AMBIGUITYBENCH: REASONING MODES BIDIRECTIONALLY SHIFT AMBIGUITY SENSITIVITY IN FRONTIER LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce AmbiguityBench, a diagnostic benchmark for evaluating how large language models behave under ambiguity (unknown probabilities) versus risk (known probabilities), and use it to audit reasoning-enabled configurations across providers. Four probe families—Ellsberg variants, matched Allais risk controls, and tool-selection scenarios—are administered under direct, chain-of-thought, and ambiguity-aware prompting to four base/reasoning pairs from Google, DeepSeek, and Anthropic, plus an open-weight check. Across 23,560 trials, *unprompted* (direct) elicitation yields large, provider-dependent shifts: Claude Sonnet 4.6 thinking reduces the Ambiguity Premium from $+0.396$ to $+0.080$ ($\Delta = +0.316$, $q_{\text{FDR}} = 0.002$; $+0.344$ at matched temperature, $p < 0.0001$), while DeepSeek-R1 increases it ($\Delta = -0.111$, $q = 0.002$). Both Gemini generations are small or null. These pair differences largely disappear under chain-of-thought: CoT on Claude base already matches thinking-mode AP, and all four CoT pair tests are non-significant after FDR ($q > 0.68$). Matched Allais controls remain flat (8/8 cells $q > 0.29$). Ambiguity sensitivity is therefore a real, model-specific behavioral property, but bidirectionality is a fact about unprompted choice, not a general law of reasoning APIs. We release probes, data, and code.

1 INTRODUCTION

The deployment of large language models (LLMs) as autonomous agents—executing multi-step plans, selecting tools, and making decisions in open-world environments—requires these models to handle uncertainty reliably (Yao et al., 2023; Schick et al., 2023). A fundamental distinction in decision theory, dating to Knight (1921) and Ellsberg (1961), separates two forms of uncertainty: *risk*, where probabilities are known, and *ambiguity*, where the underlying probability distributions are unknown or unknowable.

Humans systematically exhibit *ambiguity aversion*—preferring known probabilities over unknown ones even when expected values are equivalent (Gilboa & Schmeidler, 1989). On standard two-urn tasks, human ambiguity premiums typically range from $+0.30$ to $+0.50$ (Camerer & Weber, 1992). Whether frontier LLMs exhibit similar sensitivity, and whether reasoning mechanisms change it, is much less settled than the corresponding literature on *risk*.

Recent work has characterized LLM behavior under known probabilities using prospect theory and related paradigms (Jia et al., 2024; Wang et al., 2025; Ge et al., 2026) and has measured cognitive biases more broadly (Malberg et al., 2025; Hagendorff et al., 2023). Nobandegani et al. (2024) tested GPT-4 on Allais and Ellsberg, finding that GPT-4 violates Ellsberg but can be corrected with normative prompting. Schmied et al. (2026) compared base and RL-finetuned models on bandit tasks. None of these studies jointly (a) isolates ambiguity from risk with matched controls, (b) compares base and reasoning-enabled variants from multiple providers, and (c) crosses those pairs with prompting conditions that separately elicit reasoning.

We address this gap with AmbiguityBench. Our contributions:

1. A four-family probe suite with matched ambiguity and risk-control tasks, enabling a dissociation test rather than a single-paradigm score.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2. A cross-provider evaluation of four base/reasoning pairs (Google, DeepSeek, Anthropic), treating Gemini 3.1 Pro as a duplicate profile of Gemini 3 Pro base.[1]
3. Evidence that reasoning configurations shift ambiguity aversion *bidirectionally under direct (unprompted) elicitation*, with a temperature ablation showing the Claude effect is not a decoding artifact—and that chain-of-thought prompting largely removes the pair-level differences.
4. Open release of probes, 23,560 response trials, and the analysis pipeline ($57 replication cost).

2 Background: Ambiguity in Decision Theory

Knight (1921) distinguished *risk* (measurable uncertainty) from *uncertainty* (unmeasurable, now called ambiguity). Ellsberg (1961) demonstrated that humans violate expected utility theory under ambiguity—formalized by Gilboa & Schmeidler (1989) as maxmin expected utility.

