Humanity's First Exam: Measuring the Reliability Floor of LLMs

Oran Kelly[1], **Anastasios Kementsietsidis**[1], **Chrysovalantis Anastasiou**[1], **Alex Yakubovich**[2], **Beat Weichsler**[3], **Natalie Ha**[2], **Ryan Burnell**[1], **Samuel Albanie**[1] **and Orhan Firat**[1]
[1]Google DeepMind, [2]Google Cloud, [3]Google Research

Frontier Large Language Models (LLMs) now achieve breakthrough-level performance across domains such as mathematics, programming, and science. However, these same models can still occasionally fail on elementary tasks that humans can easily solve. We argue this is a byproduct of modern evaluation design. Standard benchmarks typically measure the "ceiling" capability of models via single-shot (pass@1) or multi-sample (pass@k) accuracy on increasingly specialized tasks, hiding deficits in floor reliability.

To address this gap, we introduce Humanity's First Exam (HFE), a benchmark consisting of 544 text and multimodal prompts across 20 foundational subcategories. Prompts are organized into four-prompt families, each built from a seed prompt filtered through three overlapping criteria (Simplicity, Relatability, and Shock Value) so that they are easy to explain, straightforward to solve, and surprising when a frontier model fails. Evaluating 18 frontier and fast-tier models from leading providers, we find that high single-shot accuracy masks substantial reliability deficits when models must succeed consistently across repeated rollouts and closely related variants. The top five models reach 96% to 98% pass@1 and 100% family pass@4, yet average only 85.8% on family pass^4 (80.5% to 92.4%), with a 7.6 to 80.1 percentage-point Family Reliability Gap across all 18 models. Remaining failures are heavily concentrated in visual perception (73% of wrong answers), where extra test-time reasoning yields far smaller gains than on text and symbolic tasks. Real-world deployment requires evaluations that enable us to measure and raise the reliability floor, not just the capability ceiling. Therefore, we release the HFE benchmark leaderboard on Kaggle, alongside a stratified ~20% public subset of evaluation prompts.

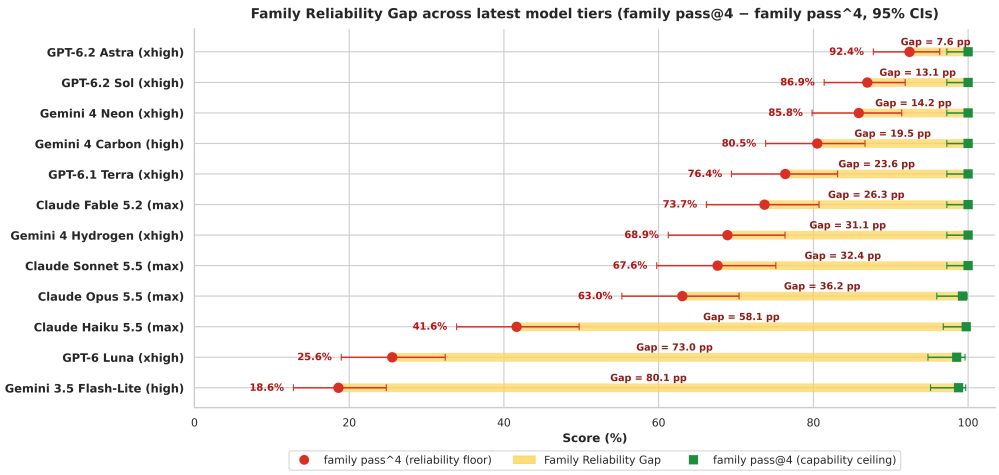


Figure 1 | Family Reliability Gap (family pass@4 − family pass^4, with 95% CIs) across 12 representative frontier and fast-tier models from Google, OpenAI, and Anthropic (four per provider). While models nearly saturate the capability ceiling (family pass@4 = 98.5% to 100%), floor reliability drops sharply (family pass^4 = 18.6% to 92.4%), leaving a 7.6 to 80.1 pp gap. Complete results appear in Table 4 and Appendix Figure A2.

Corresponding author(s): orankelly@google.com
© 2026 Google. All rights reserved

1. Introduction

Recent LLM progress has been defined by rapid gains at the boundaries of model capabilities, from gold-medal performance at the ICPC World Finals (Google DeepMind, 2025) to AI-assisted, formally verified progress on the Navier–Stokes problem (OpenAI, 2026a) and a reported Lean 4 formalization of the Generalized Riemann Hypothesis (Alpöge, 2026). Yet these same models can still fail at tasks people consider simple or routine. This discrepancy reflects the "Jagged Frontier" of artificial intelligence (Dell'Acqua et al., 2023; Karpathy, 2024), the counterintuitive phenomenon whereby a model's capability boundary is non-uniform, encompassing tasks of high technical complexity while excluding seemingly trivial and cognitively simple ones.



Prompt 1: "How many coffee mugs in the forefront do not have steam rising from them?"



Prompt 2: "List the color of the third letter of each word in the neon sign in order, separated by a semicolon."

Figure 2 | Two example prompts from HFE's public split, each drawn from a four-prompt task family: Object Counting (left; answer: *4*) and Color Identification (right; answer: *green; yellow; white; blue; blue*). Both are straightforward for humans, yet multiple models across providers fail them; the complete Color Identification family is shown in the Appendix.

LLM benchmarks currently tend to focus on model capability ceilings and efficiency metrics, but another critical component of model performance is baseline reliability. *Inconsistent* model performance (i.e., succeeding on one or even many attempts at a task, but not every time, or solving one task conclusively while failing a very closely related one) is likely to cause confusion and frustration about where LLMs are worth deploying. Moreover, for many automated workflows, time-sensitive tasks, and tasks of critical importance, LLMs can only be incorporated if people have high confidence the model will reliably complete the task without failures. Despite the importance of reliability, few frontier benchmarks explicitly measure multi-run consistency (Allcock, 2026; Barres et al., 2025; Yao et al., 2024) or variant invariance (Luo et al., 2026; Mirzadeh et al., 2024; Srivastava et al., 2024), and to our knowledge none evaluate both jointly across foundational text and multimodal tasks.

To address this blind spot, we introduce **Humanity's First Exam (HFE)**, a benchmark designed to measure the reliability floor of frontier models (Figure 1; example prompts in Figure 2). Rather than assessing whether a model can solve a difficult task at least once after multiple retries (pass@k), HFE evaluates whether a model can *consistently* solve an elementary task across both repeated runs and closely related variants (pass^k and family pass^k; building on Yao et al. 2024). Throughout this paper, reliability refers to whether a model consistently produces correct outputs, rather than serving or infrastructure stability. While reliability is equally critical in agentic workflows (Section 5.3), we scope HFE to single-turn tasks so that results reflect the model rather than external tools or agent harnesses.

In this paper, alongside introducing the HFE dataset and evaluation methodology, we investigate three core empirical questions:

1. **How low is the reliability floor across frontier LLMs?** To what extent do today's leading models fail on elementary tasks where perfect consistency is expected, and how does that floor differ across smaller and larger model tiers?
2. **How wide is the reliability gap between the capability ceiling (pass@4, family pass@4) and reliability floor (pass^4, family pass^4)?** Do models fail elementary tasks because they lack the underlying capability, or because they fail to leverage this capability consistently?
3. **Where do reliability gaps concentrate, and how much does test-time reasoning help?** Which task categories and modalities drive floor failures, how do model errors compare with human difficulty, and to what extent does increased reasoning effort close the gap?

2. Benchmark Design

Traditional LLM benchmark design prioritizes discriminative power at the upper boundary of model performance. HFE inverts this approach, measuring the baseline reliability of frontier models on elementary tasks that a majority of humans can complete.

2.1. Taxonomy

To ground our taxonomy in real-world failures, we began by curating widely shared social-media posts (X, Reddit) following frontier model releases that showed elementary LLM errors across five broad areas: logical reasoning, visual perception, basic arithmetic and counting, constraint adherence, and false-premise questions (excluding image generation tasks to focus on objectively gradable text outputs). Expanding these areas into an exploratory 500-prompt pilot set (drawn from social media, public failure lists, and elementary-level reasoning exams) revealed that coarse categories frequently entangled multiple skills within a single prompt. Fine-grained capability annotation and clustering of this 500-prompt pilot set yielded 23 atomic, single-skill sub-abilities (Table 1).

2.2. Seed Prompts

We authored new, unleaked seed prompts for each subcategory. Candidate seeds had to satisfy three requirements:

- **Single sub-ability:** Targets one isolated sub-ability so failures are diagnostic.
- **Empirical failure on calibration panel:** Elicits at least one error across four sampled rollouts on at least one model in our seed calibration panel (**Gemini 3.1 Flash-Lite**, **Gemini 3.1 Pro**, **Claude Opus 4.7**, **Claude Sonnet 4.6**, and **GPT-5.5**), filtering out already-saturated tasks at creation time (May to June 2026).
- **Virality criteria:** Passes all three binary review tags; **(a) Simplicity:** The task and failure mode can be explained easily; **(b) Relatability:** The task is easily human-solvable without domain expertise, using common sense or perception; and **(c) Shock Value:** A frontier model failure appears logically absurd and starkly incompatible with its perceived capability ceiling.

This process produced **321 seed prompts** satisfying all criteria across the 23 subcategories (Table 1).

Table 1 | Categories of tasks in HFE.

#	Category Name (Families / Prompts)	Description	Input Modality	Macro-Category
1	**Algorithmic and Symbolic Execution (8 / 32)**	Manipulating symbols strictly by rule, ignoring semantic meaning.	Text	Math & Algorithmic
2	**Analogical Reasoning (Structural)[†] (0 / 0)**	Identifying abstract relationships (A is to B as C is to D).	Text, Text & Image	Reasoning & Language
3	**Causal and Counterfactual Reasoning (2 / 8)**	Understanding cause-and-effect and hypotheticals.	Text, Text & Image	Reasoning & Language
4	**Commonsense and Physical Reasoning (4 / 16)**	Basic physics, gravity, and object permanence.	Text, Text & Image	Reasoning & Language
5	**Constraint Satisfaction (13 / 52)**	Following strict negative or formatting instructions that require suppressing the model's natural output.	Text	Math & Algorithmic
6	**General Arithmetic and Calculation (3 / 12)**	Basic integer math, money, and simple addition.	Text	Math & Algorithmic
7	**Generalization and Pattern Recognition (4 / 16)**	Identifying a rule from small examples and applying it to a new instance.	Text, Text & Image	Reasoning & Language
8	**Inductive and Probabilistic Reasoning[†] (0 / 0)**	Determining the most likely outcome based on evidence; distinguishing probability from certainty.	Text	Reasoning & Language
9	**Logical Deduction (2 / 8)**	Multi-step transitive logic.	Text	Reasoning & Language
10	**Mathematical and Quantitative Reasoning (5 / 20)**	Understanding math concepts beyond simple calculation (magnitude, comparison, units).	Text, Text & Image	Math & Algorithmic
11	**Positional Tracking (6 / 24)**	Keeping track of object state/location through a sequence of moves.	Text, Text & Image	Spatial & Positional
12	**Sequence Manipulation (6 / 24)**	Reordering, reversing, or editing strings.	Text	Math & Algorithmic
13	**Sorting and Ordering (5 / 20)**	Organizing lists by criteria (alphabetical, size).	Text, Text & Image	Math & Algorithmic
14	**Spatial and Topological Reasoning (4 / 16)**	Understanding geometry, shapes, and how objects fit together.	Text, Text & Image	Spatial & Positional
15	**Temporal Arithmetic[†] (0 / 0)**	Calculations involving dates, times, and calendars.	Text, Text & Image	Math & Algorithmic
16	**Negative Fact Verification (3 / 12)**	Correctly identifying when something did not happen or does not exist.	Text, Text & Image	Reasoning & Language
17	**Pragmatic Reasoning (3 / 12)**	Grasping implied meaning, obvious sarcasm, or the practical intent behind a statement rather than taking it literally.	Text	Reasoning & Language
18	**Color Identification (10 / 40)**	Correctly identifying specific colors/shades in an image, resisting lighting artifacts or common priors.	Text & Image	Vision
19	**Object Counting (12 / 48)**	Accurately counting distinct objects.	Text & Image	Vision
20	**Text & Symbol Recognition (11 / 44)**	Reading text (OCR) in images, especially handwriting or distorted text.	Text & Image	Vision
21	**Arrows, pointers, or indicators (11 / 44)**	Interpreting the meaning or direction of superimposed graphical elements (drawn arrows, circles) or physical pointing.	Text & Image	Vision
22	**Graph & Chart Understanding (7 / 28)**	Reading basic visual data points, legends, or obvious visual trends from simple charts.	Text & Image	Vision
23	**Spatial Reasoning (17 / 68)**	Understanding 3D physical relationships, occlusion, or relative positioning of objects depicted in a 2D image or described scene (e.g., inside/outside, in front of/behind, stacking).	Text, Text & Image	Spatial & Positional

[†] Three pilot subcategories (*Analogical Reasoning (Structural)*, *Inductive and Probabilistic Reasoning*, and *Temporal Arithmetic*), included during seed authoring (Section 2.2) and human baseline collection (Section 2.4), had no surviving families after quality subsampling (Section 2.5), leaving 20 subcategories in the final benchmark.

2.3. Prompt Families

Evaluating models on isolated phrasings risks conflating task competence with lucky surface wording. To test structural invariance, each seed prompt is expanded into a five-prompt **Prompt Family** consisting of **one seed prompt** and **four variants**: controlled surface edits (renaming entities, reordering clauses, or varying visual layout/attributes) that preserve the underlying sub-ability, cognitive difficulty, and task logic (not necessarily answer preserving).

During variant authoring (June 26 to July 26, 2026), candidate variants were calibrated across four rollouts per model on the initial five-model panel augmented with two newly released models (**Gemini 3.5 Flash** and **Claude Opus 4.8**); a variant was retained only if it elicited at least one failure on this seven-model panel. **Six** of the 321 seeds for which four model-failing variants could not be constructed were discarded during family generation, yielding **315 complete five-prompt families (1,575 prompts)**.[1]

2.4. Human Baseline

To verify that candidate tasks are easily solvable for humans without tools, matching model evaluation conditions, each seed prompt was solved independently by four humans, excluding anyone who had authored or previously reviewed the task. Raters worked self-paced using only pen and paper or a plain scratchpad, strictly prohibiting LLMs, web search, calculators, code, word or character counters, copy-paste, and writing assistants, with their completion time being logged. An independent human reviewer then graded all four human responses against the ground truth, tagging mismatches to separate genuine human errors from prompt or ground truth issues (e.g., ambiguity). Across all 1,260 human responses on these 315 seed prompts, single-rater accuracy was 80.3% with a median per-seed solve time of 1.5 minutes (IQR 1.0 to 2.5): 79.7% of seeds (251/315) were solved by at least 3 of 4 raters (50.5% at 4/4 and 29.2% at 3/4), whereas 20.3% fell below 3/4 and were subsequently removed during quality subsampling (Section 2.5). On the 136 seeds retained in the final benchmark (Section 2.5), single-rater accuracy is 90.8% (494/544, with 86 seeds at 4/4 and 50 at 3/4, and all 50 misses attributed to human rater error rather than prompt ambiguity); median per-rater solve time is 1.0 minute on both Vision and Spatial & Positional seeds (IQR 1.0 to 2.0) and 1.0 minute across the 86 unanimous 4/4 seeds.

