Self-Consistency Detects Confabulation but Does Not Prevent It in Medical QA

Anonymous ACL submission

Abstract

Large language models (LLMs) *confabulate* in clinical question answering, producing fluent but arbitrary wrong answers that are dangerous when the evidence is sparse. We ask whether dual-process prompting can *reduce* confabulation and whether a model's own answer-disagreement can *detect* it. We evaluate twelve prompting techniques, organised in a System-1 / System-2 dual-process taxonomy, on four medical question-answering benchmarks plus MedHallu, a hallucination-detection task, across six open-weight LLMs under a preregistered analysis. Deliberate prompting gives no reliable accuracy gain: single-pass chain-of-thought helps a little, multi-pass verification and integration do not, and some techniques collapse a model outright; its one clear win is hallucination *detection* (MedHallu). The more useful product of self-consistency is its *disagreement*, not its vote: resampling a question ten times detects confabulation above chance on LongHealth and BioASQ (AUROC 0.72) and moderately on PubMedQA (0.62), and abstaining on the most-uncertain half raises accuracy by up to 13 points. The signal is at chance on emergency-department disposition (0.51), where the model is no more uncertain when wrong than when right. Detection costs one ten-sample resampling pass; we report it as answer-agreement entropy on closed-label tasks and an approximation to semantic entropy on the generative one.

1 Introduction

Large language models increasingly answer medical questions in deployed assistants (Singhal et al., 2023; Maharjan et al., 2024), yet they still produce confident, incorrect outputs. Following Farquhar et al. (2024), we distinguish *hallucination*, an umbrella term for any unfaithful or incorrect generation, from *confabulation*, the subset of hallucinations that are wrong *and arbitrary*, meaning the answer is sensitive to irrelevant details such as the random seed. Confabulations are especially dangerous in medicine, where a model that does not know an answer should abstain rather than fluently invent one.

Two interventions are natural. The first is *prompting*: practitioners change the input format rather than fine-tune, and a large literature proposes deliberate, multi-step prompting to improve reasoning (Schulhoff et al., 2024; Kojima et al., 2022; Wang et al., 2023; Dhuliawala et al., 2024). The second is *uncertainty estimation*: because confabulations vary across resamples, the disagreement among multiple samples can flag them. Semantic entropy formalises this by measuring uncertainty over meanings rather than surface strings (Kuhn et al., 2023; Farquhar et al., 2024). We study both on medical QA, organising prompting techniques with the System-1 / System-2 dual-process distinction (Kahneman, 2011), recently operationalised for LLMs (Zhang et al., 2025; Ziabari et al., 2025).

Our contributions are: (i) a controlled, preregistered comparison showing that deliberate System-2 prompting gives *no reliable accuracy gain* on medical QA, with single-pass chain-of-thought helping a little on average, multi-pass verification and integration not at all, and some techniques collapsing individual models (brittleness), the one clear win being hallucination detection (MedHallu); (ii) the observation that self-consistency's resampling is more valuable as a *confabulation detector* than as a voting mechanism: a ten-sample disagreement signal flags wrong answers well above chance on three of four benchmarks and supports selective prediction that raises accuracy by up to 13 points; and (iii) a robustness check showing the prompting result holds with the one collapsing model removed (and, suggestively, on two reasoning-tuned models). The practical takeaway is to spend a self-consistency budget on detecting confabulation and abstaining, rather than on heavier prompting that does not reliably help.

2 Related work

Does deliberation help? A 2025 literature questions whether more deliberation improves accuracy. Sprague et al. (2025) find chain-of-thought helps mainly on math and symbolic tasks; Liu et al. (2025), drawing on dual-process psychology, show it can *reduce* accuracy where human deliberation also hurts; Ghosal et al. (2025) and Gema et al. (2025) report that longer test-time reasoning can plateau or invert. Evidence in medicine is mixed, with some work reporting that chain-of-thought lowers hallucination (Kim et al., 2025). We provide a controlled test on medical QA.