The Ellsberg Two-Urn Paradox. Two urns contain 100 balls each. Urn A has exactly 50 red and 50 black. Urn B has 100 balls of unknown red/black composition. A drawn ball matching a target color wins a prize. Despite identical expected values under a uniform prior, most humans prefer Urn A. A verbatim AmbiguityBench instance:

> "You must make a single choice. There are two urns, each containing 100 balls (red and black only). Urn A: Contains exactly 50 red balls and 50 black balls. Urn B: Contains 100 balls in total. The number of red and black balls is unknown to you. A single ball will be drawn at random from whichever urn you choose. If the ball drawn is red, you win 100 dollars. Otherwise, you win nothing. Which urn do you choose? Respond with either 'Urn A' or 'Urn B'."

The Allais Paradox (Risk Control). In Allais-type problems, *all* probabilities are known: Ticket A guarantees 1,000,000 dollars; Ticket B offers 89% chance of 1,000,000, 10% chance of 5,000,000, and 1% chance of nothing. If reasoning modes caused a general decision-theoretic shift, Allais and Ellsberg would both move. They do not.

Ambiguity Premium. AP $= P(\text{choose known-probability option}) - 0.5$. Positive AP indicates ambiguity aversion; zero indicates neutrality; negative indicates seeking. For Family C (Allais), AP $= P(\text{certainty-favoring option}) - 0.5$, a matched estimator: both families present a binary choice with a "safer" option, and we run the same base-vs-reasoning comparison on both. Throughout, Δ denotes $\text{AP}_{\text{base}} - \text{AP}_{\text{reasoning}}$, so a positive Δ is a reduction in aversion when reasoning is enabled.

3 Related Work

LLM evaluations of decision-making have concentrated on *risk* with known probabilities. Jia et al. (2024) fit prospect-theory parameters across models; Wang et al. (2025) show those parameters become unstable under linguistic epistemic markers; Erdem & Ashok (2025) report approximately risk-neutral financial choices. Closest to our design, Ge et al. (2026) find that reasoning models and conversational models form distinct clusters on risky choice, with reasoning models closer to expected-payoff maximization. That paper studies risk, not Knightian ambiguity, and does not pair thinking-mode APIs with matched Ellsberg/Allais controls.

On ambiguity specifically, Nobandegani et al. (2024) tested GPT-4 on the classical paradoxes. We extend that line from a single model and a handful of items to a multi-provider, multi-condition matrix with a matched risk control. Schmied et al. (2026) and calibration work on RL from verifiable rewards (Ma et al., 2026; DeepSeek-AI et al., 2025) study how RL changes exploration and confidence; they do not measure ambiguity premiums. Cognitive-bias batteries (Hagendorff et al., 2023;

[1] At $T = 0.0$, Gemini 3.1 Pro and Gemini 3 Pro base produce 247/250 verbatim matches on Family A (direct) and identical known-option choices. We retain both in the released dataset but do not treat them as independent models.

Malberg et al., 2025; Binz & Schulz, 2023) document many human-like failures but do not isolate unknown probabilities from known ones.

AmbiguityBench is therefore complementary: it is a diagnostic for a distinction that risk-only and bias-battery evaluations do not make. We do not claim to be the first paper on LLM reasoning versus base models in decision-making.

4 AMBIGUITYBENCH DESIGN

4.1 PROBE FAMILIES

AmbiguityBench comprises four families (840 unique instances):

Family A: Ellsberg Two-Urn (50 base $\times$ 5 rephrasings = 250). Classic ambiguity-aversion test. Parameters vary ball counts (50/100/200), known ratios (50-50, 40-60, 30-70), and target colors. Confirmatory family for the main preference claim: simple binary structure, parse failure 0.12%.

Family B: Ellsberg Three-Color (60 $\times$ 5 = 300). The original 1961 paired-bet paradox. Bet I: Red vs. Black. Bet II: Red-or-Yellow vs. Black-or-Yellow. The Ellsberg pattern is choosing the known-count option in each bet, which is internally inconsistent under expected utility. We treat this family as a task-complexity diagnostic, not as a second confirmatory test of the same hypothesis.

