:2609.22888*, 2026b.
- John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 50528–50652. Curran Associates, Inc., 2024a. doi: 10.52202/079017-1601. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5a7c947568c1b1328ccc5230172e1e7c-Abstract-Conference.html.
- Kaiqi Yang, Yucheng Chu, Taylor Darwin, Ahreum Han, Hang Li, Hongzhi Wen, Yasemin Copur-Gencturk, Jiliang Tang, and Hui Liu. Content knowledge identification with multi-agent large language models (llms). In *International conference on artificial intelligence in education*, pp. 284–292. Springer, 2024b.
- Ke Yang and ChengXiang Zhai. Ten principles of ai agent economics, 2025. URL <https://arxiv.org/abs/2505.20273>.
- Ke Yang, Yao Liu, Sapana Chaudhary, Rasool Fakoor, Pratik A Chaudhari, George Karypis, and Huzefa Rangwala. Agentoccam: A simple yet strong baseline for llm-based web agents. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 97533–97565, 2025c. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f2c6e459b95694a24ac69c469a4ee746-Paper-Conference.pdf.
- Ke Yang, Kevin Ros, Shankar Kumar Senthil Kumar, and ChengXiang Zhai. Jir-arena: The first benchmark dataset for just-in-time information recommendation, 2025d. URL <https://arxiv.org/abs/2505.13550>.
- Ke Yang, Zixi Chen, Xuan He, Jize Jiang, Michel Galley, Chenglong Wang, Jianfeng Gao, Jiawei Han, and ChengXiang Zhai. Plugmem: A task-agnostic plugin memory module for llm agents, 2026c. URL <https://arxiv.org/abs/2603.03296>.
- Ke Yang, Chenglong Wang, Michel Galley, Chandan Singh, Jeevana Priya Inala, ChengXiang Zhai, and Jianfeng Gao. Studentsim: Training llm-based student simulators. *arXiv preprint arXiv:2609.01591*, 2026d.
- Qianlan Yang, Xiangjun Wang, Danielle Perszyk, and Yu-Xiong Wang. Self-guided hierarchical exploration for generalist foundation model web agents. In *Advances in Neural Information Processing Systems*, volume 38, 2025e. doi: 10.52202/085713-0944.
- Qinglong Yang, Haoming Li, Haotian Zhao, Xiaokai Yan, Jingtao Ding, Fengli Xu, and Yong Li. Fingertip 20k: A benchmark for proactive and personalized mobile llm agents, 2026e. URL <https://arxiv.org/abs/2507.21071>.
- Xiliu Yang, Samuel Slezak, Lasath Siriwardena, Felix Amtsberg, and Achim Menges. MAVE: An augmented multi-agent LLM system for interactive design and robotic fabrication. In *Proceedings of the 2026 ACM Designing Interactive Systems Conference (DIS '26)*, 2026f. doi: 10.1145/3800645.3813008.
- Yuhao Yang, Tianyu Fan, and Chao Huang. CLI-Anything: Towards agent-native computer use, 2026g. URL <https://arxiv.org/abs/2606.03854>.

- Zhutian Yang, Caelan Garrett, Dieter Fox, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Guiding long-horizon task and motion planning with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16847–16853. IEEE, 2025f.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023a. URL <https://arxiv.org/abs/2305.10601>.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=WE_vluYUL-X.
- Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*, pp. 351–360, 2017. doi: 10.1145/3038912.3052577.
- Yihang Yao, Zhepeng Cen, Haohong Lin, Shiqi Liu, Zuxin Liu, Jiacheng Zhu, Zhang-Wei Hong, Laixi Shi, and Ding Zhao. Pushing forward pareto frontiers of proactive agents with behavioral agentic optimization, 2026. URL <https://arxiv.org/abs/2602.11351>.
- Yarbo Inc. *Yarbo Snow Blower: Autonomous Snow Clearing Robot*. Yarbo Inc., 2026. URL <https://www.yarbo.com/products/yarbo-snow-blower>. Product specifications and technical capabilities for the 2-stage autonomous snow removal system utilizing RTK-GPS and stereo vision.
- Chongrui Ye, Yuxiang Liu, Yu Wang, Haoifei Yu, Yining Zhao, Ge Liu, Julian McAuley, and Jiaxuan You. Auto-dreamer: Learning offline memory consolidation for language agents, 2026a. URL <https://arxiv.org/abs/2605.20616>.
- Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026b.
- Suyu Ye, Haojun Shi, Darren Shih, Hyokun Yun, Tanya G. Roosta, and Tianmin Shu. RealWebAssist: A benchmark for long-horizon web assistance with real-world users. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (40):34441–34449, 2026c. doi: 10.1609/aaai.v40i40.40742.