Confabulation and uncertainty. Black-box hallucination detectors include lexical self-checks (Manakul et al., 2023) and entropy measures. Kuhn et al. (2023) cluster samples by meaning, and Farquhar et al. (2024) use the resulting semantic entropy to detect confabulations and drive abstention. Cheng et al. (2025) report that chain-of-thought lowers hallucination rate while degrading the detector signal. Their finding is consistent with ours: they measure the rate of hallucinated *generations*, whereas we measure task accuracy on largely closed-label medical QA, where reducing a generation's hallucination rate need not produce a more correct label. We characterise where this self-consistency signal works across medical task types, and separate it from MedHallu (Pandit et al., 2025), a hallucination *detection* task.

3 Method and setup

Dual-process taxonomy. From the prompting catalogue of Schulhoff et al. (2024) we select twelve techniques and organise them by the System-1 / System-2 dual-process distinction (Kahneman, 2011; Zhang et al., 2025) into four buckets: **A** direct (zero/few-shot), **B** single-pass System-2 (chain-of-thought variants, Step-Back, Thread-of-Thought), **C** multi-pass verification (self-consistency, chain-of-verification, self-refine, self-verification), and **D** integration. Per-technique citations and definitions are in Appendix A (Table 3).

Benchmarks and models. We use four medical QA benchmarks (PubMedQA (Jin et al., 2019), MIMIC-IV-ED (Johnson et al., 2023), LongHealth (Adams et al., 2024), and BioASQ Task 13B (Nentidis et al., 2025)), plus MedHallu (Pandit et al., 2025), a hallucination-detection task. We evaluate six open-weight LLMs. Four are *primary* models for the main comparison: Qwen3-32B, the Qwen3-30B-A3B mixture-of-experts, Mistral-Small-3-24B, and the medically fine-tuned Llama-3.1-8B-UltraMedical (Zhang et al., 2024). Two are reasoning-tuned models held out for a substitution check: DeepSeek-R1-Distill-Llama-70B and QwQ-32B. Two larger pre-registered models exceed our compute budget at full precision and are excluded (§5). Each model is pinned by HuggingFace commit SHA.

Confabulation signal. For a question x we resample several answers and group them into meaning classes $\mathcal{C}(x)$. The discrete entropy

$$\mathrm{H}(x) = -\sum_{c \in \mathcal{C}(x)} p(c \mid x) \log p(c \mid x) \qquad (1)$$

is high when the resampled answers disagree and low when the model commits to one meaning. We score H against grader correctness with AUROC (0.5 is chance). On the four **closed-label** benchmarks the answer space is a small fixed set, so $\mathcal{C}(x)$ is the set of distinct labels and H is *answer-agreement entropy*. On the one **generative** benchmark (BioASQ factoid and list) one answer has many surface forms; proper semantic entropy clusters them with an entailment model (Kuhn et al., 2023; Farquhar et al., 2024), whereas our rule-based extractor (extracted-string match) only *approximates* that clustering. We obtain the samples with a dedicated pass that resamples each question's zero-shot answer ten times at $T{=}1.0$ (ten independent draws, the same resampling self-consistency performs to vote). Detection therefore costs one ten-sample resampling pass: on the order of a self-consistency call, far less than the multi-pass verification and integration techniques, and more than a single zero-shot answer.

Grading and protocol. Answers are graded by reproducible rule-based extractors (ungradable rate below 2.4% on the four short-answer benchmarks). BioASQ is graded against the official synonym sets, and its summary questions, which have no short gold, are deferred. The hypotheses, metrics, and analysis plan (paired bootstrap with 10,000 resamples over per-question scores, Wilcoxon signed-rank, and Holm-Bonferroni correction at $\alpha{=}0.05$) were registered before any aggregate statistic was computed (Appendix E). We deviate from the registration in three documented ways: the bucket

Benchmark	Type	AUROC [95% CI]
LongHealth	closed-label	**0.72** [.69,.75]
BioASQ	generative	**0.72** [.69,.76]
PubMedQA	closed-label	0.62 [.59,.64]
MIMIC-IV-ED	closed-label	0.51 [.49,.54]
MedHallu	detection	n/a

Table 1: Confabulation-detection AUROC from ten-sample answer-disagreement (Eq. 1), pooled over the four primary models, with 95% bootstrap intervals (n=1040–1990 questions). Above chance on LongHealth, BioASQ, and PubMedQA (intervals exclude 0.5); at chance on MIMIC-IV-ED, where the model is no more uncertain when wrong than when right. MedHallu is a detection task (the model judges another answer), so Eq. 1 does not apply; its result is classification accuracy (Table 2).