2.5. Quality Subsampling

Starting from the **315 seed prompts (1,575 total prompts including all variants)** satisfying all three virality criteria (*Simplicity*, *Relatability*, and *Shock Value*), we enforce three sequential filtering steps (Table 2): (1) **Human simplicity**, retaining only families whose seed prompt was solved by at least 3 of 4 independent human raters (**251 families / 1,255 prompts**); (2) **QA audit**, auditing seeds, variants, autorater edge cases, and model-flagged items, removing the entire family whenever the seed prompt (or multiple variants) exhibited visual or textual ambiguity, an unfaithful variant shift, or an inaccurate ground truth (−**115 families, leaving 136 families / 680 prompts**); and (3) **Variant standardization**, dropping one variant per remaining family (any single variant flagged during QA where present, and otherwise the highest-pass-rate redundant variant) so every family contains four verified prompts (**1 seed + 3 variants**). This yields a final benchmark of **136 complete four-prompt families (544 prompts)**.

[1] Of the 18 models evaluated in Table 4, 16 were released after data collection concluded, while Gemini 3.5 Flash-Lite and Gemini 3.6 Flash launched on July 21, 2026, in the final week of variant creation. Gemini 3.5 Flash was subsequently deprecated and routed to Gemini 3.6 Flash prior to our evaluation window and is therefore omitted from Table 4.

Table 2 | Dataset filtering process.

Step	Rule	Families	Prompts	Dropped
0	Initial complete 5-prompt families (Section 2.3)	315	1,575	—
1	**Human simplicity:** At least 3 of 4 human raters solve the seed prompt	251	1,255	64 families 320 prompts
2	**QA audit:** Remove the entire family if the seed prompt (or ≥2 variants) had prompt/ground-truth ambiguity, autorater issues, or unfaithful variant shifts	136	680	115 families 575 prompts
3	**Variant standardization:** Drop 1 variant per remaining family (QA-flagged or highest-pass-rate redundant variant) to form uniform 4-prompt families	136	544	136 prompts (1 per family)

2.6. Final Dataset

The final 544-prompt HFE dataset spans 20 foundational subcategories; three pilot subcategories had no remaining families after quality subsampling (Table 1). By input modality, HFE comprises 316 multimodal (Text & Image) prompts across 79 families (58.1%) and 228 text-only prompts across 57 families (41.9%), with every family sharing a consistent modality across its four prompts. To support community reproducibility while mitigating benchmark contamination, we release a stratified ~20% public subset of 112 prompts (28 complete families) under CC-BY-4.0 on Kaggle (https://www.kaggle.com/datasets/deepmind/hfe-public) and maintain the remaining 432 prompts (108 families) as a private holdout (Section 4.7), with live scores hosted on the Kaggle HFE leaderboard (https://www.kaggle.com/benchmarks/deepmind/hfe/leaderboard).

3. Evaluation Methodology

3.1. Metrics

For each prompt, we sample $n = 6$ independent generations and let $c \in \{0, \ldots, n\}$ denote the number of correct responses.

Prompt-level metrics. Standard evaluations report pass@k (Chen et al., 2021), the probability that *at least one* of k draws without replacement is correct (pass@1 $= c/n$ measures single-shot accuracy; pass@4 measures multi-sample capability ceiling). Because elementary tasks require reliable execution on every call, we follow Yao et al. (2024) and measure operational consistency via pass^k, the probability that *all* k draws without replacement are correct:

$$\text{pass@}k = 1 - \prod_{i=0}^{k-1} \frac{n-c-i}{n-i}, \qquad \text{pass}^k = \prod_{i=0}^{k-1} \frac{c-i}{n-i}. \tag{1}$$

At $k = 1$, pass@1 $=$ pass$^1 = c/n$; as k grows, pass@k increases (capability) while pass^k decreases (consistency). We report pass^4 for single-prompt reliability and define the *Single-Prompt Reliability Gap* as pass@k − pass^k.

Family-level metrics. To measure robustness across minor prompt variations, each task family $f \in \{1, \ldots, M\}$ ($M = 136$) comprises four logic-preserving prompts (one seed and three variants) $j \in \{1, \ldots, 4\}$. Over a joint $4 \times k$ draw (k rollouts per prompt), we define:

$$\text{family pass@}k(f) = 1 - \prod_{j=1}^{4} \big(1 - \text{pass@}k(f, j)\big), \qquad \text{family pass}^k(f) = \prod_{j=1}^{4} \text{pass}^k(f, j). \tag{2}$$

Here, family pass@1 (4×1) and family pass@4 (4×4) test whether *at least one* rollout succeeds across the family, family pass^1 (4×1) isolates *single-draw variant invariance* across all four prompts, and family pass^4 (4×4) measures *joint rollout-and-variant reliability*. Prompt and family metrics are averaged over all 544 prompts and 136 families, respectively, with 95% confidence intervals via family-clustered bootstrap. Hence we define the *Family Reliability Gap* as family pass@k − family pass^k. Figure 3 illustrates how these metrics relate for a single model, and Table 3 summarizes all six definitions.

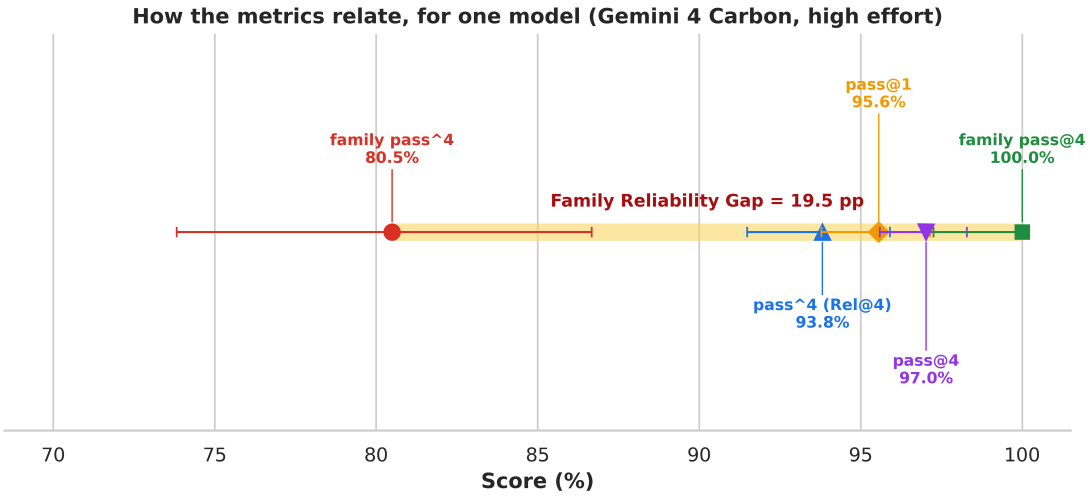


Figure 3 | The five metrics for a single model (Gemini 4 Carbon, high effort), with 95% CIs; the x-axis starts at 70%. The metrics are ordered as family pass^4 ≤ pass^4 ≤ pass@1 ≤ pass@4 ≤ family pass@4 for every model; the shaded bar is the Family Reliability Gap.

3.2. Grading

The final answer of each model response is graded against its ground-truth reference, consisting of either a golden answer or a rubric definition, by a deterministic LLM judge (**Gemini 3.1 Pro**, temperature 0). Internal reasoning traces and control tokens are stripped prior to grading, and empty visible outputs score 0. Given the prompt, stripped response, and ground-truth reference, the judge returns a binary *Yes/No*. The grading prompt specifies equivalence rules for choice formatting, numeric expressions, temporal units, and multi-criterion rubrics (provided in full in the Appendix). In a cross-provider audit of all 1,492 failed prompt–model pairs across Table 4 using Claude Fable 5.2 (weighted across 5,552 wrong-answer rollouts; Section 4.6, Figure 9A), only 3% of wrong-answer rollouts (~0.3% of all 58,752 rollouts) were flagged as potentially correct or defensible, indicating a low false-negative rate.

Table 3 | Summary of HFE evaluation metrics. Each row defines a metric for a single unit (either a single prompt for rows 1 to 3, or a single family for rows 4 to 6). Reported scores are averaged over all 544 prompts (prompt-level metrics) or all 136 families (family-level metrics); 95% CIs are from a family-clustered bootstrap.

	Metric (Scope)	Formula ($n = 6$)	What It Measures
1	**pass@1** (Single prompt)	c/n	**Single-try accuracy:** Probability **1 random rollout** is correct. Standard headline metric.
2	**pass@4** (Single prompt)	$1 - \binom{n-c}{4} / \binom{n}{4}$	**Capability ceiling:** Probability **at least 1 of 4 rollouts** is correct.
3	**pass^4 (Rel@4)** (Single prompt)	$\binom{c}{4} / \binom{n}{4}$	**Rollout consistency:** Probability **all 4 of 4 rollouts** on the same prompt are correct.
4	**family pass@4** (4-prompt family)	$1 - \prod_{j=1}^{4} \binom{n-c_j}{4} / \binom{n}{4}$	**Family capability ceiling:** Probability **at least 1 of 16 rollouts** across the family is correct.
5	**family pass^4** (4-prompt family)	$\prod_{j=1}^{4} \binom{c_j}{4} / \binom{n}{4}$	**Primary floor metric:** Probability **all 16 rollouts** (4 per prompt) across the family are correct.
6	**Family Reliability Gap** (Ceiling – Floor)	family pass@4 – family pass^4	**Main metric:** How often a model that can solve a family fails to solve it reliably.

3.3. Experiment Setup

Models are evaluated zero-shot in a single turn with no system prompt and no access to external tools (search, code execution). Multimodal prompts pass raw, uncropped image bytes. For each of the 544 prompts, we sample $n = 6$ rollouts (3,264 per model) using provider-default decoding parameters (e.g., temperature and top-p) at the reasoning effort levels specified in Tables 4 and A1, scoring failed or empty generations as 0.

4. Results

4.1. Overall Performance

Evaluating 18 models on HFE, we observe the following:

- **Frontier models nearly saturate the benchmark when asked to succeed at least once.** The top five models reach 96% to 98% pass@1 and 97% to 98% pass@4, and for every model in the top 12, family pass@4 is 99.3% to 100% (Table 4).
- **However, they still have substantial headroom when asked to succeed every time.** Family pass^4 maxes out at 92.4% (GPT-6.2 Astra), followed by 86.9% for GPT-6.2 Sol and 85.8% for Gemini 4 Neon. Because family pass@4 is close to 100% for every model, the Family Reliability Gap is almost entirely driven by family pass^4, and ranges from 7.6 pp to 80.1 pp (Figure 1; complete 18-model breakdown in Appendix Figure A2).
- **Small differences on pass@1 translate into large gaps when consistency is required.** For instance, Claude Opus 5.5 trails GPT-6.2 Astra by 6.6 pp on pass@1, yet by 29.4 pp on family pass^4. This widening is expected because any failure among the 16 sampled rollouts fails the family, illustrating how quickly small per-prompt error rates compound when consistent execution across repeated runs and minor prompt variations is required. On average across models, roughly half of the family pass^4 shortfall comes from consistent prompt failures (0 of 6 rollouts correct) and half from run-to-run flakiness (between 1 and 5 of 6 correct).

- **Floor reliability generally scales with model tier within each provider,** though not monotonically for Google or Anthropic models (Figure 4): for Google, the balanced-tier Gemini 4 Neon (85.8%) outperforms the flagship Gemini 4 Carbon (80.5%), while for Anthropic, Claude Opus 5.5 trails Claude Sonnet 5.5, largely due to refusals (see Section 4.6), before recovering on Claude Fable 5.2.

Table 4 | HFE performance across 18 frontier and fast-tier models (n = 544 prompts, 136 families; 11 additional reference models appear in Appendix Table A1).