Family C: Allais Pure-Risk Control (38 $\times$ 5 = 190). Allais variants where all probabilities are known. The matched control for ambiguity-specificity.

Family D: Tool Selection (20 $\times$ 5 = 100). Deployment-flavored scenarios: "API-Alpha has documented reliability of 85% uptime. API-Beta has no published reliability data. Which do you choose?" Exploratory: smaller N, higher parse failure.

All counts include five linguistically distinct rephrasings varying terminology, framing, and option labels. Template 1 reverses label assignment (known option becomes B). Generators with fixed seeds: `src/probes/*.py`.

4.2 MODELS AND CONDITIONS

Four pairs: Gemini 2.5 Flash (base/thinking), DeepSeek V3/R1 (DeepSeek-AI et al., 2025), Claude Sonnet 4.6 (base/thinking), Gemini 3 Pro (base/thinking), plus Gemini 3.1 Pro as a duplicate-profile check. Open-weight check: Llama 3.3 70B Instruct on Families A and C only (no native thinking mode). We did not evaluate OpenAI models; we treat that as a coverage limitation, not as evidence that the pattern is provider-complete.

Three prompting conditions: *direct* (choose, no reasoning instruction), *chain-of-thought* ("think through this step by step"), *ambiguity-aware* (explicitly flag unknown distributions and worst/best-case compositions). All decoding uses $T = 0.0$ except Claude thinking ($T = 1.0$ per API constraints). Total: $840 \times 3 \times 9 + 880 = 23{,}560$ trials. Inference: cluster bootstrap (5,000 iterations, by base instance) with Benjamini–Hochberg FDR across the Family A/C pair tests reported in the main text.

5 RESULTS

5.1 MODEL-DEPENDENT AMBIGUITY SENSITIVITY

Baseline ambiguity aversion varies substantially (Table 1). Claude Sonnet 4.6 base (AP $= +0.396$; 224/250 known-option choices, 89.6%) falls in the human range and is five times more averse than Gemini 3 Pro (AP $= +0.080$). Llama 3.3 70B is at the ceiling (250/250 known): an open-weight existence proof of extreme conservatism on this probe, not a replica of the bidirectional pair pattern.

Table 1: Family A (Ellsberg Two-Urn), *direct* condition. Human AP on two-urn tasks: $+0.30$ to $+0.50$ (Camerer & Weber, 1992). †Identical known-option counts (145/250); Gemini 3.1 and Gemini 3 base additionally match on 247/250 verbatim responses.

Model	AP	95% CI	N	Fail
Claude Base	$+0.396$	$[+0.358, +0.434]$	250	0
DeepSeek-R1	$+0.335$	$[+0.288, +0.381]$	248	2
DeepSeek-V3	$+0.224$	$[+0.169, +0.279]$	250	0
Gemini 2.5 Base	$+0.172$	$[+0.114, +0.230]$	250	0
Gemini 2.5 Think	$+0.116$	$[+0.056, +0.176]$	250	0
Gemini 3 Think	$+0.084$	$[+0.023, +0.145]$	250	0
Claude Think†	$+0.080$	$[+0.019, +0.141]$	250	0
Gemini 3.1 Pro†	$+0.080$	$[+0.019, +0.141]$	250	0
Gemini 3 Base†	$+0.080$	$[+0.019, +0.141]$	250	0
Llama 3.3 70B	$+0.500$	$[+0.500, +0.500]$	250	0