Benchmark	A direct	B 1-pass	C verify	D integ.
PubMedQA	**.645**	.635	.553	.595
MedHallu	.737	**.787**	.767	.755
MIMIC-IV-ED	.476	**.507**	**.507**	.504
LongHealth	.731	**.745**	.707	.612
BioASQ	**.769**	.766	.721	.686

Table 2: Accuracy by dual-process bucket, mean over the four primary models, the techniques in each bucket, and ten seeds. Best bucket per benchmark in bold. More verification does not buy accuracy: direct (A) or single-pass (B) prompting matches or beats the heavier verification (C) and integration (D) buckets on every benchmark, so the registered hypothesis that multi-pass beats direct (C>A) does not hold.

comparison is reported on accuracy rather than detection AUROC, two larger models are excluded for compute, and the uncertainty analysis uses a dedicated zero-shot resampling pass rather than the registered per-technique semantic entropy.

4 Results

Figure 1 summarises the accuracy landscape as change from each model's zero-shot baseline: across benchmarks the techniques mostly sit at or below the baseline, with only small gains above and large drops below, previewing the prompting result (§4.2). Per-model zero-shot baselines match previously reported performance (Appendix B), so these are technique effects, not a weak setup. We first present detection.

4.1 Self-consistency disagreement detects confabulation

A model's answer-disagreement over ten resampled answers separates wrong from right above chance on LongHealth and BioASQ (AUROC 0.72 each) and moderately on PubMedQA (0.62); the bootstrap intervals are tight and exclude 0.5 (Table 1). On MIMIC-IV-ED disposition it is at chance (0.51): there the model is no more uncertain when it is wrong than when it is right (mean entropy 0.47 vs. 0.45), so disagreement carries no signal. The detector works precisely where wrong answers are more variable than right ones (right/wrong entropy gap 0.37 and 0.75 on LongHealth and BioASQ, versus 0.02 on MIMIC-IV-ED), making the confabulation-versus-consistent-error distinction concrete. We suspect disposition prediction is multi-class over genuinely ambiguous cases: the model commits to one defensible-but-wrong label (an aleatoric error), not the seed-sensitive guessing of a confabulation. The signal applies whether the answer space is a fixed label set (answer-agreement entropy) or free text (an approximation to semantic entropy on BioASQ).

Its practical payoff is selective prediction (Figure 2): answer the confident questions and abstain on the rest. Answering only the most-confident half raises accuracy by 8/10/13 points on PubMedQA/LongHealth/BioASQ and leaves MIMIC-IV-ED flat (+0.1). This resampling is the same one self-consistency performs to vote: §4.2 shows the vote does not improve accuracy, yet the disagreement among the samples is a usable confabulation detector.

MedHallu measures something different: whether a model can flag a hallucinated candidate answer it is shown. There the fabrication is in the given candidate, not the model's own output, so Eq. 1 does not apply; MedHallu's hallucination result is its classification accuracy (Table 2), where single-pass deliberation is the one clear win for prompting (bucket B 0.79 vs. direct 0.74). Deliberation helps where the task is to *spot* a hallucination, not to avoid one.

4.2 Deliberate prompting gives no reliable gain

More System-2 deliberation does not buy accuracy (Table 2): direct (A) and single-pass (B) prompting match or beat the heavier multi-pass verification (C) and integration (D) on every benchmark, so

