SENTINEL: A Geo-Contextual AI Framework for Proactive Knowledge Synthesis and In-Situ Case Deflection in Global Enterprises

Author: Neeraj Choudhary Dónal Doyle

Abstract In large-scale distributed enterprises, traditional Knowledge Management (KM) systems face a critical failure mode: static documentation cannot keep pace with evolving operational realities and regional nuances. This "knowledge latency" forces employees out of self-service workflows and into costly support ticketing queues. This paper introduces SENTINEL, a geo-contextual AI framework designed to shift enterprise support from reactive retrieval to proactive interception. The architecture employs a novel dual-engine system integrated into an omni-present interface. The first engine utilizes Large Language Models (LLMs) to conduct pre-emptive, historical case-grounded audits of documentation, generating a "Contextual Density" score that identifies friction zones. The second engine is an autonomous Retrieval-Augmented Generation (RAG) agent that surfaces *in-situ* via a location-intelligent assistant window, resolving queries in real-time. By functioning as a strategic "defensive barrier" at the point of origin, SENTINEL demonstrates how a proactive AI assistant can drive high-fidelity, in-situ case deflection.

1. Introduction: The Imperative for Proactive Interception

Enterprise efficiency relies on centralized knowledge bases to enable employee self-service for areas such as HR and Technical support. However, current Knowledge Management (KM) systems operate on a passive "pull" paradigm—relying on users to recognize a knowledge gap, formulate a query, and successfully retrieve static information. This model fractures under the velocity of modern enterprise change and regional complexity.

As policies evolve, static documentation suffers from "information decay" [8]. When employees encounter low-context documentation that fails to address their specific localized needs, the cognitive load of continued searching often outweighs the effort of submitting a support ticket. This turns self-service platforms into mere waypoints on the path to expensive human support, creating massive operational bottlenecks centered on redundant, low-complexity inquiries.

Strategic Contribution: Addressing this requires a fundamental paradigm shift from reactive *knowledge retrieval* to proactive *knowledge synthesis*. SENTINEL leverages geo-contextual AI to create a strategic defensive barrier, utilizing an autonomous assistant window to resolve user

confusion before it escalates into a support ticket.

2. Theoretical Grounding and Related Work

2.1 Reactive vs. Proactive Information Delivery

Traditional help centers rely on users proactively searching. According to Cognitive Load Theory [1], when the effort to locate information exceeds a certain threshold, users abandon the task. In an enterprise context, "abandonment" manifests as a support ticket. SENTINEL draws inspiration from "Just-in-Time Information Retrieval Agents" [2], which argue that information is most valuable when presented proactively based on the user's immediate environment.

2.2 The Advantage of In-Situ Support

Research into "Task-Switching Costs" [7] suggests that every time an employee navigates away from their primary workspace to a secondary support portal, there is a measurable loss in productivity and focus. SENTINEL minimizes this "Switching Cost" by remaining *in-situ*—providing support on the page where the friction occurs through an integrated AI assistant.

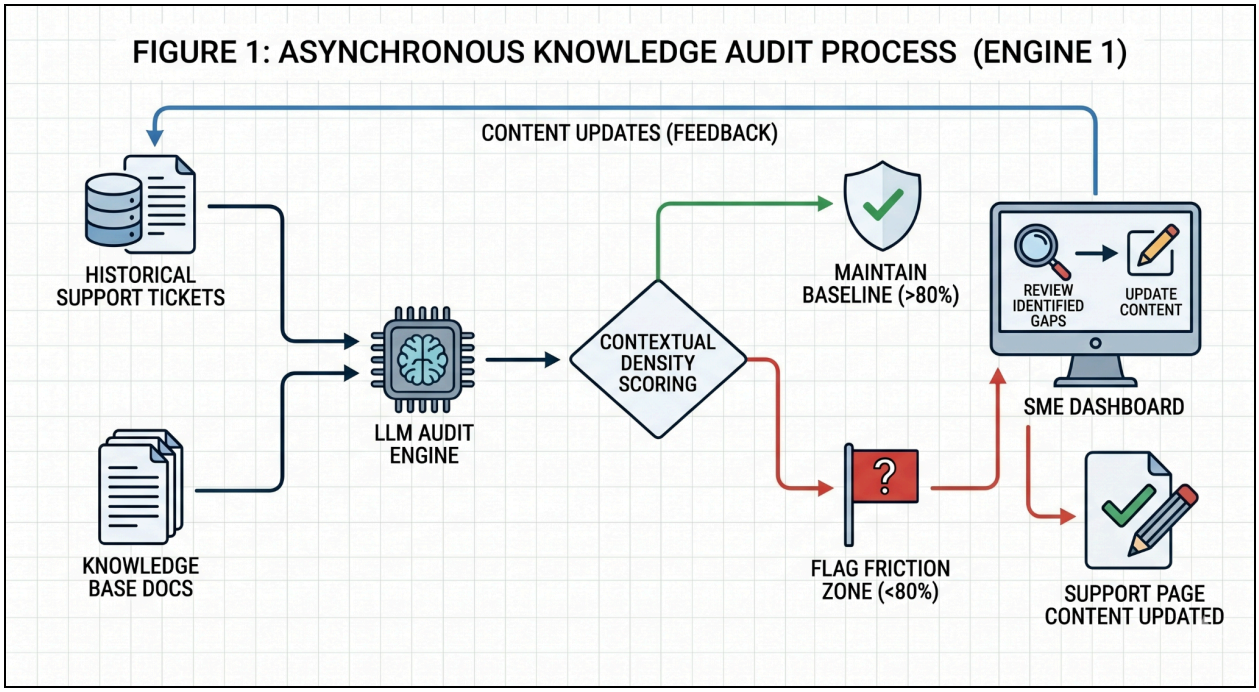
3. Methodology: The SENTINEL Dual-Engine Architecture