- Ziruo Yi, Jinyu Liu, Ting Xiao, and Mark V. Albert. A multi-agent system for complex reasoning in radiology visual question answering. *2025 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 139–147, 2025. URL <https://api.semanticscholar.org/CorpusID:280527311>.
- Fan Yin, Zifeng Wang, I-Hung Hsu, Jun Yan, Ke Jiang, Yanfei Chen, Jindong Gu, Long T. Le, Kai-Wei Chang, Chen-Yu Lee, Hamid Palangi, and Tomas Pfister. Magnet: Multi-turn tool-use data synthesis and distillation via graph translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32600–32616, 2025. doi: 10.18653/v1/2025.acl-long.1566. URL <https://aclanthology.org/2025.acl-long.1566/>.
- Kai Yin, Bo Li, Chengkai Liu, Ali Mostafavi, and Xia Hu. CrisisSense-LLM: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics. *arXiv preprint arXiv:2406.15477*, 2024a.
- Lianhao Yin, Makram Chahine, Tsun-Hsuan Wang, Tim Seyde, Chao Liu, Mathias Lechner, Ramin Hasani, and Daniela Rus. Towards cooperative flight control using visual-attention, 2023. URL <https://arxiv.org/abs/2212.11084>.
- Sheng Yin, Xianghe Pang, Yuanzhuo Ding, Menglan Chen, Yutong Bi, Yichen Xiong, Wenhao Huang, Zhen Xiang, Jing Shao, and Siheng Chen. SafeAgentBench: A benchmark for safe task planning of embodied LLM agents. *arXiv preprint arXiv:2412.13178*, 2024b.
- Zhuowen Yin, Chongyang Liu, Wenzhang Yang, Renjue Li, and Yinxing Xue. ROBOSHACKLES: A safety dataset for human-injury prevention in embodied foundation models. *arXiv preprint arXiv:2606.18632*, 2026.
- Derek Yohn, Luke Flancher, Mirajul Islam, and Khaled Slhoub. Can open-source LLM agents replace static application security testing tools? an empirical assessment, 2026. URL <https://arxiv.org/abs/2606.11672>.

- Uksang Yoo, Adam Hung, Jonathan Francis, Jean Oh, and Jeffrey Ichnowski. Ropotter: Toward robotic pottery and deformable object manipulation with structural priors. In *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*, pp. 843–850. IEEE, 2024.
- Justin Young. Effective harnesses for long-running agents. <https://www.anthropic.com/engineering/effective-harnesses-for-long-running-agents>, November 2025. Anthropic Engineering.
- YouTube. On youtube’s recommendation system. <https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/>, 2021.
- YouTube. How creators use ai for content creation. <https://www.youtube.com/howyoutubeworks/ai/>, 2026.
- Albert Yu, Chengshu Li, Luca Macesanu, Arnav Balaji, Ruchira Ray, Raymond Mooney, and Roberto Martín-Martín. Mixed-initiative dialog for human-robot collaborative manipulation, 2026a. URL <https://arxiv.org/abs/2508.05535>.
- Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3634–3640, 2018. doi: 10.24963/ijcai.2018/505.
- Chaerin Yu, Chihun Choi, Sunjae Lee, Hyosu Kim, Steven Y. Ko, Young-Bae Ko, and Sangeun Oh. Leveraging llms for efficient and personalized smart home automation. *arXiv preprint arXiv:2601.04680*, 2026b.
- Che Rin Yu, Daewon Chae, Dabin Seo, Sangwon Lee, Hyeongwoo Im, and Jinkyu Kim. Scene graph-guided proactive replanning for failure-resilient embodied agent, 2025. URL <https://arxiv.org/abs/2508.11286>.
- Jie Yu, Xinyi Zhang, Yaqiong Luo, Ray LC, Kaishun Wu, and Xin Tong. From tool to partner: Expressive behaviors as the bridge to human-robot creative collaboration. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*, 2026c. doi: 10.1145/3772318.3790270.
- Liang Yu, Weiwei Xie, Di Xie, Yulong Zou, Dengyin Zhang, Zhixin Sun, Linghua Zhang, Yue Zhang, and Tao Jiang. Deep reinforcement learning for smart home energy management. *IEEE Internet of Things Journal*, 7(4):2751–2762, 2020. doi: 10.1109/IIOT.2019.2957289.
- Peiqi Yu, Abulikemu Abuduweili, Ruixuan Liu, and Changliu Liu. Robustifying long-term human-robot collaboration through a multimodal and hierarchical framework. *arXiv preprint arXiv:2411.15711*, 2024.