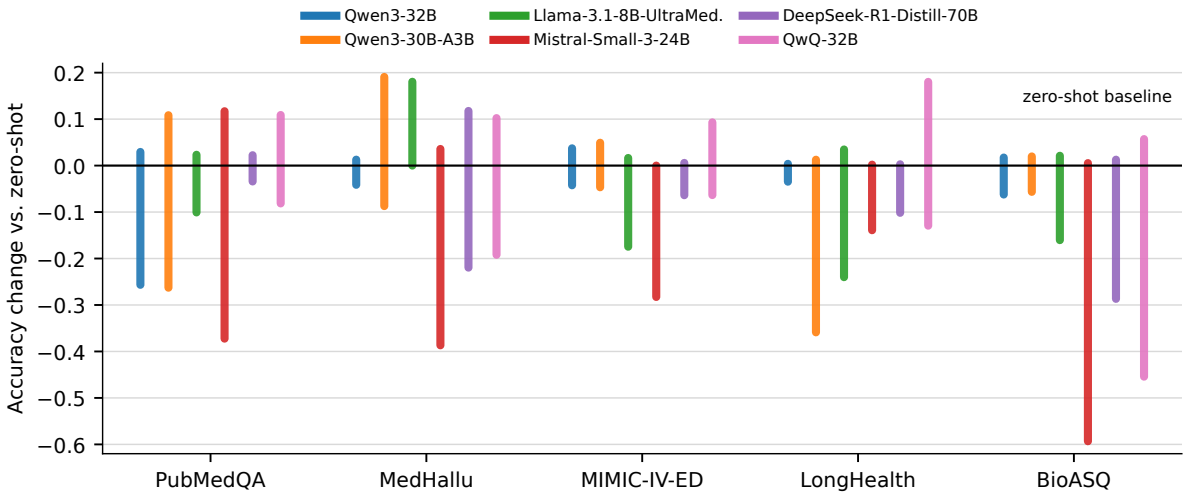

Figure 1: For each model and benchmark, the bar spans the worst and best of the twelve techniques as change from that model's zero-shot baseline (the horizontal line at 0). Bars sit mostly *below* the baseline: the best technique gives at most a small gain, while the worst can collapse a model (Mistral-Small falls 0.59 below baseline on BioASQ). MedHallu is the main exception, with real gains above baseline. Per-model baselines are in Appendix B.

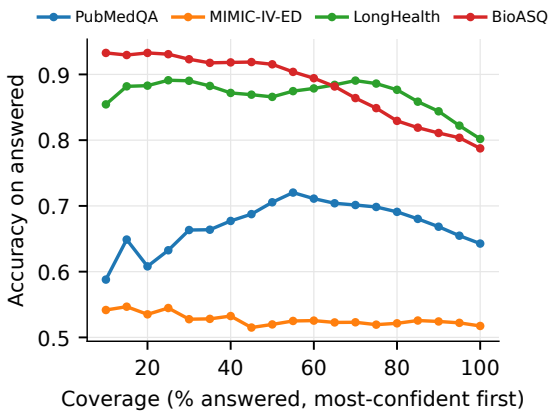

Figure 2: Selective prediction from the disagreement signal: accuracy on the answered subset as the model abstains on its most-uncertain questions (left = answer only the most confident). Accuracy rises sharply on LongHealth, BioASQ, and PubMedQA but is flat on MIMIC-IV-ED, tracking exactly where the signal is informative.

the registered hypothesis that multi-pass beats direct (C>A) fails. Under the pre-registered tests, single-pass bucket B is the only one net-positive over the direct baseline (bucket A, +1.7 points); C (−1.9) and D (−2.0) are net-negative. This net-negative average is driven largely by one model: Mistral-Small collapses under specific techniques (few-shot chain-of-thought reaches 0.21 on PubMedQA, answering "maybe" to most items; self-consistency reaches 0.25 on BioASQ). Excluding it (Appendix C), D−A moves from −2.0 to +0.3 and the technique-versus-zero-shot contrasts go from 70-worse / 49-better to a balanced 40 / 44. These are two distinct findings, not a contradiction: *on average* deliberate prompting gives no reliable gain (the best single technique, Thread-of-Thought, adds only 3.0 points and is significant on 7 of 16 cells), while *for specific model-technique pairs* it is brittle enough to collapse a model. Both are risks a practitioner should weigh.

Reasoning-tuned models. The same pattern holds on the two reasoning-tuned models: multi-pass C and integration D stay within a point of their System-1 baseline (C−A +0.7, D−A +0.2) while single-pass B still helps (+3.3); with n=2 we read this as consistent rather than conclusive (Sprague et al., 2025; Ghosal et al., 2025; Gema et al., 2025).