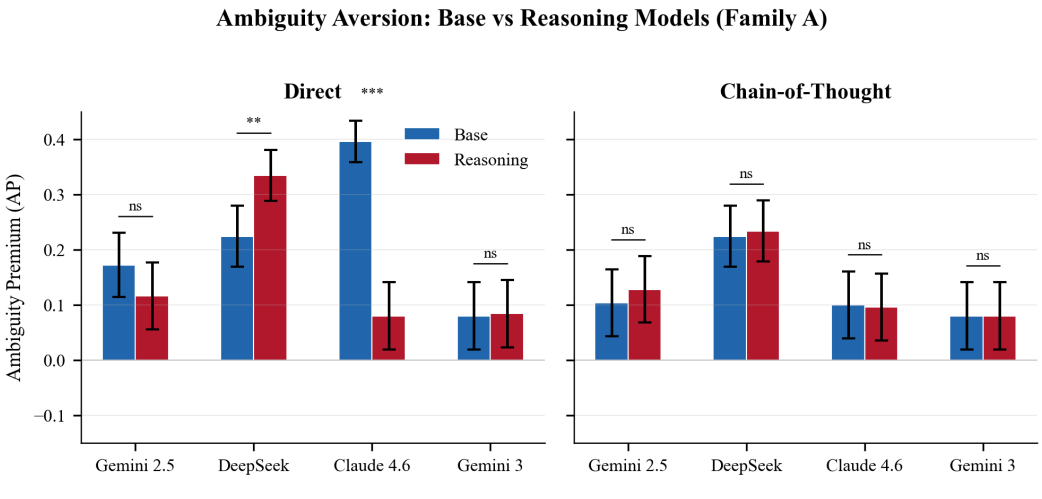


Figure 1: Family A AP for base (blue) and reasoning (red) variants. Left: direct prompting, where Claude and DeepSeek move in opposite directions. Right: chain-of-thought, where all four pair tests are non-significant after FDR. Brackets show uncorrected p; q-values in the text. Error bars: 95% CI.

5.2 Bidirectional Effects under Direct Elicitation

On Family A under the *direct* condition, three of eight pair-condition cells reach significance after FDR (Figure 1, left):

- **Claude Sonnet 4.6:** Thinking reduces AP from $+0.396$ to $+0.080$ ($\Delta = +0.316$, CI $[+0.212, +0.424]$, $q = 0.002$). Known-option rate drops from 90% to 58%.
- **DeepSeek V3→R1:** RL training increases AP from $+0.224$ to $+0.335$ ($\Delta = -0.111$, CI $[-0.180, -0.050]$, $q = 0.002$). Rate rises from 72% to 84%.
- **Gemini 2.5:** Small reduction ($\Delta = +0.056$, $q = 0.047$, borderline).
- **Gemini 3 Pro:** No effect ($\Delta = -0.004$, $q = 0.95$).

This is the bidirectional finding. It is a statement about unprompted choice, not about every way of eliciting reasoning.

Table 2: Family A AP by elicitation. Direct pair Δ values are BH-FDR cluster-bootstrap q; CoT pair tests are all $q > 0.68$. $N = 250$ except DeepSeek-R1 (direct $N = 248$, CoT $N = 244$).

Pair	Base D	Base CoT	Reas D	Reas CoT	Δ D	Δ CoT
Claude 4.6	$+0.396$	$+0.100$	$+0.080$	$+0.096$	$+0.316$	$+0.004$
DeepSeek	$+0.224$	$+0.224$	$+0.335$	$+0.234$	-0.111	-0.010
Gemini 2.5	$+0.172$	$+0.104$	$+0.116$	$+0.128$	$+0.056$	-0.024
Gemini 3	$+0.080$	$+0.080$	$+0.084$	$+0.080$	-0.004	0.000