Rank	Model	Effort	pass@1 (±95% CI)	pass@4 (±95% CI)	pass^4 / Rel@4 (±95% CI)	family pass@4 (4×4, 95% Wilson CI)	family pass^4 (4×4, ±95% CI)	Family Reliability Gap (fam pass@4 − fam pass^4)	Avg Thought Tokens
1	**GPT-6.2 Astra**	xhigh	97.8% (±1.5)	97.9% (±1.5)	97.5% (±1.6)	100.0% (97.3, 100.0)	92.4% (±4.3)	7.6 pp (±4.3)	204
2	**GPT-6.2 Sol**	xhigh	97.3% (±1.5)	98.4% (±1.2)	95.8% (±1.9)	100.0% (97.3, 100.0)	86.9% (±5.2)	13.1 pp (±5.2)	274
3	**Gemini 4 Neon**	xhigh	96.8% (±1.5)	97.9% (±1.3)	95.6% (±1.9)	100.0% (97.3, 100.0)	85.8% (±5.8)	14.2 pp (±5.8)	9,852
4	**GPT-6 Astra**	xhigh	96.8% (±1.6)	98.0% (±1.4)	94.9% (±2.0)	100.0% (97.3, 100.0)	83.4% (±5.8)	16.6 pp (±5.8)	382
5	**Gemini 4 Carbon**	high	95.6% (±1.7)	97.0% (±1.3)	93.8% (±2.2)	100.0% (97.3, 100.0)	80.5% (±6.4)	19.5 pp (±6.4)	6,616
6	**GPT-6.1 Sol**	xhigh	96.1% (±1.7)	97.8% (±1.6)	93.7% (±2.2)	100.0% (97.3, 100.0)	79.1% (±6.4)	20.9 pp (±6.4)	396
7	**Gemini 4 Argon**	high	94.1% (±2.1)	96.4% (±1.6)	91.5% (±2.8)	100.0% (97.3, 100.0)	76.6% (±6.8)	23.4 pp (±6.8)	1,838
8	**GPT-6.1 Terra**	xhigh	94.7% (±2.1)	96.6% (±1.9)	91.8% (±2.7)	100.0% (97.3, 100.0)	76.4% (±6.9)	23.6 pp (±6.9)	622
9	**Claude Fable 5.2**	max	93.4% (±2.2)	95.3% (±1.9)	91.5% (±2.6)	100.0% (97.3, 100.0)	73.7% (±7.3)	26.3 pp (±7.3)	2,951
10	**Gemini 4 Hydrogen**	xhigh	92.5% (±2.4)	95.6% (±1.7)	88.4% (±3.2)	100.0% (97.3, 100.0)	68.9% (±7.5)	31.1 pp (±7.5)	13,775
11	**Claude Sonnet 5.5**	max	91.7% (±2.4)	94.5% (±1.9)	89.2% (±2.9)	100.0% (97.3, 100.0)	67.6% (±7.7)	32.4 pp (±7.7)	6,340
12	**Claude Opus 5.5**	max	91.2% (±2.6)	94.2% (±2.3)	87.8% (±3.0)	99.3% (96.0, 99.9)	63.0% (±7.6)	36.2 pp (±7.5)	5,162
13	**Gemini 3.8 Flash**	high	87.1% (±3.2)	91.9% (±2.6)	81.7% (±4.0)	100.0% (97.3, 100.0)	56.8% (±8.1)	43.2 pp (±8.1)	8,150
14	**Claude Fable 5.1**	max	85.7% (±3.1)	89.9% (±2.7)	80.7% (±3.8)	100.0% (97.3, 100.0)	49.7% (±8.0)	50.3 pp (±8.0)	4,665
15	**Gemini 3.6 Flash**	high	83.8% (±3.4)	90.9% (±2.7)	75.8% (±4.4)	100.0% (97.3, 100.0)	44.7% (±8.0)	55.3 pp (±8.0)	3,473
16	**Claude Haiku 5.5**	max	83.6% (±3.3)	89.4% (±2.8)	76.3% (±4.0)	99.8% (96.8, 100.0)	41.6% (±7.9)	58.1 pp (±7.9)	13,351
17	**GPT-6 Luna**	xhigh	74.8% (±4.1)	85.4% (±3.7)	60.9% (±5.0)	98.5% (94.8, 99.6)	25.6% (±6.7)	73.0 pp (±6.9)	1,545
18	**Gemini 3.5 Flash-Lite**	high	68.7% (±4.4)	83.6% (±3.6)	52.6% (±5.3)	98.8% (95.1, 99.7)	18.6% (±6.0)	80.1 pp (±6.1)	3,094

Family-level metrics (family pass@4, Family Reliability Gap, and family pass^4) are averaged across all four-prompt families with 95% confidence intervals computed via family-clustered bootstrap, except family pass@4, which uses a Wilson score interval (to avoid degenerate confidence intervals where family pass@4 is 100.0% in every bootstrap sample). Family pass@4 reaches 100% for most models because in essentially every family, at least one prompt is solved on 3 or more of its 6 rollouts, so the empirical probability that all 16 draws fail is zero. Differences are calculated from unrounded values. Thought-token counts reflect each provider's native tokenizer and API reporting (e.g., OpenAI `reasoning_tokens`; Gemini 4 Argon low in Tables A1 and A2 reports 0 thought tokens) and are not directly comparable across providers. Models are ordered by descending family pass^4 at highest evaluated reasoning effort.

4.2. Performance Disaggregated by Task Categories and Modality

To isolate the drivers of floor reliability, we disaggregate family pass^4 performance across four macro-categories: **Math & Algorithmic** (n = 40 families), **Reasoning & Language** (n = 18 families), **Spatial & Positional** (n = 27 families), and **Vision** (n = 51 families) (Table 5), as well as by input modality (**Text-only** vs. **Multimodal**; Figure 5). We find the following:

- **Multimodal tasks are a major source of reliability failures across all model classes.** While the top 9 models reach 83% to 97% family pass^4 on text-only families, their performance drops to 64% to 89% on multimodal families (Figure 5).
- **The size of this multimodal drop differs sharply across models.** It is smallest for GPT-6.2 Astra (8.0 pp) and GPT-6.1 Terra (11.5 pp), and largest for Claude Haiku 5.5, Gemini 4 Hydrogen, and Claude Fable 5.1, whose family pass^4 drops by 44.8, 45.5, and 46.4 percentage points, respectively.
- **Frontier separation concentrates in Vision and Spatial & Positional tasks.** Across the top 9 models in Table 5, scores span 15.0 pp on Math & Algorithmic (83.3% to 98.3%, led by GPT-6.2 Astra at 98.3% and Gemini 4 Neon and Gemini 4 Carbon at 97.5%) and 14.7 pp on Reasoning & Language (74.2% to 88.9%, with GPT-6.2 Astra, GPT-6 Astra, and Gemini 4 Argon tied at 88.9%), compared with 18.5 pp on Spatial & Positional (71.6% to 90.1%) and 27.8 pp on Vision (62.4% to 90.2%).

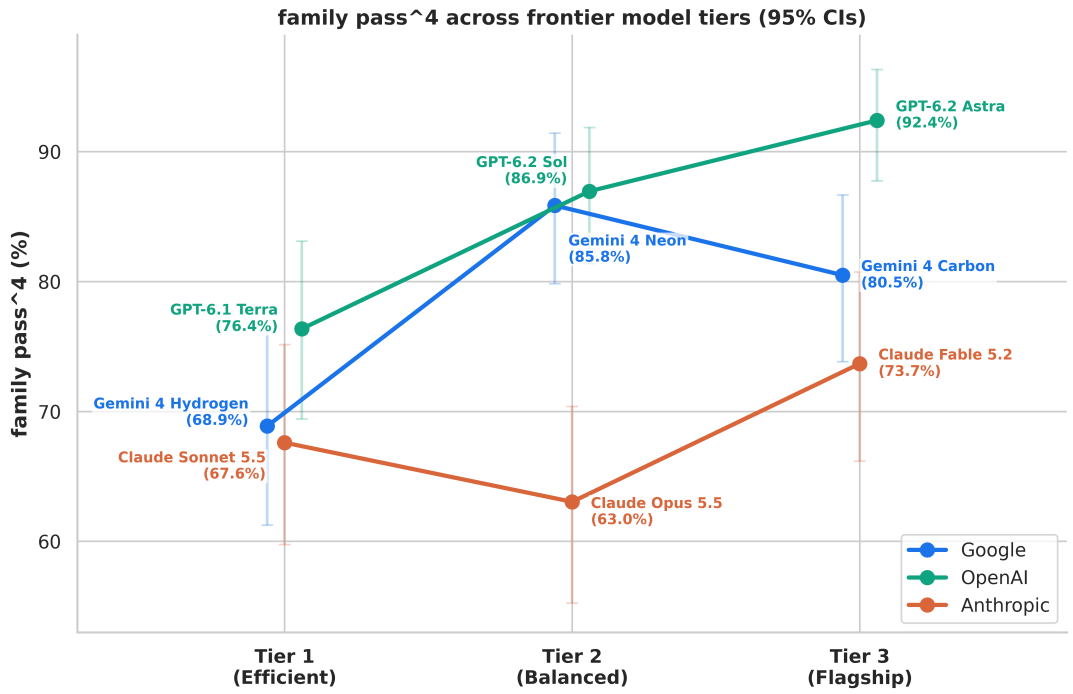


Figure 4 | Family pass[4] (with 95% CIs) across the three latest frontier-family tiers per provider (Tier 1: Efficient, Tier 2: Balanced, and Tier 3: Flagship) at highest evaluated reasoning effort, excluding fast/lite models.

Table 5 | Family pass[4] disaggregated by macro-category (95% CIs in parentheses).

Rank	Model	Effort	Math & Algorithmic (N = 160)	Reasoning & Language (N = 72)	Spatial & Positional (N = 108)	Vision (N = 204)
1	**GPT-6.2 Astra**	xhigh	**98.3%** (95.0, 100.0)	**88.9%** (72.2, 100.0)	**90.1%** (77.8, 100.0)	**90.2%** (82.4, 98.0)
2	**GPT-6.2 Sol**	xhigh	96.0% (89.7, 100.0)	83.7% (67.0, 100.0)	85.6% (71.2, 96.7)	81.7% (71.2, 90.8)
3	**Gemini 4 Neon**	xhigh	97.5% (92.5, 100.0)	83.3% (66.7, 100.0)	86.4% (72.8, 97.5)	77.3% (65.4, 88.2)
4	**GPT-6 Astra**	xhigh	92.5% (85.0, 98.3)	**88.9%** (72.2, 100.0)	84.0% (71.6, 95.1)	73.9% (62.0, 85.0)
5	**Gemini 4 Carbon**	high	97.5% (92.5, 100.0)	83.7% (67.0, 100.0)	73.1% (56.8, 87.9)	69.9% (56.9, 81.7)
6	**GPT-6.1 Sol**	xhigh	93.3% (85.0, 100.0)	80.4% (62.2, 94.8)	78.4% (63.4, 91.4)	67.9% (55.6, 79.9)
7	**Gemini 4 Argon**	high	87.0% (76.7, 95.8)	**88.9%** (72.2, 100.0)	71.6% (54.3, 86.4)	66.7% (54.3, 79.1)
8	**GPT-6.1 Terra**	xhigh	86.6% (76.1, 95.8)	74.2% (52.2, 92.6)	72.0% (55.6, 86.8)	71.4% (59.5, 82.5)
9	**Claude Fable 5.2**	max	83.3% (70.8, 93.3)	83.3% (66.7, 100.0)	74.3% (56.0, 89.1)	62.4% (49.0, 75.3)
10	**Gemini 4 Hydrogen**	xhigh	93.3% (85.0, 100.0)	83.7% (66.7, 100.0)	62.4% (44.9, 79.3)	47.9% (34.9, 60.6)
11	**Claude Sonnet 5.5**	max	85.3% (74.5, 94.3)	78.1% (56.7, 94.4)	56.8% (38.3, 75.3)	55.7% (42.0, 68.8)
12	**Claude Opus 5.5**	max	75.2% (62.7, 86.7)	79.6% (61.1, 94.4)	55.8% (38.5, 73.1)	51.5% (38.7, 64.7)
13	**Gemini 3.8 Flash**	high	76.8% (64.3, 88.3)	69.6% (48.1, 89.3)	46.4% (28.6, 64.7)	42.0% (28.8, 55.6)
14	**Claude Fable 5.1**	max	70.8% (57.5, 83.3)	81.5% (64.8, 96.3)	33.4% (15.1, 51.9)	30.5% (18.7, 42.9)
15	**Gemini 3.6 Flash**	high	69.6% (56.0, 82.3)	48.5% (27.3, 69.8)	34.6% (17.3, 51.9)	29.2% (17.6, 41.5)
16	**Claude Haiku 5.5**	max	70.3% (56.4, 83.3)	37.8% (17.8, 58.5)	31.0% (14.9, 48.3)	26.1% (15.7, 37.2)
17	**GPT-6 Luna**	xhigh	30.0% (18.6, 43.3)	52.2% (31.9, 72.6)	9.2% (0.7, 20.2)	21.3% (10.9, 32.7)
18	**Gemini 3.5 Flash-Lite**	high	46.5% (32.5, 60.6)	17.2% (0.7, 34.2)	5.7% (0.4, 14.1)	4.1% (0.8, 8.8)

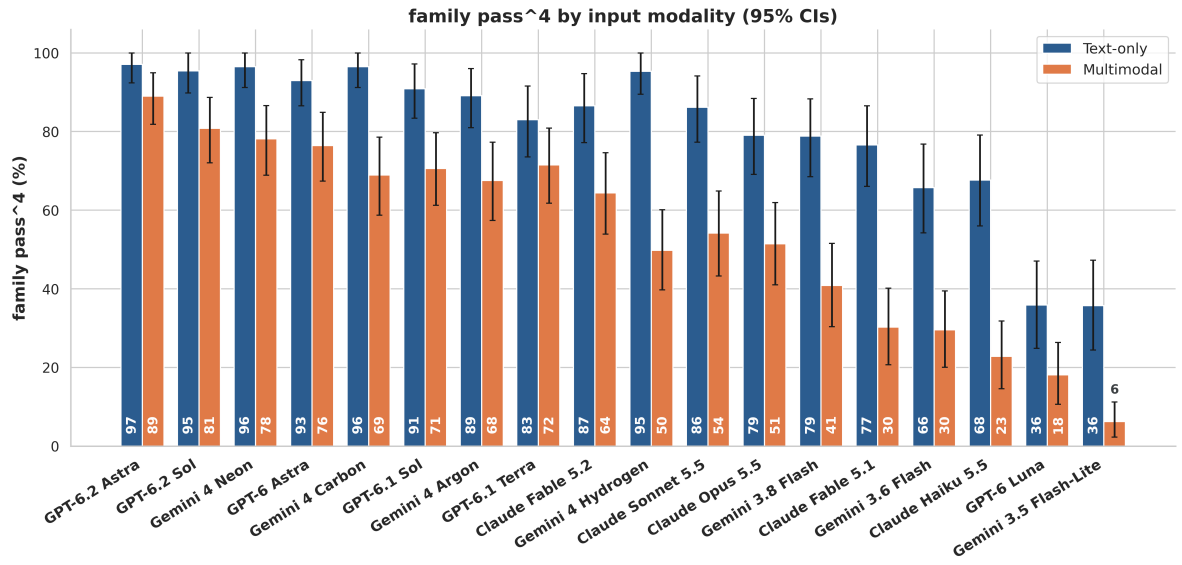


Figure 5 | Family pass^4 disaggregated by input modality.

4.3. Impact of Thinking Levels (Reasoning Effort) on Floor Reliability

Because HFE tasks are intentionally simple for humans, a natural question is whether test-time reasoning ("thinking") is necessary to achieve floor reliability, or whether models can solve these tasks in a single fast forward pass. To answer this, we evaluate 21 models across multiple reasoning effort configurations (Table A2) and find:

- **Increasing thinking effort primarily raises the reliability floor** (family pass^4) rather than the capability ceiling (family pass@4). Family pass@4 stays close to 100% at both low and high effort, while family pass^4 rises by 17.4 pp on Gemini 4 Carbon (63.1% to 80.5%) and 32.3 pp on Claude Sonnet 5.5 (35.3% to 67.6%; Figure 6A).
- **Disabling thinking causes a sharp drop in performance, but even a small thinking budget recovers most of the gain.** On Gemini 3.6 Flash, minimal effort (0 thought tokens) yields only 8.0% family pass^4, compared to 44.7% at high effort, with low effort already capturing 78% of that improvement (Figure 6A; macro-category averages across both Flash models in Figure 6C).
- **Thinking benefits symbolic and text tasks far more than visual perception.** On Gemini 4 Carbon, reducing effort from high to low drops family pass^4 by 42.4 pp on Math & Algorithmic tasks (97.5% to 55.1%), compared to just 3.1 pp on Vision tasks (69.9% to 66.8%; Table A2; 18-model category averages in Figure 6B; subcategory breakdown in Table A3).
- **Sensitivity to thinking effort varies widely, even between similar models.** At low effort, Gemini 4 Neon and Gemini 4 Hydrogen are statistically flat (+1.0 and +1.3 pp vs. xhigh, within sampling noise) and GPT-6 Astra loses only 2.5 pp family pass^4, whereas GPT-6.2 Astra, GPT-6.2 Sol, and Gemini 4 Argon lose 18.1 to 19.9 pp.