- Peiqi Yu, Philip Huang, Chaitanya Chawla, Guanya Shi, Jiaoyang Li, and Changliu Liu. Autonomous integration and improvement of robotic assembly using skill graph representations. *arXiv preprint arXiv:2603.12649*, 2026d.
- Ryan Yu, Pushi Zhang, Starrick Liu, Brae Liu, Miracle Kang, Shalfun Li, Lights Shi, Ellie Ma, Ping Yang, Chris Pan, et al. Wall-oss-0.5 technical report. *arXiv preprint arXiv:2605.30877*, 2026e.
- Xinle Yu, Fan Bai, Kaiser Sun, Hengshuo Miao, Abhay Anand, Zhongyan Luo, Kun Zhou, and Zhen Wang. PrimeScientist: Strategic allocation of research effort in autonomous research, 2026f. URL <https://arxiv.org/abs/2609.17846>.
- Bo Yuan, Zelin Zhao, Petr Molodyk, Bin Hu, and Yongxin Chen. Clarify before you draw: Proactive agents for robust text-to-cad generation, 2026a. URL <https://arxiv.org/abs/2602.03045>.
- Mengqi Yuan, Zilong Zhou, Xinzhuang Xiong, Weiming Wu, Jiayang Sun, Jiamin Song, Kaiqian Cui, Bowen Wang, Haoyuan Wu, Yitong Li, Dunjie Lu, Haikong Lu, Qi Zhen, Xinyuan Wang, Jiaqi Deng, Yuhao Yang, Cheng Chen, Boyuan Zheng, Alex Su, Xiao Yu, Hao Zou, Saaket Agashe, Xing Han Lu, Manpreet Kaur, Zhengyang Qi, Vincent Sunn Chen, Frederic Sala, Dayiheng Liu, Junyang Lin, Zhou Yu, Yu Su, Siva Reddy, Xin Eric Wang, Peng Qi, Tianbao Xie, and Tao Yu. OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks, 2026b. URL <https://arxiv.org/abs/2606.29537>.
- Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. R-judge: Benchmarking safety risk awareness for llm agents, 2024. URL <https://arxiv.org/abs/2401.10019>.
- Shengbin Yue, Ting Huang, Zheng Jia, Siyuan Wang, Shujun Liu, Yun Song, Xuanjing Huang, and Zhongyu Wei. Multi-agent simulator drives language models for legal intensive interaction. In Luis Chiruzzo, Alan Ritter, and Lu Wang

- (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6552–6585, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.365. URL <https://aclanthology.org/2025.findings-naacl.365/>.
- Setareh Zafari and Sabine T. Koeszegi. Attitudes toward attributed agency: Role of perceived control. *International Journal of Social Robotics*, 13(8):2071–2080, July 2020. ISSN 1875-4805. doi: 10.1007/s12369-020-00672-7. URL <http://dx.doi.org/10.1007/s12369-020-00672-7>.
- Kevin Zakka, Baruch Tabanpour, Qiayuan Liao, Mustafa Haiderbhai, Samuel Holt, Jing Yuan Luo, Arthur Allshire, Erik Frey, Koushil Sreenath, Lueder A Kahrs, et al. Mujoco playground. *arXiv preprint arXiv:2502.08844*, 2025.
- Hongzhi Zang, Mingjie Wei, Si Xu, Yongji Wu, Zhen Guo, Yuanqing Wang, Hao Lin, Peihong Wang, Liangzhi Shi, Yuqing Xie, et al. Rlinf-vla: A unified and efficient framework for reinforcement learning of vision-language-action models. *arXiv preprint arXiv:2510.06710*, 2025. Accepted to RSS 2026.
- Xinshi Zang, Huaxiu Yao, Guanjie Zheng, Nan Xu, Kai Xu, and Zhenhui Li. MetaLight: Value-based meta-reinforcement learning for traffic signal control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 1153–1160, 2020. doi: 10.1609/aaai.v34i01.5467.
- Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Voelkel, Johannes Schöning, Rainer Malaka, and Yvonne Rogers. Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces (CUI)*, pp. 3:1–3:14, 2022. doi: 10.1145/3543829.3543834.
- Sharare Zehtabian, Siavash Khodadadeh, Ladislau Bölöni, and Damla Turgut. Privacy-preserving learning of human activity predictors in smart environments, 2021. URL <https://arxiv.org/abs/2101.06564>.
- Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10471–10506, 2024. URL <https://aclanthology.org/2024.findings-acl.624/>.