SENTINEL operates on a continuous, closed-loop architecture split into two core phases: an asynchronous audit phase and a real-time interception phase.

3.1 Part 1: Engine 1 – The Asynchronous Audit Loop

The first part of the framework is dedicated to the pre-emptive identification of knowledge gaps before they impact the user journey.

- Historical Data Ingestion: The engine continuously ingests anonymized historical support tickets and existing knowledge base documentation.
- LLM Audit Engine: A Large Language Model (LLM) cross-references documentation against the support corpus to identify recurring queries that are absent or poorly addressed in the current text.
- Contextual Density Scoring: The engine computes a score for each page. Pages maintaining a score >80% are considered high-fidelity and remain in the "baseline" state.
- Friction Zone Flagging: Pages scoring <80% are flagged as "Friction Zones." This data is pushed to a Domain Expert Dashboard, providing insights into specific gaps (e.g., missing regional SLAs) and auto-drafting remediations to "heal" the content.



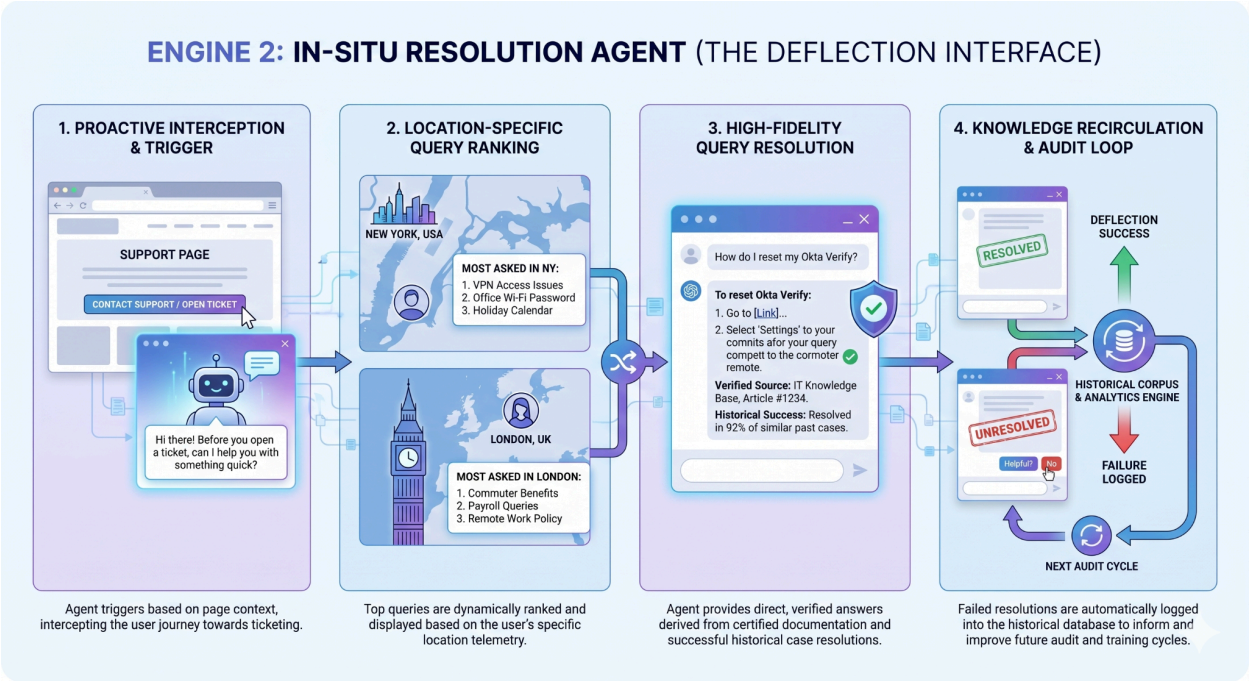
3.2 Part 2: Engine 2 – The Real-Time Interception Loop