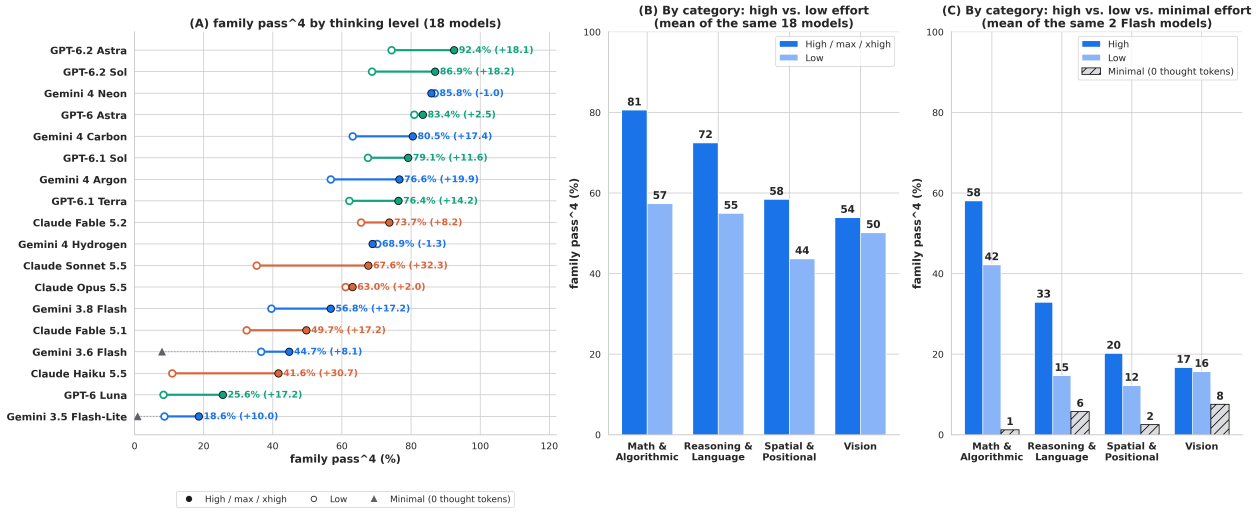


Figure 6 | Thinking levels across the 18 models from Table 4. (A) Family pass^4 per model at each effort level. (B) High vs. low effort by category, averaged over the same 18 models. (C) High vs. low vs. minimal effort by category, averaged over the two Flash models that were run at minimal effort.

4.4. Seed vs. Variant Consistency

To test whether floor failures reflect a missing capability or prompt phrasing, we look at how failures are distributed across the four prompts in each family, first within a single model and then across models. We observe the following:

- **Seeds and variants are similarly difficult.** Across the top 12 models, seed and variant pass^4 differ at most by 4.5 pp (e.g., 95.6% vs. 93.2% on Gemini 4 Carbon), and the seed prompt is the most-failed prompt in about a quarter of failing families, matching the 25% expected by chance.
- **Within a model, failing one prompt only weakly predicts failure on similar prompts.** If failures stemmed from a missing capability, a model that fails one prompt would tend to fail other prompts in the same family, and family pass^4 would sit well above the independent-failure baseline ((pass^4)^4). Instead, it is only 1.9 to 7.8 pp higher across the top 12 models (e.g., 80.5% vs. 77.4% on Gemini 4 Carbon; within-family ICC = 0.09 to 0.20).
- **Most failing families miss by a single prompt.** Among the top 12 models, 61% to 85% of families with any failed rollout have exactly one prompt with a failure, while the other three prompts are solved on all 6 rollouts (Figure 7).
- **Across models, failures concentrate on the same prompt.** In 83 of the 122 families where at least two of the 18 models fail (Table 4), a single prompt accounts for at least half of all failed rollouts. In families where at least three models fail, the most-failed prompt receives 61% of the failed rollouts on average, compared to 43% if each model's failures fell on a random prompt of the family. For example, in one spatial reasoning family that passed manual QA for consistent task logic and unambiguous ground truth (Section 2.5), 15 of the 18 models fail the same variant (78 of 108 rollouts), while the other three prompts account for only 11 failed rollouts in total.

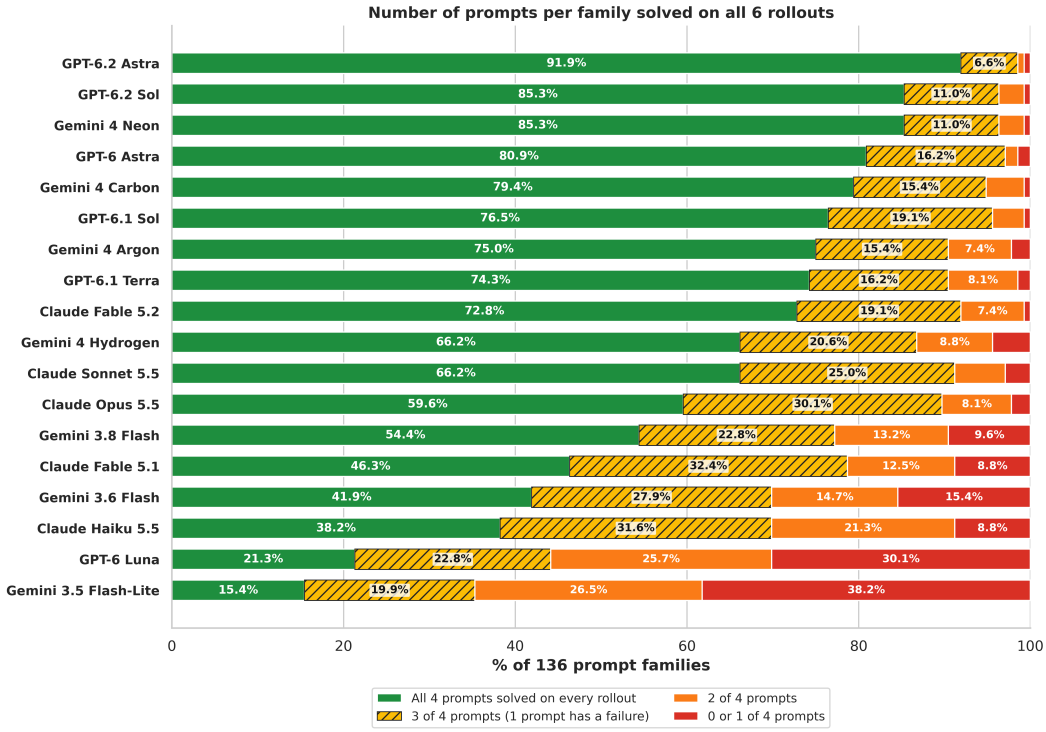


Figure 7 | Share of the 136 task families with all 4, 3, 2, or ≤ 1 prompts solved across all 6 rollouts per model.

4.5. Human vs. Model Comparison

To examine discrepancies between human and model error patterns, we compare single-rater accuracy on all 315 candidate seed prompts (Section 2.4, before prompts that fewer than 3 of 4 raters solved were removed) with pass@1 of the 18 models on the seed prompts of the 136 final families. Because these are different prompt sets (and all candidate seeds were written to make models fail), only the ordering of categories is comparable, not the absolute accuracy. We observe the following:

- **Humans and models find different categories difficult.** Humans are least accurate on Math & Algorithmic and Reasoning & Language seeds (76% and 77%) and most accurate on Vision seeds (84%). Models show the reverse order, from 98% on Math & Algorithmic to 83% on Vision, and the five models used during authoring show the same order across all 315 candidate seeds.
- **The same reversal holds at the subcategory level.** Across the 19 subcategories with at least five candidate seeds, human and model accuracy are negatively correlated (Spearman $\rho = -0.64$; Figure 8 shows human agreement by subcategory). Humans are least accurate on logical deduction, which models almost always solve, whereas models are among the least accurate on object counting and reading text in images, which are among the easiest subcategories for humans.
- **Human errors reflect rater mistakes rather than flaws in the prompt.** The reviewer attributed 244 of the 248 incorrect human responses to human rater error rather than a prompt or ground-truth issue. For example, three of four raters named a real treaty when asked about an event that never happened, and three of four omitted or added a letter when listing the capital letters in a sentence.
- **Among the final seeds, human agreement does not predict model accuracy.** Seeds solved by all four raters are no easier for models than seeds solved by three (89% vs. 93% pass@1; this holds within Vision at 81% vs. 88%, Spatial & Positional at 87% vs. 93%, and Reasoning & Language at 92% vs. 96%, while Math & Algorithmic is near ceiling for both at 99% vs. 98%, suggesting this

is not a category-mix artifact), and on 60 of the 86 seeds that all four raters solved, at least one model fails at least one rollout. The same conclusion holds at the full-family level: restricting the entire 18-model evaluation to the 86 families whose seed all four raters solved leaves the Family Reliability Gap virtually unchanged at 7.0 to 80.6 pp (compared with 7.6 to 80.1 pp overall), and 13 of the 18 models actually score lower on family pass[4] across these unanimous 4/4 families than across the 50 3/4 families (18-model mean 64.0% vs. 66.8%; for example, Gemini 4 Carbon scores 77.7% vs. 85.3% and Claude Opus 5.5 scores 59.0% vs. 70.0%).

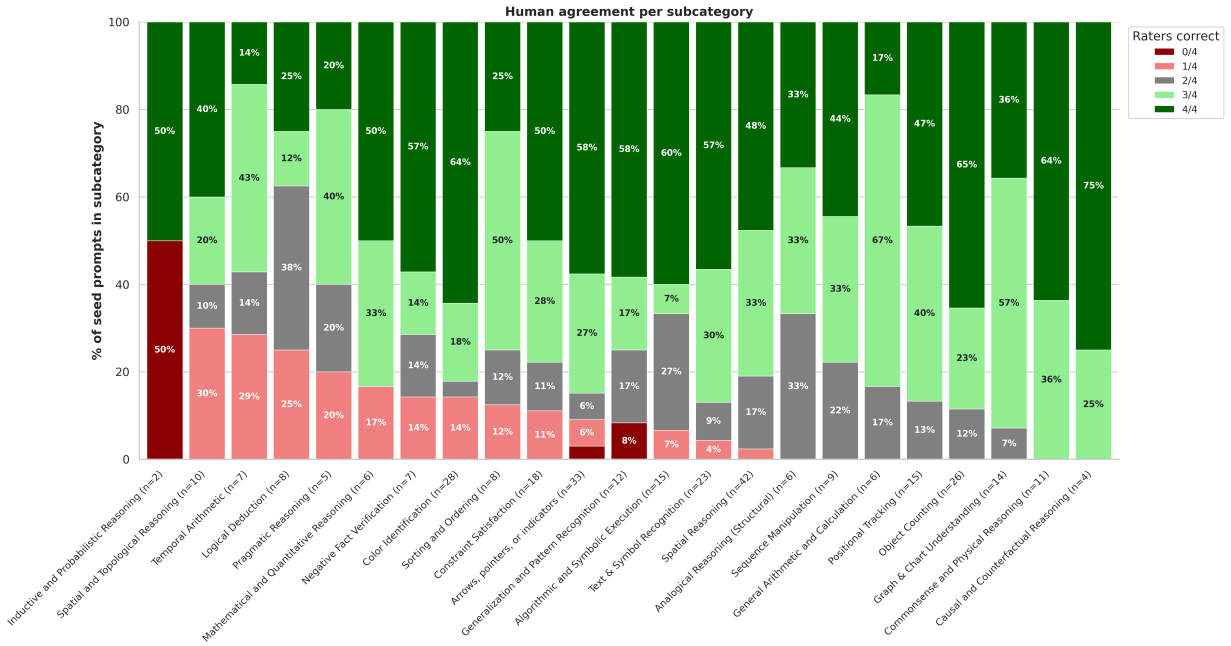


Figure 8 | Human agreement per subcategory on the 315 candidate seed prompts: share of seeds solved by 0, 1, 2, 3, or 4 of the 4 raters (n = number of seeds). Seeds below 3/4 were removed (Section 2.5).

4.6. Failure Modes

A rollout can fail in four ways: (1) a **wrong answer**, where the model responds and the autorater grades the response as incorrect; (2) a **refusal**, where the provider blocks the response with a safety stop reason or content filter; (3) an **output-limit truncation**, where the response reaches the output-token limit before giving an answer; and (4) an **evaluation-harness error**, such as a malformed response or an API quota error. Only the first is a model error; the other three are failures without an answer. Across the 18 models evaluated, 95% of the 5,823 failed rollouts are wrong answers, 3% are refusals, 2% are output-limit truncations, and 0.1% are harness errors (Table 4; Table 6). We first describe the wrong answers and then the other three types.

Wrong answers (model errors). To characterize incorrect responses, we examined every prompt on which one of the 18 models gave at least one wrong answer (385 prompts, 1,492 prompt–model pairs, 5,552 wrong rollouts). For each pair, an LLM judge (Claude Fable 5.2, medium reasoning effort) read one failed response together with the prompt, the reference answer, and the image (if there is one), and assigned a primary cause and a set of behavior flags. Shares of wrong rollouts weight each pair by its number of wrong rollouts, which assumes the other wrong rollouts of a pair fail for the same reason. The labels were spot-checked manually but not validated by human annotators, and

because the judge is itself one of the evaluated models, the shares below are approximate; a second judge cross-checked the labels. We observe the following:

- **About three-quarters of wrong answers are visual perception errors** (73% of wrong rollouts, Figure 9A). The two largest causes are misjudging a position, direction, or path in an image (31%) and miscounting objects (23%).
- **Many wrong answers are near misses.** In 40% of failed prompt–model pairs, the answer is off by one or has a single wrong element; this holds for 173 of the 303 counting errors.
- **Occlusion is a recurring cause.** In 16% of failed pairs (concentrated in 53 prompts), the error involves an object that is partly hidden, overlapping, or drawn over (for example, overlapping objects in a cluster, such as keys bunched on a ring).
- **Models default to answering the familiar version of a question.** In 8% of wrong rollouts, the model solves the well-known puzzle or accepts a false premise instead of reading the changed detail (for example, giving the textbook answer to a well-known riddle whose key constraint has been altered).
- **Smaller models slip while executing simple text procedures.** 111 of the 132 such failures come from the smallest fast-tier models across all three providers (54 across the three Gemini Flash/Flash-Lite models, 48 from GPT-6 Luna, and 9 from Claude Haiku 5.5), and in 42 of the 132, the response derives the correct result step by step before writing a different final answer.
- **Hedging differs by provider** (Figure 9B). Anthropic models express doubt or offer alternatives in 39% of their failed pairs, compared with 14% for Google and 7% for OpenAI models, while declining to give any definite answer is rare (4% of wrong rollouts). Because we did not label correct responses, we cannot say whether hedging is more common when a model is wrong.