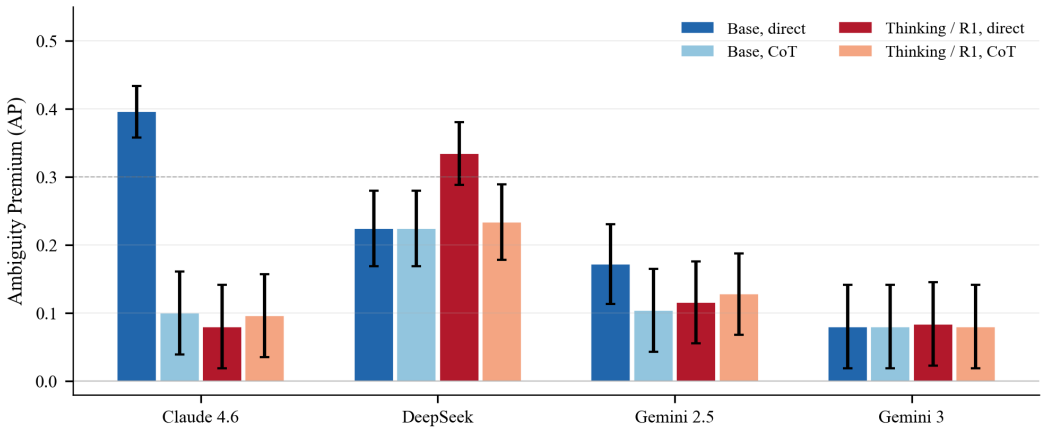


Figure 2: Family A AP by model and elicitation. Claude's thinking-mode collapse is already produced by CoT on the base model. DeepSeek's reverse is direct-only: CoT on R1 returns AP to V3. Gemini pairs barely move. Dashed line: lower end of typical human AP. Error bars: 95% CI.

5.3 Chain-of-Thought Removes the Pair Differences

The same 250 Family A instances under CoT tell a different story (Table 2, Figure 2; Figure 1, right). Prompting Claude *base* to think step by step reduces AP from $+0.396$ to $+0.100$—statistically indistinguishable from Claude thinking-mode AP $+0.080$. DeepSeek-V3 is invariant to CoT (AP $= +0.224$ in both conditions); DeepSeek-R1's extra conservatism is direct-only and falls from $+0.335$ to $+0.234$ under CoT, matching V3. After FDR, none of the four base-vs-reasoning comparisons under CoT is significant (all $q > 0.68$).

Two implications follow. First, Claude thinking-mode is not a unique API effect on this task: eliciting reasoning with a prompt is enough. Second, the bidirectional pattern should not be summarized as "reasoning modes always shift AP." It is a pattern of *unprompted* policies, which converge once reasoning is elicited.

5.4 Temperature Ablation

Claude thinking requires $T = 1.0$ versus $T = 0.0$ for base. We ran Claude base at $T = 1.0$ on the same 250 Family A items: AP $= +0.424$ vs. $+0.396$ at $T = 0.0$ ($\Delta_{\text{temp}} = -0.028$, $p = 0.274$). Temperature alone is not significant. At matched $T = 1.0$, base AP $= +0.424$ vs. thinking AP $= +0.080$, $\Delta = +0.344$ ($p < 0.0001$). Supporting evidence that is not Claude-specific: DeepSeek's reverse occurs with both models at $T = 0.0$; both Gemini pairs are null at matched temperature. (A Gemini 2.5 ablation at $T = 0.7$ does attenuate that pair; Appendix B.)

5.5 Ambiguity-Specificity

On matched Allais tasks under direct and CoT, none of the four pairs shows a significant base-vs-reasoning difference after FDR (all $q > 0.29$; Figure 3). We exclude the ambiguity-aware condition from this dissociation because it is a general framing intervention (it also moves Allais for some

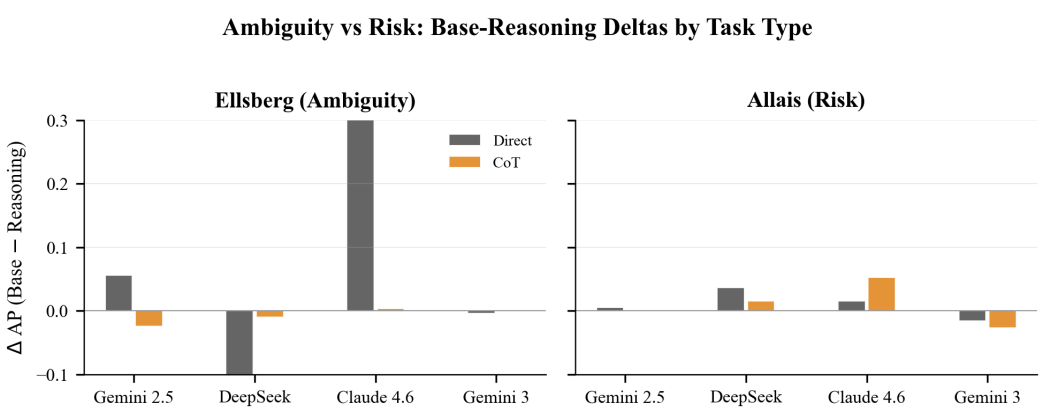


Figure 3: Base-reasoning AP deltas on Ellsberg (left) vs. Allais (right). Gray: direct; orange: CoT. Large deltas appear only on ambiguity tasks under direct prompting; Allais deltas stay near zero.

models; Appendix A). The shifts that survive this control are specific to unknown probabilities, not a generic change in caution.