- Andy K. Zhang, Neil Perry, Riya Dulepet, Joey Ji, Celeste Menders, Justin W. Lin, Eliot Jones, Gashon Hussein, Samantha Liu, Donovan Jasper, Pura Peetathawatchai, Ari Glenn, Vikram Sivashankar, Daniel Zamoshchin, Leo Glikbarg, Derek Askaryar, Mike Yang, Teddy Zhang, Rishi Alluri, Nathan Tran, Rinnara Sangpisit, Polycarpus Yiorkadjis, Kenny Osele, Gautham Raghupathi, Dan Boneh, Daniel E. Ho, and Percy Liang. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL <https://arxiv.org/abs/2408.08926>.
- Barry Zhang, Keith Lazuka, and Mahesh Murag. Equipping agents for the real world with agent skills. <https://www.anthropic.com/engineering/equipping-agents-for-the-real-world-with-agent-skills>, October 2025b. Anthropic Engineering.
- Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, Xiaojun Chang, Junge Zhang, Feng Yin, Yitao Liang, and Yaodong Yang. Proagent: Building proactive cooperative agents with large language models, 2024a. URL <https://arxiv.org/abs/2308.11339>.
- Chaoyun Zhang, He Huang, Chiming Ni, Jian Mu, Si Qin, Shilin He, Lu Wang, Fangkai Yang, Pu Zhao, Chao Du, Liqun Li, Yu Kang, Zhao Jiang, Suzhen Zheng, Rujia Wang, Jiaxu Qian, Minghua Ma, Jian-Guang Lou, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. UFO2: The Desktop AgentOS, 2025c. URL <https://arxiv.org/abs/2504.14603>.
- Fengji Zhang, Bei Chen, Yue Zhang, Jacky Keung, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou, and Weizhu Chen. RepoCoder: Repository-level code completion through iterative retrieval and generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2471–2484, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.151. URL <https://aclanthology.org/2023.emnlp-main.151/>.
- Fulei Zhang and Zhou Yu. Mind the gap: Linguistic divergence and adaptation strategies in human-llm assistant vs. human-human interactions, 2025. URL <https://arxiv.org/abs/2510.02645>.

- Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say ‘i don’t know’, 2024b. URL <https://arxiv.org/abs/2311.09677>.
- Haobo Zhang, Yutao Zhu, Kelong Mao, Tianhao Li, and Zhicheng Dou. Recthinker: An agentic framework for tool-augmented reasoning in recommendation, 2026a. URL <https://arxiv.org/abs/2603.09843>.
- Haoran Zhang, Luxin Xu, Zhilin Wang, Runquan Gui, Shunkai Zhang, Haodi Lei, Zihao He, Bingsu He, Chicheng Qin, Tong Zhu, Xiaoye Qu, Yang Yang, Yu Cheng, and Yafu Li. π -bench: Evaluating proactive personal assistant agents in long-horizon workflows, 2026b. URL <https://arxiv.org/abs/2605.14678>.
- Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-VAU: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13843–13853, June 2025d. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Holmes-VAU_Towards_Long-term_Video_Anomaly_Understanding_at_Any_Granularity_CVPR_2025_paper.html.
- Jenny Zhang, Joel Lehman, Kenneth Stanley, and Jeff Clune. Omni: Open-endedness via models of human notions of interestingness, 2024c. URL <https://arxiv.org/abs/2306.01711>.
- Jingwei Zhang, Thomas Lampe, Abbas Abdolmaleki, Jost Tobias Springenberg, and Martin Riedmiller. Game on: Towards language models as rl experimenters, 2024d. URL <https://arxiv.org/abs/2409.03402>.
- Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1655–1661, 2017.
- Kaiyuan Zhang, Jiaqi Li, Yueyue Wu, Haitao Li, Cheng Luo, Shaokun Zou, Yujia Zhou, Weihang Su, Yiqun Liu, and Qingyao Ai. Chinese court simulation with LLM-based agents system. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8429–8454, San Diego, California, United States, July 2026c. Association for Computational Linguistics. ISBN 979-8-89176-395-1. URL <https://aclanthology.org/2026.findings-acl.411/>.
- Lefan Zhang, Weijia He, Jesse Martinez, Noah Brackenburg, Shan Lu, and Blase Ur. AutoTap: Synthesizing and repairing trigger-action programs using LTL properties. In *Proceedings of the 41st International Conference on Software Engineering (ICSE)*, pp. 281–291, 2019. doi: 10.1109/ICSE.2019.00043.
- Lefan Zhang, Weijia He, Olivia Morkved, Valerie Zhao, Michael L. Littman, Shan Lu, and Blase Ur. Trace2tap: Synthesizing trigger-action programs from traces of behavior. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(3), September 2020a. doi: 10.1145/3411838. URL <https://doi.org/10.1145/3411838>.