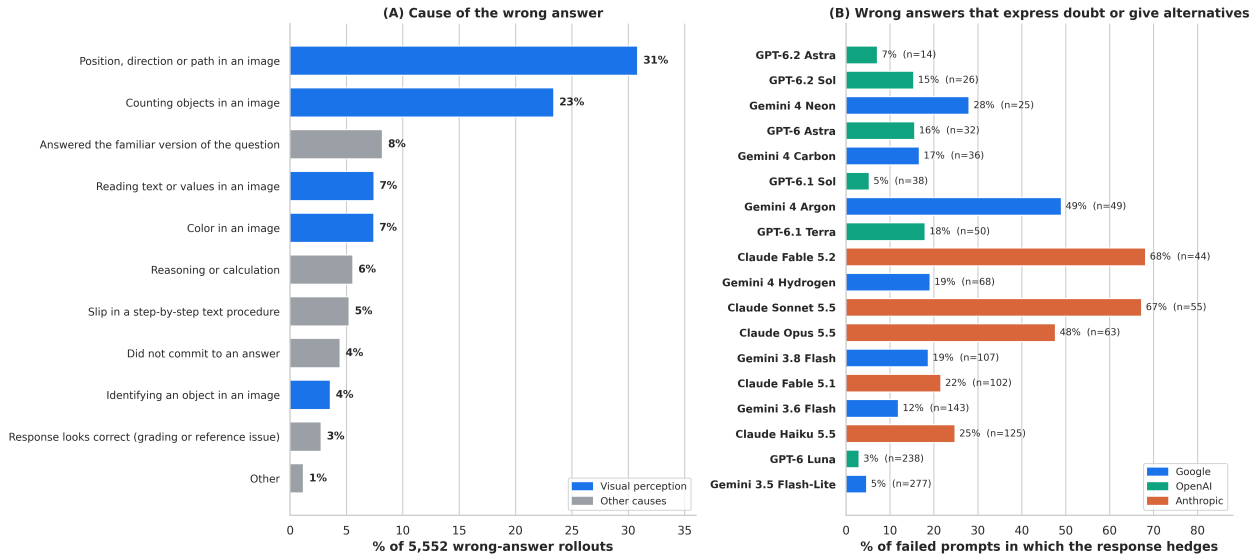


Figure 9 | Behaviors behind wrong answers, as labeled by an LLM judge. (A) Cause, as a share of wrong-answer rollouts. (B) Share of each model's failed prompts in which the response hedges; n is the number of failed prompts.

Agreement with a second judge. Repeating the labeling on all 1,492 failed prompt–model pairs with Gemini 4 Carbon (high reasoning effort) confirms that the primary failure taxonomy is stable: the two judges assign the same cause on 84% of pairs (Cohen's $\kappa = 0.81$) and report nearly identical

shares of visual perception errors (72% vs. 73%). Behavior-flag rates are similar or somewhat lower under the second judge (near misses 32% vs. 40%; hedging 13% vs. 19%; occlusion 16% vs. 16%), while the cross-provider hedging gap persists (26% for Anthropic vs. 10% for Google and 6% for OpenAI).

Refusals, output limits, and harness errors. Because family pass^4 requires 16 consecutive successes, a single rollout that fails without an answer lowers the score for the whole family. We report how often each non-answer failure type occurs and how much family pass^4 would change if they were counted as correct (Table 6; Figure 10). We observe the following:

- **Harness errors are rare.** They occur in only 5 of 58,752 rollouts across the 18 models in Table 6 (and 10 of 169,728 across all 52 model runs in Table A1).
- **Refusals occur only for Anthropic models and concentrate in a few families.** Claude Opus 5.5 refuses on 1.6% of rollouts, affecting 4.4% of families and lowering its family pass^4 by 4.4 pp (67.4% to 63.0%). Claude Fable 5.1 is affected similarly (1.9% of rollouts, 4.5 pp from refusals alone), while Claude Fable 5.2 refuses less often (0.6% of rollouts across 2.9% of families, 2.5 pp from refusals alone). We retain refusals as failures in our primary score as we feel that models should not refuse on the benign, basic tasks in HFE.
- **Output-limit truncations mainly affect smaller Anthropic models with very long reasoning.** Claude Haiku 5.5 reaches the output limit on 2.0% of rollouts, in 14.7% of families.
- **Counting every refusal, output-limit truncation and harness error as correct yields the largest gains for Anthropic models, yet swaps only three adjacent pairs:** GPT-6.1 Terra (#8 → #9) with Claude Fable 5.2 (#9 → #8, +3.2 pp to 76.9%), Gemini 4 Hydrogen (#10 → #11) with Claude Sonnet 5.5 (#11 → #10, +2.4 pp to 70.0%), and Gemini 3.6 Flash (#15 → #16) with Claude Haiku 5.5 (#16 → #15, +4.0 pp to 45.6%).

Table 6 | Failure types per model and their effect on family pass^4.

Model	Effort	Wrong answers (% of rollouts)	Refusals (%)	Output limit reached (%)	Harness errors (%)	Families with ≥1 refusal	Families with ≥1 output-limit or harness failure	family pass^4	family pass^4 if refusals, output limits and harness errors counted correct
GPT-6.2 Astra	xhigh	2.24%	0.00%	0.00%	0.00%	0.0%	0.0% / 0.0%	92.4%	92.4% (+0.0)
GPT-6.2 Sol	xhigh	2.73%	0.00%	0.00%	0.00%	0.0%	0.0% / 0.0%	86.9%	86.9% (+0.0)
Gemini 4 Neon	xhigh	3.03%	0.00%	0.15%	0.00%	0.0%	0.7% / 0.0%	85.8%	86.6% (+0.7)
GPT-6 Astra	xhigh	3.12%	0.00%	0.00%	0.03%	0.0%	0.0% / 0.7%	83.4%	83.4% (+0.0)
Gemini 4 Carbon	high	4.44%	0.00%	0.00%	0.00%	0.0%	0.0% / 0.0%	80.5%	80.5% (+0.0)
GPT-6.1 Sol	xhigh	3.86%	0.00%	0.00%	0.03%	0.0%	0.0% / 0.7%	79.1%	79.2% (+0.0)
Gemini 4 Argon	high	5.88%	0.00%	0.03%	0.00%	0.0%	0.7% / 0.0%	76.6%	77.1% (+0.5)
GPT-6.1 Terra	xhigh	5.30%	0.00%	0.00%	0.00%	0.0%	0.0% / 0.0%	76.4%	76.4% (+0.0)
Claude Fable 5.2	max	5.85%	0.64%	0.12%	0.00%	2.9%	1.5% / 0.0%	73.7%	76.9% (+3.2)
Gemini 4 Hydrogen	xhigh	7.14%	0.00%	0.40%	0.00%	0.0%	1.5% / 0.0%	68.9%	68.9% (+0.0)
Claude Sonnet 5.5	max	7.44%	0.28%	0.61%	0.00%	0.7%	4.4% / 0.0%	67.6%	70.0% (+2.4)
Claude Opus 5.5	max	7.14%	1.62%	0.06%	0.00%	4.4%	1.5% / 0.0%	63.0%	67.4% (+4.4)
Gemini 3.8 Flash	high	12.87%	0.00%	0.00%	0.00%	0.0%	0.0% / 0.0%	56.8%	56.8% (+0.0)
Claude Fable 5.1	max	12.19%	1.87%	0.21%	0.00%	5.1%	2.2% / 0.0%	49.7%	54.7% (+5.0)
Gemini 3.6 Flash	high	16.24%	0.00%	0.00%	0.00%	0.0%	0.0% / 0.0%	44.7%	44.7% (+0.0)
Claude Haiku 5.5	max	14.25%	0.12%	2.02%	0.00%	1.5%	14.7% / 0.0%	41.6%	45.6% (+4.0)
GPT-6 Luna	xhigh	25.18%	0.00%	0.00%	0.00%	0.0%	0.0% / 0.0%	25.6%	25.6% (+0.0)
Gemini 3.5 Flash-Lite	high	31.19%	0.00%	0.00%	0.09%	0.0%	0.0% / 2.2%	18.6%	19.1% (+0.5)

4.7. Release Subset Selection

To prevent benchmark overfitting, we release a ~20% public subset of the dataset constructed via stratified sampling across capability categories and difficulty levels. First, we apply a Two-Parameter Logistic (2PL) Item Response Theory (IRT) model (Maia Polo et al., 2024) to 84,864 evaluation outcomes, encompassing 544 prompts evaluated across an earlier release-selection pool of 26 model–effort configurations with 6 rollouts to estimate item difficulty and discriminability for each prompt.

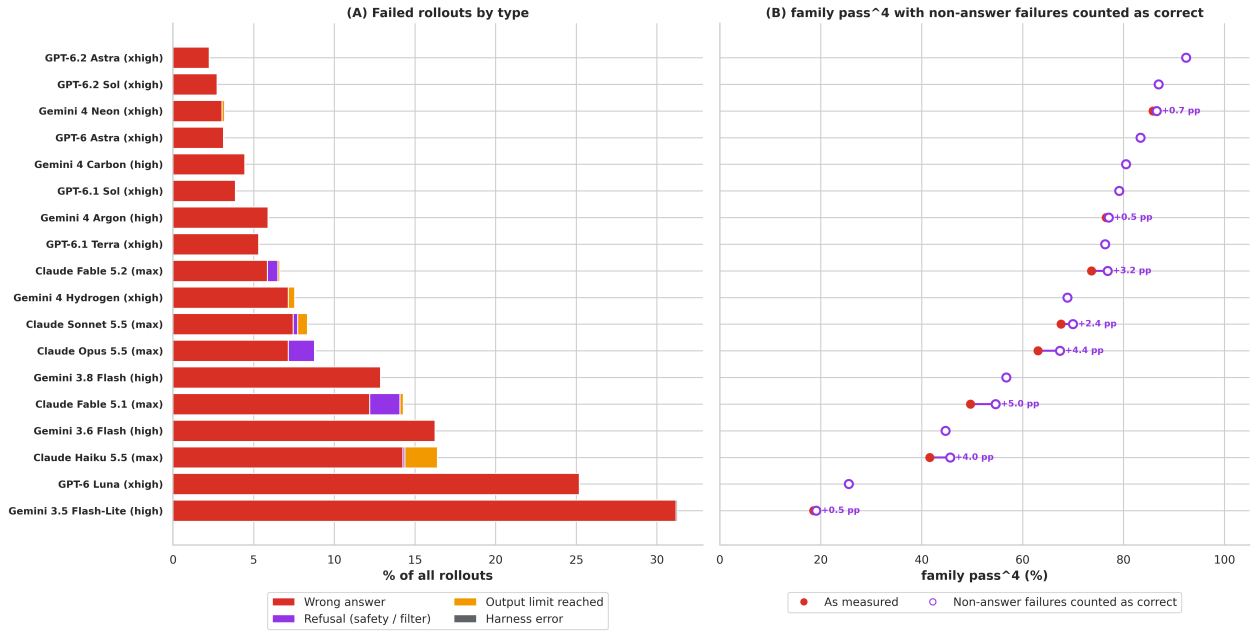


Figure 10 | Failure types and their effect on family pass[4].

We selected the public partition in two steps using the 136 seed prompts (excluding the three variants per family from the 544-prompt pool):

- **Stratified Seed Sampling:** Difficulty estimates define the sampling strata per capability category. Within each stratum, we select the top 20% of prompts ranked by discriminability score, yielding 28 representative seed prompts.
- **Public Set Expansion:** We include the 3 variants derived from each chosen seed prompt, completing a 112-prompt (28-family) public evaluation set while holding out the remaining 432 prompts (108 families) privately.

Consequently, the public sample subset exhibits strong alignment with the complete dataset, achieving a Spearman's rank correlation of 0.97.

5. Related Work

5.1. Capability Ceilings vs. Elementary Failures

Frontier evaluations largely target specialized reasoning ceilings and contamination-resistant tasks (Jain et al., 2024; Phan et al., 2025; Rein et al., 2023; Wang et al., 2024; White et al., 2024; Xu et al., 2025; Zhou et al., 2023), yet high expert scores coexist with a "jagged" profile on trivial problems (Dell'Acqua et al., 2023; Karpathy, 2024). Prior benchmarks have probed specific failure modes across character-level manipulation (Edman et al., 2024), rule induction (Hua et al., 2025), and simple commonsense or constraint-following tasks (Allcock, 2026; Nezhurina et al., 2024; SimpleBench, 2024) as well as basic visual perception (Fu et al., 2024; Li et al., 2024; Tong et al., 2024). Building on adversarial model-in-the-loop collection (Kiela et al., 2021; Nie et al., 2020), HFE unifies text and multimodal elementary tasks across 20 human-verified subcategories.

5.2. Perturbation Invariance and Repeated-Sampling Reliability

Perturbation-based benchmarks test whether models generalize across surface variations rather than memorizing specific phrasings (Mirzadeh et al., 2024), though recent studies largely perturb difficult capability-boundary problems such as competition math (Luo et al., 2026; Srivastava et al., 2024). Meanwhile, repeated-sampling studies measure multi-try coverage (pass@k; Brown et al. 2024; Chen et al. 2021) and test-time compute scaling optimizes single-output accuracy (Snell et al., 2024), rather than consistency across repeated calls. Extending multi-trial consistency (pass^k; Allcock 2026; Yao et al. 2024) across four-prompt task families (family pass^4), HFE jointly measures rollout reliability and variant robustness on elementary tasks that humans solve easily.

5.3. Reliability in Agentic and Real-World Knowledge Work

The jaggedness in model reliability we observe on elementary prompts extends to agentic tasks and real-world knowledge work. Benchmarks and deployment reports evaluating performance in these areas show that frontier models frequently fail to complete multi-step workflows end-to-end across tool use, knowledge work, and research engineering (Barres et al., 2025; OpenAI, 2026b; Patwardhan et al., 2025; Xu et al., 2024; Yao et al., 2024), and long-horizon evaluations show that reliable completion drops sharply as task length grows (Kwa et al., 2025). On various agentic benchmarks, run-to-run consistency and model robustness to perturbations have improved at a slower pace compared to overall accuracy (Rabanser et al., 2026). Whereas these studies evaluate end-to-end performance on complex agentic workflows, HFE isolates the underlying single-turn reliability floor on elementary skills where basic perceptual or procedural slips can cause downstream tasks to fail.