5.6 Task Complexity, Open-Weight Check, and Tool Selection

Family A is confirmatory for the preference claim. Family C is confirmatory for the control. Family B is a caveat: on the harder three-color probes, Gemini 2.5 base is near chance (AP $= +0.017$ direct) while thinking reaches $+0.330$, consistent with a comprehension gap rather than an ambiguity-preference reversal. We do not fold Family B into the headline bidirectional claim.

Llama 3.3 70B. 880 trials, 0 parse failures. Family A: AP $= +0.500$ direct (ceiling), $+0.428$ under CoT. Family C: AP $= -0.132$ direct, -0.168 CoT (risk-seeking). This shows that (i) extreme ambiguity aversion on two-urn items is not unique to closed Claude-class models, and (ii) the Allais dissociation can appear in an open-weight model. It is not a fifth reasoning pair and it does not reproduce bidirectionality.

Family D (exploratory). DeepSeek-R1 prefers known-reliability tools more than V3 under direct prompting ($\Delta = -0.187$, uncorrected $p = 0.005$), but this cell has 18/100 parse failures. Other pairs are not significant. We do not treat Family D as confirmatory deployment evidence.

6 Discussion

The results cut against two slogans: that reasoning uniformly improves decisions under uncertainty, and that it uniformly degrades them. Under direct prompting, Claude becomes much less ambiguity-averse when thinking is on; DeepSeek-R1 becomes more averse; Gemini 3 does not move. Under CoT, those pair differences collapse. The scientifically conservative summary is: *how* a model is asked to choose—unprompted versus reasoning-elicited—can change its policy toward unknown probabilities, and the direction is a property of the model, not of "reasoning" as a category.

Practical significance. For Claude, direct base choice is 90% known-option; thinking (or CoT on base) is about 58–60%. That is a qualitative shift from a conservative policy to near-indifference on this probe. We do not claim the same percentages will appear in production tool-calling; Family D is too noisy to carry that load. The operational implication is narrower and still useful: if an agent will face missing probabilities, measure that model's AP under the actual decoding and prompting configuration, rather than assuming thinking mode is always safer or always riskier.

Human comparison. Claude base (AP $= +0.396$) sits in the human two-urn range. Gemini 3 Pro ($+0.080$) is far less averse than typical humans. Llama's ceiling ($+0.500$) is the upper bound of that

range. These are snapshot comparisons to published human aggregates (Camerer & Weber, 1992), not a matched human experiment.

7 LIMITATIONS

All claims are conditional on the tested models, items, and decoding settings.

Elicitation dependence. The bidirectional pair effects are concentrated in the direct condition. A reader who only compares thinking APIs under CoT will see little. We now treat that as a result, not a nuisance.

Construct validity. Probes describe known vs. unknown probabilities in language. Models may respond to surface features (explicit numbers vs. their absence) rather than to ambiguity as decision theorists define it. Five rephrasings, label reversal, and the Allais dissociation reduce but do not eliminate that concern.

Coverage. No OpenAI models. Gemini 3.1 is not an independent profile. Llama has no thinking-mode counterpart. Preview APIs may change.

Mechanism. We do not explain why Claude and DeepSeek move in opposite directions when unprompted. Trace analysis on open-weight reasoners is future work, in the same sense that Ellsberg (1961) established the behavioral phenomenon without a cognitive mechanism.

Parse failures. 174/22,680 (0.77%), concentrated in DeepSeek-R1 (139/174), especially Families C and D. Worst-case AP shift from all failures pooled is 0.008; cell-level failure rates on Family D are large enough that we already discarded that family as confirmatory.