The second part of the architecture manages the user-facing interception and resolution process, functioning as the primary vehicle for case deflection.

- **Instant Sentinel Triggering:** The moment an employee navigates to a documentation page flagged in Part I as a "Friction Zone," the SENTINEL AI assistant window proactively pops up. This immediate intervention intercepts the user before they can abandon the page to search for ticketing options.
- **Location-Intelligent Engagement:** To actively engage the user and maximize the probability of deflection, the interface utilizes hyper-local telemetry to preemptively display the **top 5 most recurring queries** specific to the user's geographic location (e.g., "Top 5 trending queries in the Dublin Hub today"). This ensures the initial interaction is highly relevant to the localized nuances of the user's region, addressing "unspoken" questions instantly.
- **In-Situ RAG Resolution:** Leveraging Retrieval-Augmented Generation (RAG) [3], the agent synthesizes answers from both the knowledge base and verified historical resolutions, providing an instant, high-fidelity answer without requiring the user to leave the page.
- **Reinforcement & Escalation:**
    - **Success (Deflection Recorded):** If the query is resolved, a "Ticket Avoided" signal is recorded, providing a reinforcement signal to the audit engine confirming the content's efficacy.
    - **Failure (Case Escalation):** If the AI assistant is unable to resolve the query, it

seamlessly asks the user to raise a support case directly through the chat window.

- **Closed-Loop Deflection:** Once the ticket is resolved, the system identifies the specific reason why the case was created (the "knowledge gap"). This insight is fed back to the audit engine part 1 to update the support page content, audits, ensure the system continuously learns from unhandled edge cases ensuring that the same query is successfully deflected for the next user.



4. Expected Impact: Maximizing Case Deflection

4.1 Case Deflection through In-Situ Interception

The primary operational impact of SENTINEL is the substantial deflection of low-complexity inquiries. By providing high-fidelity resolutions directly within the support page through the integrated AI assistant, the framework effectively neutralizes inquiries before the user initiates a ticket.

Based on initial modeling, we target a 40-50% deflection rate of all redundant cases. This is achieved by:

1. Predictive Engagement: The AI assistant engages the user with localized trending queries, solving the "unspoken" questions that lead to ticket creation.
2. Workflow Preservation: By resolving the issue in the current viewport, the assistant

removes the friction of navigating to a support portal.

4.2 Strategic Resource Allocation

The most significant organizational benefit is the shielding of administrative teams from "ticket noise." By filtering out the high volume of low-complexity queries through the proactive AI assistant, domain experts can reallocate their focus toward high-value, complex policy strategy and organizational development.

5. Conclusion

This research demonstrates that minimizing internal support friction requires a shift from reactive help to proactive interception at the point of origin. By deploying a geo-contextual dual-engine system with a focus on location-intelligent AI assistant windows, organizations can create a strategic defensive barrier that resolves user confusion at its source. This high-fidelity deflection model ensures that internal documentation evolves alongside user needs, fundamentally improving the global self-service ecosystem.

6. References and Extended Research

[1] Sweller, J. (1988). "Cognitive load during problem solving: Effects on learning." *Cognitive Science*.

[2] Rhodes, B. J., & Maes, P. (2000). "Just-in-time information retrieval agents." *IBM Systems Journal*.

[3] Lewis, P., et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *NeurIPS*.

[4] Gartner (2024). "Service and Support Leaders Priorities: Improving Self-Service to Handle Complexity."

[5] Forrester (2025). "Total Economic Impact of AI-Native Enterprise Service Management."

[6] ResearchGate (2023). "LLM-Assisted Content Analysis: Supporting Deductive Coding."

[7] Monsell, S. (2003). "Task Switching." *Trends in Cognitive Sciences*.

[8] Davenport, T. H. (2005). *Thinking for a Living: How to Get Better Performance and Results from Knowledge Workers*. Harvard Business Press.

[9] McKinsey & Company (2023). "The economic potential of generative AI: The next

productivity frontier."