6. Discussion and Conclusion

6.1. Discussion

Returning to the questions posed in Section 1, frontier LLMs rarely fail elementary tasks from a lack of underlying capability: top models achieve 99.3% to 100% family pass@4, solving nearly every task family on at least one rollout. Yet their reliability floor remains well below ceiling (family pass^4 $\leq$ 92.4%, leaving a 7.6 to 80.1 pp Family Reliability Gap), with most family failures triggered by a single prompt, often the same prompt across providers (Sections 4.1 and 4.4). Scaling test-time thinking only partially closes this gap: increased reasoning effort substantially lifts family pass^4 on symbolic and text tasks by catching procedural slips, yet yields small or negligible gains on Vision (Section 4.3), which accounts for 73% of wrong answers (Section 4.6) and inverts human difficulty ordering (Spearman $\rho = -0.64$; Section 4.5). Raising baseline reliability therefore requires stronger visual perception and robustness to surface perturbations, not test-time compute alone.

6.2. Limitations

Four methodological choices bound the interpretation of our results. First, candidate seeds and variants were adversarially filtered to elicit at least one failure on our calibration panel during authoring (Sections 2.2 and 2.3), and outputs from evaluated models also informed the QA audit and the choice of which variant to drop (Section 2.5). HFE thus measures reliability on elementary tasks selected to expose model blind spots rather than a uniform sample of everyday queries, though 16 of the 18 evaluated models (Table 4), including all frontier tiers, post-date collection and perform similarly on seeds and variants (Section 4.4). Second, most multimodal prompts pair text with Gemini-generated images alongside diagrams, screenshots, and photographs. Synthetic imagery can introduce visual generation artifacts and does not span the full distribution of natural photographs, though

human raters achieved their highest accuracy on Vision seeds (84%; Section 4.5) and non-Google models achieve the highest multimodal family pass[4] (Section 4.2), showing no systematic advantage for Gemini models. Third, responses are graded by a single deterministic Gemini 3.1 Pro autorater (Section 3.2); although tasks use short, objective reference answers or rubrics and passed multi-stage quality review (Section 2.5), occasional residual prompt ambiguities, edge-case reference answers, or autorater misclassifications cannot be fully ruled out, and relying on one model family for grading could introduce subtle same-provider bias. Fourth, our human baseline evaluated seed prompts only (25% of the final benchmark, not the three sibling variants) using single-attempt, unassisted raters rather than repeated multi-variant trials (human family pass[4]). While working without tools such as calculators can occasionally cause human clerical slips even on conceptually elementary tasks, Section 4.4 shows that models fail seeds and variants at similar rates, and Section 4.5 confirms that restricting HFE strictly to the 86 families with unanimous 4/4 human agreement preserves the full reliability gap (7.0 to 80.6 pp).

6.3. Future Work

What constitutes the "reliability floor" is a moving target: as frontier capabilities advance, expectations of what counts as a "basic" model competency will rise over time, eventually including tasks that are difficult for humans, such as challenging mathematics or coding problems, where consistent reliability will increasingly be expected as routine. Tracking this rising baseline motivates three extensions of HFE:

1. Floor evaluations should be continually refreshed (for example, through open, community-driven task collection) as yesterday's capability ceiling becomes tomorrow's expected floor.
2. Whereas we restricted HFE to objectively gradable text outputs (Section 2.1), extending family pass^k to multimodal generation (such as image generation and editing, diagram synthesis, and audio) would test whether elementary counting, spatial, and attribute-binding failures similarly limit generative reliability.
3. Extending floor-reliability evaluation from standalone, single-turn prompts (Section 3.3) to agentic tasks (including tool and computer use) would measure whether models can complete everyday multi-step workflows consistently, where a single basic mistake can derail the entire task.

6.4. Conclusion

We introduced **Humanity's First Exam (HFE)**, a benchmark of 544 text and multimodal prompts (136 four-prompt task families) for measuring the foundational reliability of LLMs on elementary tasks that humans can easily solve. Across 18 models, near-perfect multi-sample solvability coexists with substantial inconsistency across repeated rollouts and minor surface variations, concentrated in basic visual perception where extra test-time thinking offers limited relief. Trustworthy deployment depends on models succeeding every time rather than merely once. Through our family-reliability methodology, 112-prompt public subset, and held-out private evaluation, we aim to encourage practices that raise the reliability floor alongside the capability ceiling.

Acknowledgments

We thank our colleagues across Google DeepMind and Kaggle for their support and feedback throughout this project, with special thanks to:

- **Sponsors:** Demis Hassabis, Logan Kilpatrick, and Tulsee Doshi.
- **Advisors:** Kate Olszewska, Apurv Suman, and Nico Duduta.
- **Operational, Editorial & Release Support:** Alison Snyder, Parth Modi, Sean Nakamoto, and Anand Rao.
- **Public Leaderboard:** The Kaggle team for hosting the public HFE leaderboard and supporting community task contributions.
- **AI Assistance:** Gemini was used for coding, dataset auditing, data analysis, and editorial assistance during the preparation of this manuscript.

References

A. Allcock. ObviousBench: The cost of not making obvious mistakes. Preprint, https://obviousbench.com (July 2026); Zenodo doi:10.5281/zenodo.21781601, 2026.

L. Alpöge. Remark [Lean 4 formalization of the Riemann Hypothesis and Generalized Riemann Hypothesis]. https://alpo.ge/remark.pdf, Nov. 20, 2026.

V. Barres, H. Dong, S. Ray, X. Si, and K. Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025.

B. Brown, J. Juravsky, R. Ehrlich, R. Clark, Q. V. Le, C. Ré, and A. Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

F. Dell'Acqua, E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working Paper 24-013, Harvard Business School, 2023.

L. Edman, H. Schmid, and A. Fraser. CUTE: Measuring LLMs' understanding of their tokens. *arXiv preprint arXiv:2409.15452*, 2024.

X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, and R. Krishna. BLINK: Multimodal large language models can see but not perceive. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. arXiv:2404.12390.

Google DeepMind. Gemini achieves gold-medal level at the International Collegiate Programming Contest World Finals. https://deepmind.google/discover/blog/gemini-achieves-gold-level-performance-at-the-international-collegiate-programming-contest-world-finals/, Sept. 17, 2025.

W. Hua, T. Wong, F. Sun, L. Pan, A. Jardine, and W. Y. Wang. InductionBench: LLMs fail in the simplest complexity class. *arXiv preprint arXiv:2502.15823*, 2025.

N. Jain, K. Han, A. Gu, W.-D. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen, and I. Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

A. Karpathy. Jagged intelligence. Public commentary on irregular LLM capability profiles, July 2024.

D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, Z. Ma, T. Thrush, S. Riedel, Z. Waseem, P. Stenetorp, R. Jia, M. Bansal, C. Potts, and A. Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021. arXiv:2104.14337.

T. Kwa, B. West, J. Becker, A. Deng, K. Garcia, M. Hasin, et al. Measuring AI ability to complete long software tasks. *arXiv preprint arXiv:2503.14499*, 2025.

B. Li, Z. Lin, W. Peng, J. d. D. Nyandwi, D. Jiang, Z. Ma, S. Khanuja, R. Krishna, G. Neubig, and D. Ramanan. NaturalBench: Evaluating vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2410.14669.

D. Luo, C. Zhang, S. Li, et al. JaggedBench: Measuring reasoning invariance via automated weakness exploration. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026.

F. Maia Polo, L. Weber, L. Choshen, Y. Sun, G. Xu, and M. Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.14992.

I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

M. Nezhurina, L. Cipolina-Kun, M. Cherti, and J. Jitsev. Alice in Wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models. *arXiv preprint arXiv:2406.02061*, 2024.

Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. arXiv:1910.14599.

OpenAI. On the Navier–Stokes Millennium Prize Problem. https://openai.com/index/navier-stokes-solution/, Sept. 8, 2026a.

OpenAI. Research acceleration: The jagged frontier. https://openai.com/index/research-acceleration-jagged-frontier/, Nov. 18, 2026b.

T. Patwardhan, R. Dias, E. Proehl, G. C. Kim, M. Wang, O. Watkins, et al. GDPval: Evaluating AI model performance on real-world economically valuable tasks. *arXiv preprint arXiv:2510.04374*, 2025.

L. Phan, A. Gatti, Z. Han, N. Li, J. Hu, H. Zhang, C. B. C. Zhang, M. Shaaban, J. Ling, S. Shi, et al. Humanity's Last Exam. *arXiv preprint arXiv:2501.14249*, 2025.

S. Rabanser, S. Kapoor, P. Kirgis, K. Liu, S. Utpala, and A. Narayanan. Towards a science of AI agent reliability. *arXiv preprint arXiv:2602.16666*, 2026.

D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

SimpleBench. SimpleBench: The text benchmark in which unspecialized human performance exceeds that of current frontier models. Technical report, https://simple-bench.com, 2024.

C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

S. Srivastava, A. M B, A. P V, S. Menon, A. Sukumar, A. Samod T, A. Philipose, S. Prince, and S. Thomas. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap. *arXiv preprint arXiv:2402.19450*, 2024.

S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes Wide Shut? Exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2401.06209.

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2406.01574.

C. White, S. Dooley, M. Roberts, A. Pal, B. Feuer, S. Jain, R. Shwartz-Ziv, N. Jain, K. Saifullah, S. Dey, S. Agrawal, S. S. Sandha, S. Naidu, C. Hegde, Y. LeCun, T. Goldstein, W. Neiswanger, and M. Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. *arXiv preprint arXiv:2406.19314*, 2024.

F. F. Xu, Y. Song, B. Li, Y. Tang, K. Jain, M. Bao, et al. TheAgentCompany: Benchmarking LLM agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*, 2024.

X. Xu, J. Zhang, T. Chen, Z. Chao, J. Hu, and C. Yang. UGMathBench: A diverse and dynamic benchmark for undergraduate-level mathematical reasoning with large language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2501.13766.

S. Yao, N. Shinn, P. Razavi, and K. Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

7. Appendix

Table A1 | Complete HFE results across all 29 models and 52 reasoning-effort configurations (n = 544 prompts, 136 families).