8 CONCLUSION

AmbiguityBench shows that ambiguity-sensitive behavior in frontier LLMs is structured, model-dependent, and invisible to risk-only evaluations. Under direct prompting, reasoning configurations shift that behavior in opposite directions across providers; chain-of-thought prompting largely removes the pair-level differences; a temperature ablation indicates the Claude shift is not a $T = 1.0$ artifact. Ambiguity-specific evaluation should be part of deployment assessment for LLM agents, and it should be reported under the elicitation conditions that will actually be used.

REPRODUCIBILITY STATEMENT

Probe generators (`src/probes/*.py`), analysis (`analysis/robust_stats.py`), and raw responses (23,560 trials) are released. Decoding settings, model IDs, and seeds are in `config/`. Replication cost is approximately $57 in API credits. Anonymized code: `https://anonymous.4open.science/r/ambiguity-bench-5BF6`. Cluster-bootstrap and FDR procedures are described in Section 4 and Appendix B.

ETHICS STATEMENT

This work evaluates existing commercial and open-weight APIs. It does not involve human subjects, new training, or release of personal data. Released artifacts are prompts and model outputs on stylized decision problems. Findings could be misread as a recommendation to disable thinking mode; we make no such recommendation. Sponsor identity is withheld for double-blind review.

AI USE STATEMENT

In this work, we used generative AI tools to draft and edit manuscript text, search and summarize related work, implement experiment and analysis code, and produce figures. We used AI tools to

help organize results and refine claim language. All numerical claims were recomputed from the frozen JSONL by the authors. We have not used generative AI tools to generate the experimental responses under study (those come from the evaluated model APIs), to create synthetic datasets, to formulate proofs, or to invent untested hypotheses that appear as results. We take responsibility for the final content of this work.

REFERENCES

Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), 2023.

Colin Camerer and Martin Weber. Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, 5(4):325–370, 1992.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Daniel Ellsberg. Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75 (4):643–669, 1961.

Orhan Erdem and Ragavi Pobbathi Ashok. Artificial finance: How AI thinks about money. *arXiv preprint arXiv:2507.10933*, 2025.

Luise Ge, Yongyan Zhang, and Yevgeniy Vorobeychik. Mind the (DH) gap! a contrast in risky choices between reasoning and conversational LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.

Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 2023.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E. McNamara, and Deming Chen. Decision-making behavior evaluation framework for LLMs under uncertain context. In *Advances in Neural Information Processing Systems*, 2024.

Frank H. Knight. *Risk, Uncertainty and Profit*. Houghton Mifflin, Boston, 1921.

Zhengzhao Ma, Xueru Wen, Boxi Cao, Yaojie Lu, Hongyu Lin, Jinglin Yang, Min He, Xianpei Han, and Le Sun. Decoupling reasoning and confidence: Resurrecting calibration in reinforcement learning from verifiable rewards. *arXiv preprint arXiv:2603.09117*, 2026.

Simon Malberg, Roman Poletukhin, Carolin M. Schuster, and Georg Groh. A comprehensive evaluation of cognitive biases in LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pp. 578–613, Albuquerque, USA, 2025. Association for Computational Linguistics.

Ardavan S. Nobandegani, Irina Rish, and Thomas R. Shultz. Decision-making paradoxes in humans vs machines: The case of the Allais and Ellsberg paradoxes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2024.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 2023.

Thomas Schmied, Jörg Bornschein, Jordi Grau-Moya, Markus Wulfmeier, and Razvan Pascanu. LLMs are greedy agents: Effects of RL fine-tuning on decision-making abilities. In *International Conference on Learning Representations*, 2026.