- Li Zhang, Longxi Gao, and Mengwei Xu. Does chain-of-thought reasoning help mobile GUI agent? an empirical study. *arXiv preprint arXiv:2503.16788*, 2025e. doi: 10.48550/arXiv.2503.16788.
- Lingzhe Zhang, Tong Jia, Mengxi Jia, Yifan Wu, Aiwei Liu, Yong Yang, Zhonghai Wu, Xuming Hu, Philip S. Yu, and Ying Li. A survey of aiops for failure management in the era of large language models, 2024e. URL <https://arxiv.org/abs/2406.11213>.
- Michael J. Q. Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with LMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5541–5558, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.306. URL <https://aclanthology.org/2025.findings-naacl.306/>.
- Michael J. Q. Zhang, W. Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach llms to ask clarifying questions, 2025f. URL <https://arxiv.org/abs/2410.13788>.
- Mike Zhang, Amalie Pernille Dilling, Léon Gondelman, Niels Erik Ruan Lyngdorf, Euan D. Lindsay, and Johannes Bjerva. Self: A framework for generating synthetic educational assignment feedback with llm agents, 2026d. URL <https://arxiv.org/abs/2502.12927>.
- Shao Zhang, Xihuai Wang, Wenhao Zhang, Chaoran Li, Junru Song, Tingyu Li, Lin Qiu, Xuezhi Cao, Xunliang Cai, Wen Yao, Weinan Zhang, Xinbing Wang, and Ying Wen. Leveraging dual process theory in language agent framework for real-time simultaneous human-ai collaboration, 2025g. URL <https://arxiv.org/abs/2502.11882>.

- Shibo Zhang, Yaxuan Li, Shen Zhang, Farzad Shahabi, Stephen Xia, Yu Deng, and Nabil Alshurafa. Deep learning in human activity recognition with wearable sensors: A review on advances, 2022. URL <https://arxiv.org/abs/2111.00418>.
- Shiliang Zhang and Sabita Maharjan. Security, privacy, and agentic ai in a regulatory view: From definitions and distinctions to provisions and reflections, 2026. URL <https://arxiv.org/abs/2603.18914>.
- Siyao Zhang, Daocheng Fu, Wenzhe Liang, Zhao Zhang, Bin Yu, Pinlong Cai, and Baozhen Yao. Trafficgpt: Viewing, processing and interacting with traffic foundation models. *Transport Policy*, 150:95–105, 2024f. ISSN 0967-070X. doi: <https://doi.org/10.1016/j.tranpol.2024.03.006>. URL <https://www.sciencedirect.com/science/article/pii/S0967070X24000726>.
- Taiyu Zhang, Xinnian Zhao, and Adalberto L Simeone. " what can i do for you": How should ai companions provide assistance to players in virtual reality games. *arXiv preprint arXiv:2609.13727*, 2026e.
- Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10746–10766, Bangkok, Thailand, August 2024g. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.578. URL <https://aclanthology.org/2024.acl-long.578/>.
- Weijia Zhang, Jindong Han, Zhao Xu, Hang Ni, Tengfei Lyu, Hao Liu, and Hui Xiong. Towards urban general intelligence: A review and outlook of urban foundation models, 2026f. URL <https://arxiv.org/abs/2402.01749>.
- Xiangyue Zhang. Deep researcher agent: An autonomous framework for 24/7 deep learning experimentation with zero-cost monitoring, 2026a. URL <https://arxiv.org/abs/2604.05854>.
- Xinrong Zhang, Yingfa Chen, Shengding Hu, Xu Han, Zihang Xu, Yuanwei Xu, Weilin Zhao, Maosong Sun, and Zhiyuan Liu. Beyond the turn-based game: Enabling real-time conversations with duplex models, 2026g. URL <https://arxiv.org/abs/2406.15718>.
- Xuan Zhang, Yang Deng, Zifeng Ren, See-Kiong Ng, and Tat-Seng Chua. Ask-before-plan: Proactive language agents for real-world planning, 2024h. URL <https://arxiv.org/abs/2406.12639>.
- Yao Zhang, Zijian Ma, Yunpu Ma, Zhen Han, Yu Wu, and Volker Tresp. WebPilot: A Versatile and Autonomous Multi-Agent System for Web Task Execution with Strategic Exploration, 2024i. URL <https://arxiv.org/abs/2408.15978>.
- Yi-Fan Zhang, Hang Li, Dingjie Song, Lichao Sun, Tianlong Xu, and Qingsong Wen. From correctness to comprehension: Ai agents for personalized error diagnosis in education. *arXiv preprint arXiv:2502.13789*, 2025h.