Rank	Model	Effort	pass@1 (±95% CI)	pass@4 (±95% CI)	pass^4 / Rel@4 (±95% CI)	family pass@4 (4×4, 95% Wilson CI)	family pass^4 (4×4, ±95% CI)	Family Reliability Gap (fam pass@4 − fam pass^4)	Avg Thought Tokens
1	GPT-6.2 Astra	xhigh	97.8% (±1.5)	97.9% (±1.5)	97.5% (±1.6)	100.0% (97.3, 100.0)	92.4% (±4.3)	7.6 pp (±4.3)	204
1	GPT-6.2 Astra	low	94.9% (±1.9)	97.9% (±1.5)	90.7% (±2.9)	100.0% (97.3, 100.0)	74.3% (±7.0)	25.7 pp (±7.0)	14
2	GPT-6.2 Sol	xhigh	97.3% (±1.5)	98.4% (±1.2)	95.8% (±1.9)	100.0% (97.3, 100.0)	86.9% (±5.2)	13.1 pp (±5.2)	274
2	GPT-6.2 Sol	low	93.1% (±2.5)	96.8% (±1.9)	87.3% (±3.7)	100.0% (97.2, 100.0)	68.7% (±7.1)	31.3 pp (±7.1)	15
3	Gemini 4 Neon	xhigh	96.8% (±1.5)	97.9% (±1.3)	95.6% (±1.9)	100.0% (97.3, 100.0)	85.8% (±5.8)	14.2 pp (±5.8)	9,852
3	Gemini 4 Neon	low	96.9% (±1.7)	97.6% (±1.4)	96.0% (±1.9)	100.0% (97.3, 100.0)	86.8% (±5.5)	13.2 pp (±5.5)	787
4	GPT-6 Astra	xhigh	96.8% (±1.6)	98.0% (±1.4)	94.9% (±2.0)	100.0% (97.3, 100.0)	83.4% (±5.8)	16.6 pp (±5.8)	382
4	GPT-6 Astra	low	96.1% (±1.8)	97.9% (±1.5)	93.8% (±2.3)	100.0% (97.3, 100.0)	80.9% (±6.3)	19.1 pp (±6.3)	65
5	Gemini 4 Carbon	high	95.6% (±1.7)	97.0% (±1.3)	93.8% (±2.2)	100.0% (97.3, 100.0)	80.5% (±6.4)	19.5 pp (±6.4)	6,616
5	Gemini 4 Carbon	low	90.9% (±2.5)	95.6% (±1.8)	84.4% (±3.9)	100.0% (97.3, 100.0)	63.1% (±7.7)	36.9 pp (±7.7)	29
6	GPT-6.1 Sol	xhigh	96.1% (±1.7)	97.8% (±1.6)	93.7% (±2.2)	100.0% (97.3, 100.0)	79.1% (±6.4)	20.9 pp (±6.4)	396
6	GPT-6.1 Sol	low	94.0% (±2.0)	96.7% (±1.7)	90.0% (±2.5)	100.0% (97.3, 100.0)	67.5% (±7.3)	32.5 pp (±7.3)	79
7	Gemini 4 Argon	high	94.1% (±2.1)	96.4% (±1.6)	91.5% (±2.8)	100.0% (97.3, 100.0)	76.6% (±6.8)	23.4 pp (±6.8)	1,838
7	Gemini 4 Argon	low	88.7% (±2.9)	93.9% (±2.3)	81.3% (±4.1)	99.8% (96.8, 100.0)	56.7% (±7.9)	43.0 pp (±7.8)	0
8	GPT-6.1 Terra	xhigh	94.7% (±2.1)	96.6% (±1.9)	91.8% (±2.7)	100.0% (97.3, 100.0)	76.4% (±6.9)	23.6 pp (±6.9)	622
8	GPT-6.1 Terra	low	91.3% (±2.5)	95.2% (±2.1)	86.3% (±3.3)	100.0% (97.3, 100.0)	62.1% (±7.7)	37.9 pp (±7.7)	89
9	Claude Fable 5.2	max	93.4% (±2.2)	95.3% (±1.9)	91.5% (±2.6)	100.0% (97.3, 100.0)	73.7% (±7.3)	26.3 pp (±7.3)	2,951
9	Claude Fable 5.2	low	91.0% (±2.5)	93.7% (±2.2)	87.7% (±3.1)	100.0% (97.3, 100.0)	65.5% (±7.7)	34.5 pp (±7.7)	125
10	Gemini 4 Hydrogen	xhigh	92.5% (±2.4)	95.6% (±1.7)	88.4% (±3.2)	100.0% (97.3, 100.0)	68.9% (±7.5)	31.1 pp (±7.5)	13,775
10	Gemini 4 Hydrogen	low	92.9% (±2.3)	95.6% (±2.0)	89.6% (±3.0)	100.0% (97.3, 100.0)	70.2% (±7.4)	29.8 pp (±7.4)	601
11	Claude Sonnet 5.5	max	91.7% (±2.4)	94.5% (±1.9)	89.2% (±2.9)	100.0% (97.3, 100.0)	67.6% (±7.7)	32.4 pp (±7.7)	6,340
11	Claude Sonnet 5.5	low	79.3% (±3.9)	86.4% (±3.4)	71.4% (±4.7)	99.3% (95.9, 99.9)	35.3% (±7.6)	63.9 pp (±7.6)	32
12	Claude Opus 5.5	max	91.2% (±2.6)	94.2% (±2.3)	87.8% (±3.0)	99.3% (96.0, 99.9)	63.0% (±7.6)	36.2 pp (±7.5)	5,162
12	Claude Opus 5.5	low	89.3% (±3.0)	92.5% (±2.5)	85.1% (±3.6)	99.8% (96.8, 100.0)	61.1% (±7.9)	38.7 pp (±7.9)	252
13	Gemini 3.8 Flash	high	87.1% (±3.2)	91.9% (±2.6)	81.7% (±4.0)	100.0% (97.3, 100.0)	56.8% (±8.1)	43.2 pp (±8.1)	8,150
13	Gemini 3.8 Flash	low	82.2% (±3.6)	90.0% (±3.0)	73.4% (±4.5)	100.0% (97.2, 100.0)	39.6% (±7.7)	60.4 pp (±7.7)	861
14	Claude Fable 5.1	max	85.7% (±3.1)	89.9% (±2.7)	80.7% (±3.8)	100.0% (97.3, 100.0)	49.7% (±8.0)	50.3 pp (±8.0)	4,665
14	Claude Fable 5.1	low	78.2% (±3.9)	86.5% (±3.1)	69.6% (±4.7)	100.0% (97.3, 100.0)	32.4% (±7.4)	67.6 pp (±7.4)	120
15	Gemini 3.6 Flash	high	83.8% (±3.4)	90.9% (±2.7)	75.8% (±4.4)	100.0% (97.3, 100.0)	44.7% (±8.0)	55.3 pp (±8.0)	3,473
15	Gemini 3.6 Flash	low	81.3% (±3.5)	90.3% (±2.6)	70.5% (±4.7)	100.0% (97.3, 100.0)	36.6% (±7.5)	63.4 pp (±7.5)	1,000
15	Gemini 3.6 Flash	minimal	55.0% (±4.8)	71.2% (±4.8)	39.6% (±5.0)	93.7% (88.3, 96.7)	8.0% (±4.0)	85.8 pp (±5.4)	0
16	Claude Haiku 5.5	max	83.6% (±3.3)	89.4% (±2.8)	76.3% (±4.0)	99.8% (96.8, 100.0)	41.6% (±7.9)	58.1 pp (±7.9)	13,351
16	Claude Haiku 5.5	low	68.7% (±3.8)	81.2% (±3.5)	53.9% (±4.4)	99.0% (95.5, 99.8)	10.9% (±4.5)	88.1 pp (±4.7)	752
17	GPT-6 Luna	xhigh	74.8% (±4.1)	85.4% (±3.7)	60.9% (±5.0)	98.5% (94.8, 99.6)	25.6% (±6.7)	73.0 pp (±6.9)	1,545
17	GPT-6 Luna	low	63.0% (±4.2)	79.0% (±4.1)	43.3% (±4.7)	98.6% (94.9, 99.6)	8.4% (±4.1)	90.2 pp (±4.3)	307
18	Gemini 3.5 Flash-Lite	high	68.7% (±4.4)	83.6% (±3.6)	52.6% (±5.3)	98.8% (95.1, 99.7)	18.6% (±6.0)	80.1 pp (±6.1)	3,094
18	Gemini 3.5 Flash-Lite	low	58.5% (±4.7)	74.4% (±4.6)	41.4% (±5.0)	95.0% (90.0, 97.6)	8.6% (±4.4)	86.4 pp (±5.3)	840
18	Gemini 3.5 Flash-Lite	minimal	37.9% (±4.1)	56.9% (±4.8)	20.8% (±3.8)	89.2% (82.9, 93.4)	0.9% (±1.2)	88.3 pp (±5.1)	0
19*	GPT-5.6 Sol	max	87.2% (±3.4)	91.5% (±3.0)	81.9% (±4.0)	98.3% (94.4, 99.5)	54.2% (±7.9)	44.0 pp (±8.0)	1,653
20*	GPT-5.5	xhigh	84.5% (±3.6)	89.4% (±2.9)	79.2% (±4.3)	99.9% (97.0, 100.0)	51.7% (±8.0)	48.2 pp (±8.0)	2,299
21*	Claude Opus 5	max	81.3% (±4.0)	85.8% (±3.6)	76.8% (±4.5)	99.8% (96.8, 100.0)	47.3% (±8.1)	52.4 pp (±8.1)	2,326
22*	Gemini 3.1 Pro	high	83.6% (±3.5)	92.2% (±2.5)	73.1% (±4.8)	100.0% (97.3, 100.0)	43.6% (±7.7)	56.4 pp (±7.7)	2,834
22*	Gemini 3.1 Pro	low	83.3% (±3.6)	91.6% (±2.5)	73.3% (±5.0)	100.0% (97.3, 100.0)	46.1% (±8.0)	53.9 pp (±8.0)	1,219
23*	GPT-5.6 Terra	max	81.6% (±3.7)	88.8% (±3.3)	72.6% (±4.6)	99.2% (95.9, 99.9)	39.0% (±7.7)	60.2 pp (±7.7)	3,501
24*	Claude Opus 4.8	max	71.3% (±5.0)	80.0% (±4.2)	62.0% (±5.9)	99.1% (95.7, 99.8)	36.9% (±7.9)	62.2 pp (±7.9)	4,613
25*	Claude Sonnet 5	max	66.6% (±5.5)	76.2% (±4.8)	57.1% (±6.3)	97.7% (93.6, 99.2)	32.9% (±7.8)	64.8 pp (±7.7)	12,225
26*	Gemini 3 Flash	high	77.3% (±3.9)	88.5% (±3.0)	64.9% (±5.0)	100.0% (97.3, 100.0)	31.7% (±7.3)	68.3 pp (±7.3)	4,071
26*	Gemini 3 Flash	low	74.5% (±4.1)	86.7% (±3.3)	60.7% (±5.2)	99.0% (95.6, 99.8)	26.0% (±6.9)	73.1 pp (±7.0)	1,016
27*	GPT-5.6 Luna	max	75.9% (±4.0)	85.4% (±3.5)	64.4% (±4.9)	98.8% (95.2, 99.7)	28.6% (±7.2)	70.2 pp (±7.2)	4,198
28*	Claude Sonnet 4.6	max	60.0% (±5.7)	68.8% (±5.4)	50.6% (±6.2)	92.9% (87.3, 96.1)	24.6% (±7.0)	68.3 pp (±7.5)	6,240
29*	Gemini 3.1 Flash-Lite	high	64.7% (±4.8)	79.0% (±4.3)	49.2% (±5.5)	96.9% (92.5, 98.8)	18.3% (±6.1)	78.6 pp (±6.5)	3,381
29*	Gemini 3.1 Flash-Lite	low	39.0% (±3.8)	59.4% (±4.7)	20.9% (±3.6)	92.2% (86.4, 95.6)	0.0% (±0.0)	92.2 pp (±4.1)	126

Thought-token counts reflect each provider's native tokenizer and API reporting (e.g., OpenAI `reasoning_tokens`; Gemini 4 Argon low reports 0 thought tokens) and are not directly comparable across providers. Models are ordered by descending family pass^4 at highest evaluated reasoning effort (current tier first, then prior-generation models marked with *). Reported ± values denote half-widths (hi − lo)/2 of the 95% family-clustered percentile bootstrap interval. * Prior-generation models, reported for reference.

Table A2 | Effect of reasoning effort on reliability. Each model is shown at its highest evaluated effort (`high/max/xhigh`), at `low` effort, and, where available, at `minimal` effort (0 thought tokens).

Model	Effort	pass@1 (±CI)	pass@4 (±CI)	pass^4 / Rel@4 (±CI)	family pass@4	family pass^4 (±CI)	Δ family pass^4, this effort − low	Δ family pass^4, this effort − minimal	Avg Thought Tokens	Math fam pass^4 (N = 160)	Reas fam pass^4 (N = 72)	Spatial fam pass^4 (N = 108)	Vision fam pass^4 (N = 204)
GPT-6.2 Astra	xhigh	97.8% (±1.5)	97.9% (±1.5)	97.5% (±1.6)	100.0%	**92.4%** (±4.3)	**+18.1 pp**	—	204	98.3%	88.9%	90.1%	90.2%
GPT-6.2 Astra	low	94.9% (±1.9)	97.9% (±1.5)	90.7% (±2.9)	100.0%	74.3% (±7.0)	—	—	14	57.2%	88.9%	72.3%	83.7%
GPT-6.2 Sol	xhigh	97.3% (±1.5)	98.4% (±1.2)	95.8% (±1.9)	100.0%	**86.9%** (±5.2)	**+18.2 pp**	—	274	96.0%	83.7%	85.6%	81.7%
GPT-6.2 Sol	low	93.1% (±2.5)	96.8% (±1.9)	87.3% (±3.7)	100.0%	68.7% (±7.1)	—	—	15	57.2%	59.6%	62.3%	84.3%
Gemini 4 Neon	xhigh	96.8% (±1.5)	97.9% (±1.3)	95.6% (±1.9)	100.0%	**85.8%** (±5.8)	**−1.0 pp**	—	9,852	97.5%	83.3%	86.4%	77.3%
Gemini 4 Neon	low	96.9% (±1.7)	97.6% (±1.4)	96.0% (±1.9)	100.0%	86.8% (±5.5)	—	—	787	95.8%	88.9%	81.5%	81.8%
GPT-6 Astra	xhigh	96.8% (±1.6)	98.0% (±1.4)	94.9% (±2.0)	100.0%	**83.4%** (±5.8)	**+2.5 pp**	—	382	92.5%	88.9%	84.0%	73.9%
GPT-6 Astra	low	96.1% (±1.8)	97.9% (±1.5)	93.8% (±2.3)	100.0%	80.9% (±6.3)	—	—	65	91.7%	85.6%	72.2%	75.4%
Gemini 4 Carbon	high	95.6% (±1.7)	97.0% (±1.3)	93.8% (±2.2)	100.0%	**80.5%** (±6.4)	**+17.4 pp**	—	6,616	97.5%	83.7%	73.1%	69.9%
Gemini 4 Carbon	low	90.9% (±2.5)	95.6% (±1.8)	84.4% (±3.9)	100.0%	63.1% (±7.7)	—	—	29	55.1%	63.1%	67.9%	66.8%
GPT-6.1 Sol	xhigh	96.1% (±1.7)	97.8% (±1.6)	93.7% (±2.2)	100.0%	**79.1%** (±6.4)	**+11.6 pp**	—	396	93.3%	80.4%	78.4%	67.9%
GPT-6.1 Sol	low	94.0% (±2.0)	96.7% (±1.7)	90.0% (±2.5)	100.0%	67.5% (±7.3)	—	—	79	84.7%	67.0%	57.0%	59.7%
Gemini 4 Argon	high	94.1% (±2.1)	96.4% (±1.6)	91.5% (±2.8)	100.0%	**76.6%** (±6.8)	**+19.9 pp**	—	1,838	87.0%	88.9%	71.6%	66.7%
Gemini 4 Argon	low	88.7% (±2.9)	93.9% (±2.3)	81.3% (±4.1)	99.8%	56.7% (±7.9)	—	—	0	47.8%	52.8%	53.3%	66.9%
GPT-6.1 Terra	xhigh	94.7% (±2.1)	96.6% (±1.9)	91.8% (±2.7)	100.0%	**76.4%** (±6.9)	**+14.2 pp**	—	622	86.6%	74.2%	72.0%	71.4%
GPT-6.1 Terra	low	91.3% (±2.5)	95.2% (±2.1)	86.3% (±3.3)	100.0%	62.1% (±7.7)	—	—	89	53.9%	66.8%	58.4%	68.9%
Claude Fable 5.2	max	93.4% (±2.2)	95.3% (±1.9)	91.5% (±2.6)	100.0%	**73.7%** (±7.3)	**+8.2 pp**	—	2,951	83.3%	83.3%	74.3%	62.4%
Claude Fable 5.2	low	91.0% (±2.5)	93.7% (±2.2)	87.7% (±3.1)	100.0%	65.5% (±7.7)	—	—	125	77.7%	76.3%	53.8%	58.3%
Gemini 4 Hydrogen	xhigh	92.5% (±2.4)	95.6% (±1.7)	88.4% (±3.2)	100.0%	**68.9%** (±7.5)	**−1.3 pp**	—	13,775	93.3%	83.7%	62.4%	47.9%
Gemini 4 Hydrogen	low	92.9% (±2.3)	95.6% (±2.0)	89.6% (±3.0)	100.0%	70.2% (±7.4)	—	—	601	85.5%	88.9%	50.4%	62.1%
Claude Sonnet 5.5	max	91.7% (±2.4)	94.5% (±1.9)	89.2% (±2.9)	100.0%	**67.6%** (±7.7)	**+32.3 pp**	—	6,340	85.3%	78.1%	56.8%	55.7%
Claude Sonnet 5.5	low	79.3% (±3.9)	86.4% (±3.4)	71.4% (±4.7)	99.3%	35.3% (±7.6)	—	—	32	43.9%	34.5%	22.1%	35.9%
Claude Opus 5.5	max	91.2% (±2.6)	94.2% (±2.3)	87.8% (±3.0)	99.3%	**63.0%** (±7.6)	**+2.0 pp**	—	5,162	75.2%	79.6%	55.8%	51.5%
Claude Opus 5.5	low	89.3% (±3.0)	92.5% (±2.5)	85.1% (±3.6)	99.8%	61.1% (±7.9)	—	—	252	72.8%	76.3%	50.8%	51.9%
Gemini 3.8 Flash	high	87.1% (±3.2)	91.9% (±2.6)	81.7% (±4.0)	100.0%	**56.8%** (±8.1)	**+17.2 pp**	—	8,150	76.8%	69.6%	46.4%	42.0%
Gemini 3.8 Flash	low	82.2% (±3.6)	90.0% (±3.0)	73.4% (±4.5)	100.0%	39.6% (±7.7)	—	—	861	44.7%	30.5%	31.6%	43.0%
Claude Fable 5.1	max	85.7% (±3.1)	89.9% (±2.7)	80.7% (±3.8)	100.0%	**49.7%** (±8.0)	**+17.2 pp**	—	4,665	70.8%	81.5%	33.4%	30.5%
Claude Fable 5.1	low	78.2% (±3.9)	86.5% (±3.1)	69.6% (±4.7)	100.0%	32.4% (±7.4)	—	—	120	45.8%	59.3%	18.8%	19.7%
Gemini 3.6 Flash	high	83.8% (±3.4)	90.9% (±2.7)	75.8% (±4.4)	100.0%	**44.7%** (±8.0)	**+8.1 pp**	**+36.8 pp**	3,473	69.6%	48.5%	34.6%	29.2%
Gemini 3.6 Flash	low	81.3% (±3.5)	90.3% (±2.6)	70.5% (±4.7)	100.0%	36.6% (±7.5)	—	+28.6 pp	1,000	56.5%	28.8%	24.0%	30.5%
Gemini 3.6 Flash	minimal	55.0% (±4.8)	71.2% (±4.8)	39.6% (±5.0)	93.7%	8.0% (±4.0)	—	—	0	2.2%	5.9%	4.8%	14.9%
Gemini 3.1 Pro*	high	83.6% (±3.5)	92.2% (±2.5)	73.1% (±4.8)	100.0%	**43.6%** (±7.7)	**−2.5 pp**	—	2,834	80.9%	43.3%	31.1%	21.1%
Gemini 3.1 Pro*	low	83.3% (±3.6)	91.6% (±2.5)	73.3% (±5.0)	100.0%	46.1% (±8.0)	—	—	1,219	73.7%	42.0%	42.1%	28.1%
Claude Haiku 5.5	max	83.6% (±3.3)	89.4% (±2.8)	76.3% (±4.0)	99.8%	**41.6%** (±7.9)	**+30.7 pp**	—	13,351	70.3%	37.8%	31.0%	26.1%
Claude Haiku 5.5	low	68.7% (±3.8)	81.2% (±3.5)	53.9% (±4.4)	99.0%	10.9% (±4.5)	—	—	752	21.1%	9.3%	7.4%	5.4%
Gemini 3 Flash*	high	77.3% (±3.9)	88.5% (±3.0)	64.9% (±5.0)	100.0%	**31.7%** (±7.3)	**+5.7 pp**	—	4,071	66.9%	17.5%	25.2%	12.4%
Gemini 3 Flash*	low	74.5% (±4.1)	86.7% (±3.3)	60.7% (±5.2)	99.0%	26.0% (±6.9)	—	—	1,016	48.6%	13.3%	16.5%	17.7%
GPT-6 Luna	xhigh	74.8% (±4.1)	85.4% (±3.7)	60.9% (±5.0)	98.5%	**25.6%** (±6.7)	**+17.2 pp**	—	1,545	30.0%	52.2%	9.2%	21.3%
GPT-6 Luna	low	63.0% (±4.2)	79.0% (±4.1)	43.3% (±4.7)	98.6%	8.4% (±4.1)	—	—	307	13.2%	11.5%	1.5%	7.1%
Gemini 3.5 Flash-Lite	high	68.7% (±4.4)	83.6% (±3.6)	52.6% (±5.3)	98.8%	**18.6%** (±6.0)	**+10.0 pp**	**+17.7 pp**	3,094	46.5%	17.2%	5.7%	4.1%
Gemini 3.5 Flash-Lite	low	58.5% (±4.7)	74.4% (±4.6)	41.4% (±5.0)	95.0%	8.6% (±4.4)	—	+7.7 pp	840	27.8%	0.5%	0.4%	0.8%
Gemini 3.5 Flash-Lite	minimal	37.9% (±4.1)	56.9% (±4.8)	20.8% (±3.8)	89.2%	0.9% (±1.2)	—	—	0	0.2%	5.6%	0.2%	0.2%
Gemini 3.1 Flash-Lite*	high	64.7% (±4.8)	79.0% (±4.3)	49.2% (±5.5)	96.9%	**18.3%** (±6.1)	**+18.3 pp**	—	3,381	48.6%	13.6%	2.1%	4.7%
Gemini 3.1 Flash-Lite*	low	39.0% (±3.8)	59.4% (±4.7)	20.9% (±3.6)	92.2%	0.0% (±0.0)	—	—	126	0.0%	0.0%	0.0%	0.0%