Rui Wang, Qihan Lin, Jiayu Liu, Qing Zong, Tianshi Zheng, Dadi Guo, Haochen Shi, Peixuan Han, Weiqi Wang, and Yangqiu Song. Rethinking prospect theory for LLMs: Revealing the instability of decision-making under epistemic uncertainty. *arXiv preprint arXiv:2508.08992*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

A Full AmbiguityBench Leaderboard

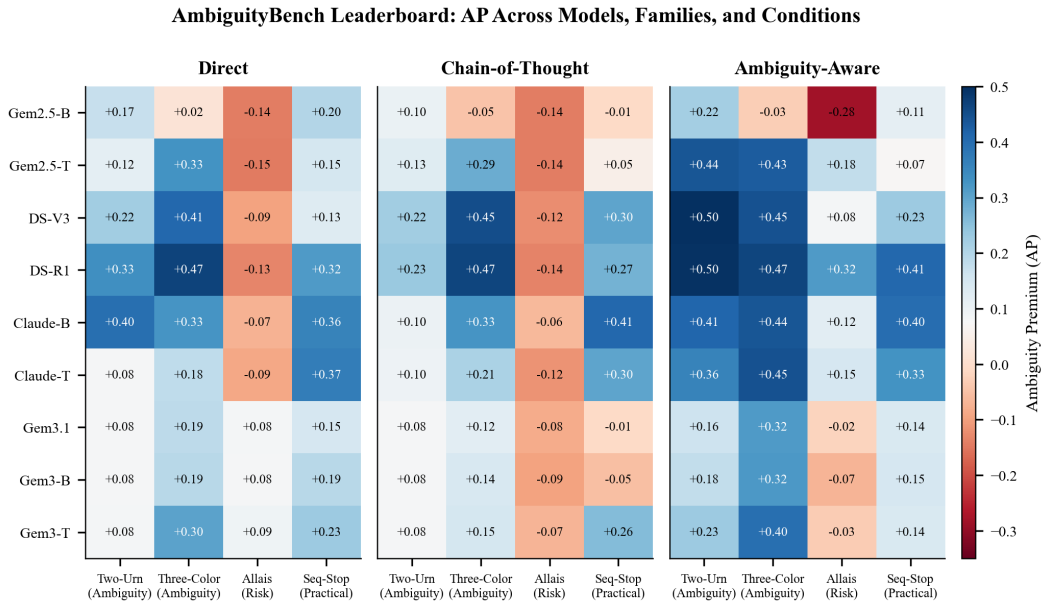


Figure 4: AP across models, families, and prompting conditions. Blue: ambiguity aversion / certainty-favoring; red: seeking. Allais is the matched risk control. Ambiguity-aware prompting (rightmost hue within each family) often raises AP but also moves Allais for some models, so we treat it as a framing intervention rather than an ambiguity-isolated mitigation.

B Parse Failures and Robustness

Table 3: Parse failures by family and model. Total: 174/22,680 (0.77%). DeepSeek-R1 accounts for 139/174. Family D DeepSeek-R1 direct: 18/100 failures on the cell cited as exploratory in the main text.

Model	A: Two-Urn	B: 3-Color	C: Allais	D: Tool
Gem2.5-B	0	11	3	2
Gem2.5-T	0	0	0	0
DS-V3	0	0	0	0
DS-R1	8	1	60	70
Claude-B	0	0	0	0
Claude-T	0	0	0	0
Gem3.1	0	1	0	10
Gem3-B	0	1	0	7
Gem3-T	0	0	0	0
Total	8	14	63	89
Rate	0.1%	0.2%	1.2%	3.3%

Parser methodology. Multi-strategy extraction: short-answer detection, declaration pattern matching, last-occurrence fallback. Pooled worst-case AP shift from all 174 failures: 0.008.

Clustered inference. Pair tests: cluster bootstrap (5,000 iterations, by base instance) with BH-FDR, accounting for non-independence across rephrasings. Per-cell CIs in Table 1 use binomial-normal approximation.

Temperature sensitivity. Gemini 2.5 at $T = 0.7$ (Family A, direct, 3 samples per instance): base AP $= +0.049$, thinking $= +0.084$, $\Delta = -0.035$, $p = 0.18$. Claude base at $T = 1.0$ (250 trials): AP $= +0.424$ vs. $+0.396$ at $T = 0.0$ ($\Delta = -0.028$, $p = 0.274$). Reasoning-mode gap at matched $T = 1.0$: $+0.344$ ($p < 0.0001$).