- Yifan Zhang, Yunheng Zou, Shaokun Zhang, Jian Hu, Hao Zhang, Binfeng Xu, Jan Kautz, and Yi Dong. Agora: Git as shared memory for collective autoresearch, 2026h. URL <https://arxiv.org/abs/2609.18094>.
- Yilei Zhang. Agent manufacturing: Foundation-model agents as first-class industrial entities. *arXiv preprint arXiv:2605.24823*, 2026b.
- Yuedi Zhang, Shuanghao Bai, Wanqi Zhou, Haoran Zhang, Qi Zhang, Zhirong Luan, and Badong Chen. Assistance without interruption: A benchmark and LLM-based framework for non-intrusive human-robot assistance. *arXiv preprint arXiv:2605.01368*, 2026i.
- Yuhao Zhang, Borong Zhang, Jiaming Fan, Jiachen Shen, Yishuai Cai, Yaodong Yang, and Jiaming Ji. RedVLA: Physical red teaming for vision-language-action models. *arXiv preprint arXiv:2604.22591*, 2026j.
- Yuji Zhang, Duo Zhou, Bo Chen, Aditya Chalasani, Noah Schroeder, H Chad Lane, and ChengXiang Zhai. Scaffolding human learning by shaping visual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 10922–10931, June 2026k.
- Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and*

- Transparency*, pp. 295–305. Association for Computing Machinery, January 2020b. doi: 10.1145/3351095.3372852. URL <https://doi.org/10.1145/3351095.3372852>.
- Yuntong Zhang, Haifeng Ruan, Zhiyu Fan, and Abhik Roychoudhury. AutoCodeRover: Autonomous program improvement. In *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis*, pp. 1592–1604, New York, NY, USA, 2024j. Association for Computing Machinery. doi: 10.1145/3650212.3680384. URL <https://doi.org/10.1145/3650212.3680384>.
- Zeyi Zhang, Zixi Kang, Ruijie Zhao, Yusen Feng, Biao Jiang, and Libin Liu. Proact: A dual-system framework for proactive embodied social agents, 2026l. URL <https://arxiv.org/abs/2602.14048>.
- Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents, 2024k. URL <https://arxiv.org/abs/2404.13501>.
- Zhao-Yu Zhang, Qingying Chen, Chunyuan Zheng, Jing Zhou, Jian Sun, Siqi Chen, Leiying Chen, Chuan Zhou, Huiyou Jiang, Xin Tao, et al. A self-triggered agentic push recommendation system. *arXiv preprint arXiv:2608.01949*, 2026m.
- Zhong Zhang, Yaxi Lu, Yikun Fu, Yupeng Huo, Shenzhi Yang, Yesai Wu, Han Si, Xin Cong, Haotian Chen, Yankai Lin, Jie Xie, Wei Zhou, Wang Xu, Yuanheng Zhang, Zhou Su, Zhongwu Zhai, Xiaoming Liu, Yudong Mei, Jianming Xu, Hongyan Tian, Chongyi Wang, Chi Chen, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Agentcpm-gui: Building mobile-use agents with reinforcement fine-tuning, 2025i. URL <https://arxiv.org/abs/2506.01391>.
- Dora Zhao, Michelle S. Lam, Diyi Yang, and Michael S. Bernstein. Behavior latticing: Inferring user motivations from unstructured interactions, 2026a. URL <https://arxiv.org/abs/2604.07629>.
- Ruixiang Zhao, Jie Yang, Zijie Xin, Tianyi Wang, Fengyun Rao, Jing Lyu, and Xirong Li. OmniPro: A comprehensive benchmark for omni-proactive streaming video understanding, 2026b. URL <https://arxiv.org/abs/2605.18577>.
- Sebastian Zhao, Alan Zhu, Hussein Mozannar, David Sontag, Ameet Talwalkar, and Valerie Chen. Codinggenie: A proactive llm-powered programming assistant, 2025a. URL <https://arxiv.org/abs/2503.14724>.
- Shiyan Zhao, Xinwei Song, Tianyu Guo, Xiaomeng Gao, Mingyuan Liu, Xu Han, Yuanyuan Zhang, Zhenliang Zhang, Xue Feng, and Bo Dai. Normact: Benchmarking embodied agents’ proactive compliance with unspoken social norms. *arXiv preprint arXiv:2606.27826*, 2026c.
- Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.
- Yuheng Zhao, Xueli Shu, Liwen Fan, Lin Gao, Yu Zhang, and Siming Chen. Proactiveva: Proactive visual analytics with llm-based ui agent, 2025b. URL <https://arxiv.org/abs/2507.18165>.