The Δ columns give the change in family pass^4 of that row's effort level relative to the same model's `low` or `minimal` run. * Prior-generation models, reported for reference. Models are ordered by descending family pass^4 at highest evaluated reasoning effort. Reported ± values denote half-widths (hi − lo)/2 of the 95% family-clustered percentile bootstrap interval.

Table A3 | Subcategory sensitivity to thinking effort (20 subcategories ranked by mean Δ family pass[4], high/max/xhigh − low, across 18 models).

Rank	Subcategory	Macro-Category	Families	Mean Δ family pass[4] (High/Max/xHigh − Low, 18 Models)
1	**Logical Deduction**	Reasoning & Language	2	**+29.7 pp**
2	**Constraint Satisfaction**	Math & Algorithmic	13	**+28.5 pp**
3	**Algorithmic and Symbolic Execution**	Math & Algorithmic	8	**+26.2 pp**
4	**Sequence Manipulation**	Math & Algorithmic	6	**+26.0 pp**
5	**Sorting and Ordering**	Math & Algorithmic	5	**+21.6 pp**
6	**General Arithmetic and Calculation**	Math & Algorithmic	3	+19.6 pp
7	**Negative Fact Verification**	Reasoning & Language	3	+19.6 pp
8	**Positional Tracking**	Spatial & Positional	6	+18.7 pp
9	**Causal and Counterfactual Reasoning**	Reasoning & Language	2	+18.3 pp
10	**Generalization and Pattern Recognition**	Reasoning & Language	4	+18.3 pp
11	**Spatial and Topological Reasoning**	Spatial & Positional	4	+15.8 pp
12	**Commonsense and Physical Reasoning**	Reasoning & Language	4	+13.8 pp
13	**Spatial Reasoning**	Spatial & Positional	17	+13.1 pp
14	**Pragmatic Reasoning**	Reasoning & Language	3	+10.7 pp
15	**Text & Symbol Recognition**	Vision	11	+5.7 pp
16	**Graph & Chart Understanding**	Vision	7	+5.0 pp
17	**Mathematical and Quantitative Reasoning**	Math & Algorithmic	5	+4.9 pp
18	**Color Identification**	Vision	10	+4.9 pp
19	**Arrows, pointers, or indicators**	Vision	11	+2.0 pp
20	**Object Counting**	Vision	12	+1.9 pp

Example Four-Prompt Task Family (Public Split)

To illustrate how HFE pairs each seed prompt with three logic-preserving variants, Figure A1 shows the complete four-prompt Color Identification family highlighted in Figure 2 (right), whose seed was solved by 4/4 human raters.

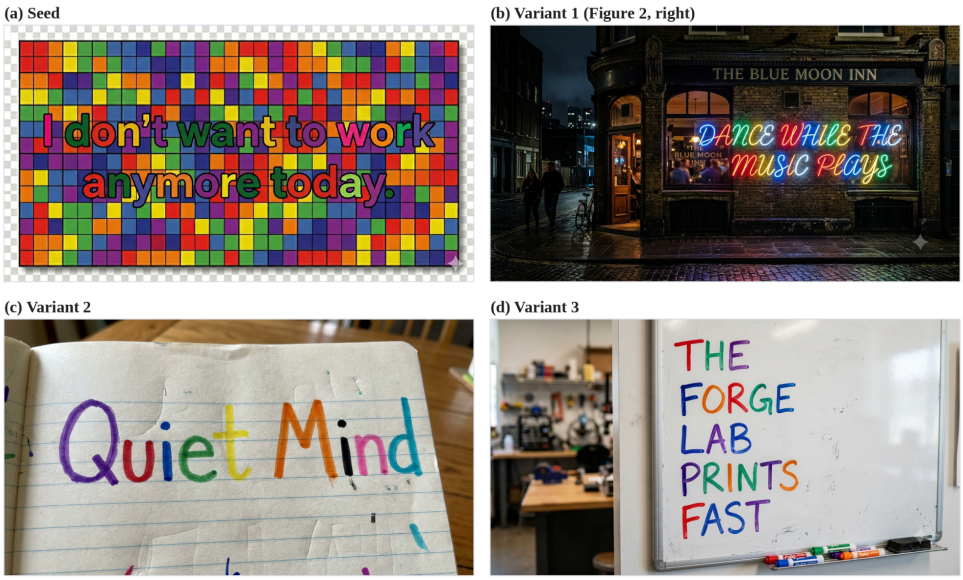


Figure A1 | All four images for the Color Identification task family (public split). Table A4 lists the corresponding prompts and ground-truth answers.

Table A4 | Prompts and ground-truth answers for the four-prompt Color Identification family in Figure A1.

Panel	Prompt	Ground Truth
(a) Seed	What color is the second letter in every word? If the word only has one letter, ignore that word. Return your answer by listing only the identified colors from the beginning of the sentence to the end.	Red; Green; Purple; Orange; Green; Red
(b) Variant 1 (Fig. 2, right)	List the color of the third letter of each word in the neon sign in order, separated by a semicolon.	green; yellow; white; blue; blue
(c) Variant 2	List the colors of every vowel in the phrase, separated by commas.	red, blue, green, black
(d) Variant 3	List the color of the second consonant of every word on the whiteboard, in order from top to bottom, separated by a comma.	green, red, blue, green, blue

Autorater Grading Prompt

Below is the verbatim grading prompt template used by the deterministic Gemini 3.1 Pro autorater (Section 3.2), evaluated after stripping internal reasoning traces and control tokens.

```
[Task]: You are an expert tasked with grading a puzzle solution.
You will be provided with:
1. A question that was used as a prompt for the model.
2. A model's response
3. A reference answer (ground truth).

Your role is to determine whether the final answer in the model's response is equivalent
to the reference answer. Respond only with Yes or No.

[Guidelines for Equivalence]:
1) Answer Choices and Values: If the reference answer is a choice (e.g., B), and model's
response is enclosed in brackets (e.g., (B), [B], or $\boxed[B]$, consider the answer to
be correct (Yes).
2) Numerics: If the reference answer is a numeric value (e.g., 10), and the model's
response is the same numeric value spelled out in words (e.g., ten), or vice-versa,
consider the answer to be correct (Yes).
3) Temporal quantities: If the reference answer is a temporal quantity at a certain
resolution (e.g., 3 hours), and the model's response is at a different resolution (e.g.,
180 minutes) but the two are equivalent (e.g., 3 hours is 180 minutes), consider the
answer to be correct (Yes).
4) Rubrics: If the ground truth is a rubric (e.g., the response must contain exactly 10
words) and the model's response satisfies all the criteria in the rubric, consider the
answer to be correct (Yes).

[Input]:
Question:
{prompt}
--------------------
Model response:
{response}
--------------------
Ground truth:
{target}

[Output]:
Is the final answer in the model's response equivalent to the reference answer, answer
Yes or No:
```

Full 18-Model Family Reliability Gap

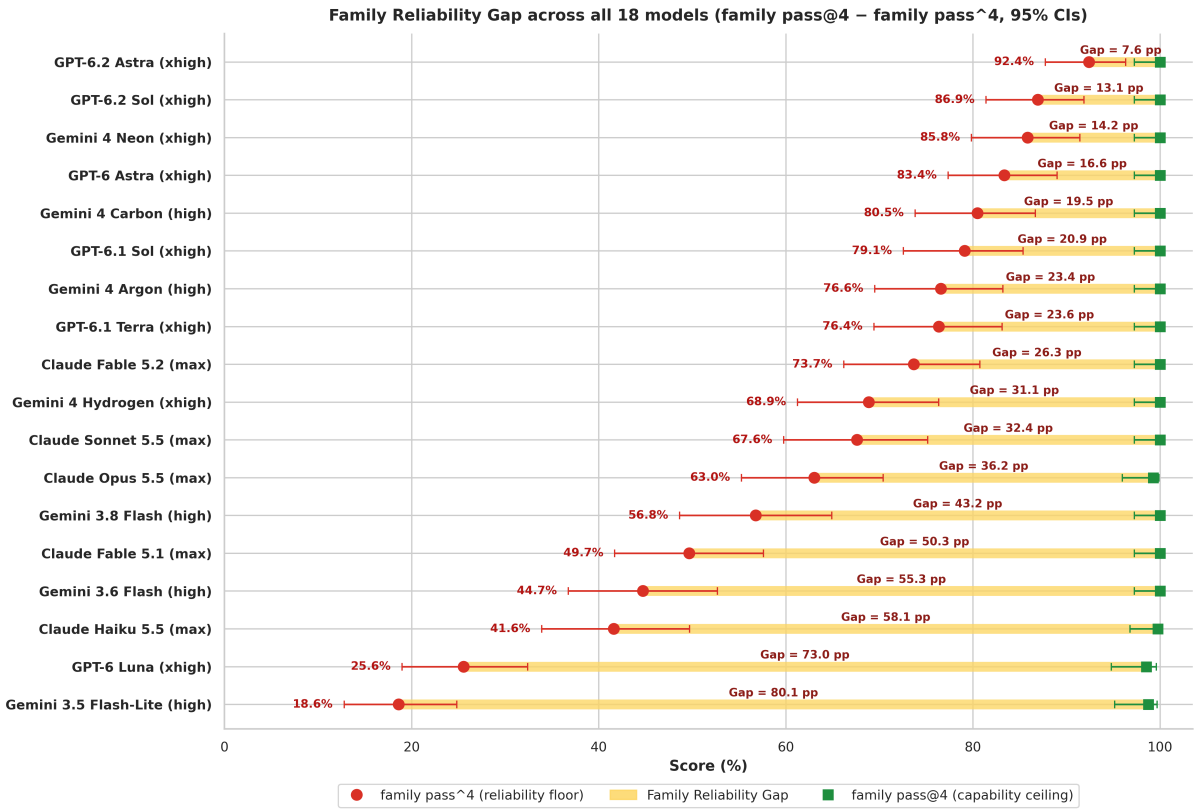


Figure A2 | Family Reliability Gap (family pass@4 – family pass^4, with 95% CIs) across all 18 models in Table 4 at highest evaluated reasoning effort.