- Yuyang Zhao, Wentao Shi, Fuli Feng, and Xiangnan He. AppAgent-Pro: A proactive GUI agent system for multidomain information integration and user assistance. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM ’25*, pp. 6767–6771. ACM, November 2025c. doi: 10.1145/3746252.3761473. URL <https://doi.org/10.1145/3746252.3761473>.
- Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B. Cohen, and Emine Yilmaz. PersonaLens: A benchmark for personalization evaluation in conversational AI assistants. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18023–18055, Vienna, Austria, 2025d. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.927. URL <https://aclanthology.org/2025.findings-acl.927/>.
- Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025e.
- Boyuan Zheng, Michael Y. Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, and Yu Su. Skillweaver: Web agents can self-improve by discovering and honing skills, 2025a. URL <https://arxiv.org/abs/2504.07079>.

- Hongjie Zheng, Zesheng Shi, and Ping Yi. Medcoact: Confidence-aware multi-agent collaboration for complete clinical decision. *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 4525–4528, 2025b. URL <https://api.semanticscholar.org/CorpusID:282058197>.
- Hongwei Zheng, Weiqi Wu, Zhengjia Wang, Guanyu Jiang, Haoming Li, Tianyu Wu, Yongchun Zhu, Jingwu Chen, and Feng Zhang. Icebreaker for conversational agents: Breaking the first-message barrier with personalized starters, 2026a. URL <https://arxiv.org/abs/2604.18375>.
- Huaixiu Steven Zheng, Swaroop Mishra, Hugh Zhang, Xinyun Chen, Minmin Chen, Azade Nova, Le Hou, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, and Denny Zhou. NATURAL PLAN: Benchmarking LLMs on natural language planning. *arXiv preprint arXiv:2406.04520*, 2024a. doi: 10.48550/arXiv.2406.04520. URL <https://arxiv.org/abs/2406.04520>.
- Longwei Zheng, Fei Jiang, Xiaoqing Gu, Yuanyuan Li, Gong Wang, and Haomin Zhang. Teaching via llm-enhanced simulations: Authenticity and barriers to suspension of disbelief. *Internet High. Educ.*, 65:100990, 2024b. URL <https://api.semanticscholar.org/CorpusID:274695185>.
- Mingwei Zheng, Chengpeng Wang, Xuwei Liu, Jinyao Guo, Shiwei Feng, and Xiangyu Zhang. RFCAudit: AI agent for auditing protocol implementations against RFC specifications. In *2025 40th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 1221–1233. IEEE, 2025c. doi: 10.1109/ASE63991.2025.00105. URL <https://doi.org/10.1109/ASE63991.2025.00105>.
- Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, et al. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026b.
- Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. User perceptions of smart home IoT privacy. *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 2:200:1–200:20, 2018. doi: 10.1145/3274469.
- Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):38:1–38:55, 2014. doi: 10.1145/2629592.
- Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 414–431, 2025d.
- Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let’s think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- Shanshan Zhong, Kate Shen, and Chenyan Xiong. Agentwebbench: Benchmarking multi-agent coordination in agentic web. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.
- Wanjuan Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory, 2023. URL <https://arxiv.org/abs/2305.10250>.
- Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models, 2024a. URL <https://arxiv.org/abs/2310.04406>.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023. URL <https://arxiv.org/abs/2205.10625>.
- Hanzhang Zhou, Panrong Tong, Xu Zhang, Quyu Kong, Chenglin Cai, Tianyu Xia, Gongjie Zhang, Jianan Zhang, Long Li, Long Chen, Lei Wang, Gaole Dai, Pengxiang Li, Liangyu Chen, Yue Wang, and Steven Hoi. Qwen-UI-Agent technical report: Toward next-generation real-world centric foundation GUI agents, 2026a. URL <https://arxiv.org/abs/2607.28227>.
- Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024b. URL <https://openreview.net/forum?id=oKn9c6ytLx>.

- Xiao Zhou, Siyue Zhang, Yilun Zhao, Jinbiao Wei, Tingyu Song, Arman Cohan, and Chen Zhao. Gui vs. cli: Execution bottlenecks in screen-only and skill-mediated computer-use agents, 2026b. URL <https://arxiv.org/abs/2606.24551>.
- Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. Sotopia: Interactive evaluation for social intelligence in language agents, 2024c. URL <https://arxiv.org/abs/2310.11667>.
- Andrew Zhu and Chris Callison-Burch. Overhearing llm agents: A survey, taxonomy, and roadmap, 2025. URL <https://arxiv.org/abs/2509.16325>.
- Brian Zhu, Philipp Schmitt, Philine Meister, Lukas Gensler, Momen Khalil, Emmanuele Poggi, Johannes Hechtl, Carsten Braunroth, Kai Wurm, Gokul Narayanan, Eugen Solowjow, Georg von Wichert, Andre Scholz, Felix Albrecht, and Maximilian Metzner. A factory-floor deployment case study of vla pipelines for industrial packaging task: Workflow, failures, and lessons. *arXiv preprint arXiv:2605.27461*, 2026a.
- Fangqi Zhu, Zhengyang Yan, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. Wmpo: World model-based policy optimization for vision-language-action models. *arXiv preprint arXiv:2511.09515*, 2025a.
- Jiarun Zhu, Yijun Hong, Xiaoquan Sun, Zetian Xu, Qijun He, Haijier Chen, Zhiyong Wang, Mingqi Yuan, Wenjun Zeng, and Jiayu Chen. Can vision-language-action models learn from real-world data continually without forgetting? *arXiv preprint arXiv:2605.26820*, 2026b.
- Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.
- Shitong Zhu, Chenhao Fang, Derek Larson, Neel Reddy Pochareddy, Rajeev Rao, Sophie Zeng, Yanqing Peng, Wendy Summer, Alex Goncalves, Arya Pudota, et al. Compliance brain assistant: Conversational agentic ai for assisting compliance tasks in enterprise environments. *arXiv preprint arXiv:2507.17289*, 2025b.
- Sibo Zhu, Wenyi Wu, Kun Zhou, Stephen Wang, and Biwei Huang. Hybrid self-evolving structured memory for gui agents, 2026c. URL <https://arxiv.org/abs/2603.10291>.
- Xiaomeng Zhu, Fengming Zhu, Weijie Zhou, Ye Tian, Zhenlin Hu, Yufei Huang, Yuchun Guo, Xinyu Wu, Zhengyou Zhang, Fangzhen Lin, and Xuantang Xiong. ProAct: A benchmark and multimodal framework for structure-aware proactive response. *arXiv preprint arXiv:2602.03430*, 2026d.
- Yuxuan Zhu, Antony Kellermann, Dylan Bowman, Philip Li, Akul Gupta, Adarsh Danda, Richard Fang, Conner Jensen, Eric Ihli, Jason Benn, Jet Geronimo, Avi Dhir, Sudhit Rao, Kaicheng Yu, Twm Stone, and Daniel Kang. CVE-Bench: A benchmark for AI agents’ ability to exploit real-world web application vulnerabilities. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025c. URL <https://arxiv.org/abs/2503.17332>.
- Zilin Zhu, Longteng Guo, Yanghong Mei, Bowen Pang, Zongxun Zhang, Xingjian He, Ruyi Ji, and Jing Liu. When robots do the chores: A benchmark and agent for long-horizon household task execution. *arXiv preprint arXiv:2605.14504*, 2026e.
- Albert Ziegler, Eirini Kalliamvakou, Shawn Simister, Ganesh Sittampalam, Alice Li, Andrew Rice, Devon Rifkin, and Edward Aftandilian. Productivity assessment of neural code completion, 2022. URL <https://arxiv.org/abs/2205.06537>.
- Zoo. Announcing Zookeeper, Our Conversational CAD Agent. <https://zoo.dev/blog/announcing-zookeeper>, February 2026. Accessed: 2026-10-05.
- Zoox. Zoox Robotaxi Launches in Las Vegas. <https://zoox.com/journal/las-vegas>, September 2025. Accessed: 2026-07-20.
- Zoox. Zoox Expands Robotaxi Service and Unveils New Product Features. <https://zoox.com/journal/zoox-service-updates-and-expansions>, March 2026. Accessed: 2026-07-20.
- Yuxin Zuo, Zikai Xiao, Li Sheng, Fei Huang, Jianhong Tu, Yuxuan Liu, Tianyi Tang, Xiaomeng Hu, Yang Su, Qingfeng Lan, Yantao Liu, Qin Zhu, Yinger Zhang, Bowen Yu, Haiquan Zhao, Haiyang Xu, Jianxin Yang, Jiayang Cheng,

Junyang Wang, Lianghao Deng, Mingfeng Xue, Tianyi Bai, Yang Fan, Yubo Ma, Yucheng Li, Zeyu Cui, Zhihai Wang, Zhihui Xie, Zhuorui Ye, An Yang, Dayiheng Liu, Jingren Zhou, and Ning Ding. Qwen-agentworld: Language world models for general agents, 2026. URL <https://arxiv.org/abs/2606.24597>.