5 Conclusion

On medical QA, deliberate prompting is not a reliable lever: multi-pass verification and integration give no accuracy gain over direct prompting, single-pass chain-of-thought helps only slightly, and effects are model-dependent. The resampling that self-consistency performs is more useful for *detecting* confabulation than for voting: its disagreement flags errors above chance on three of four benchmarks and supports abstention that lifts accuracy by up to 13 points. Practically, spend a resampling budget on detection, not heavier prompting.

Limitations

Detection is measured on a dedicated ten-sample resampling pass over the zero-shot answer; we did not extend it to the deliberate-prompting techniques, or to MedHallu, where the model judges another answer so own-answer disagreement does not apply. Our clustering uses extracted-answer matching rather than a trained entailment model, so on the one generative benchmark it only approximates meaning-level clustering. We report six open-weight LLMs; two larger pre-registered models exceed our compute budget, so the model-scale question is untested, and closed-source frontier models are out of scope. For LongHealth, the few-shot and self-consistency accuracy cells use YaRN-rescaled context for the Qwen models (uneven coverage in Table 2); the detection result uses the zero-shot answer and is unaffected. We benchmark the disagreement signal only against chance, not against other black-box uncertainty estimators (token entropy, P(true), or lexical self-checks Manakul et al., 2023); we do not claim it is the best such signal. All claims concern accuracy and the detectability of confabulation, not parametric faithfulness, a separate axis.

Ethical considerations

MIMIC-IV-ED is governed by the PhysioNet Data Use Agreement. We release neither raw MIMIC text nor any derived per-patient row, only the regeneration script and the exemplar identifiers needed to reproduce the split under independently obtained credentials. All models are open-weight and no patient data was sent to hosted APIs. The paper makes no claim of clinical deployment readiness. The uncertainty AUROCs are correlates of confabulation detectability, not safety certifications, and should not be read as deployment-ready abstention thresholds.

References

Lisa Christine Adams, Keno K. Bressem, and 1 others. 2024. LongHealth: A question answering benchmark with long clinical documents. *arXiv preprint arXiv:2401.14490*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jiahao Cheng, Tiancheng Su, Jia Yuan, Guoxiu He, Jiawei Liu, Xinqi Tao, Jingwen Xie, and Huaxia Li. 2025. Chain-of-thought prompting obscures hallucination cues in large language models: An empirical evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1272–1305, Suzhou, China. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630.

Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Arditi, Jacob Goldman-Wetzler, Kit Fraser-Taliente, Henry Sleight, Linda Petrini, Julian Michael, Beatrice Alex, Pasquale Minervini, Yanda Chen, Joe Benton, and Ethan Perez. 2025. Inverse scaling in test-time compute. *Transactions on Machine Learning Research*. ArXiv:2507.14417.

Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, Yifu Lu, Mengdi Wang, Dinesh Manocha, Furong Huang, Mohammad Ghavamzadeh, and Amrit Singh Bedi. 2025. Does thinking more always help? mirage of test-time scaling in reasoning models. *arXiv preprint arXiv:2506.04210*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Liwei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, and 1 others. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. 2025. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *International Conference on Machine Learning (ICML)*. ArXiv:2410.21333.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jenish Maharjan, Anurag Garikipati, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Rahul Thapa, Qingqing Mao, and Ritankar Das. 2024. OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *npj Digital Medicine*, 7(1):196.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Anastasios Nentidis, Anastasia Krithara, Konstantinos Bougiatiotis, and Georgios Paliouras. 2025. Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF)*.

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. 2025. MedHallu: A comprehensive benchmark for detecting medical hallucinations in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics. ArXiv:2502.14302.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and 12 others. 2024. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations (ICLR)*. ArXiv:2409.12183.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Kaiyan Zhang, Sihang Ding, Yuxin Wang, Zhennan Sun, Junzhe Chen, Xuekai Jin, Hao Zhou, Bowen Zhou, and Yu Wu. 2024. UltraMedical: Building specialized generalists in biomedicine. *arXiv preprint arXiv:2406.03949*.

Shao Zhang, Xihuai Wang, Wenhao Zhang, Chaoran Li, Junru Song, Tingyu Li, Lin Qiu, Xuezhi Cao, Xunliang Cai, Wen Yao, Weinan Zhang, Xinbing Wang, and Ying Wen. 2025. Leveraging dual process theory in language agent framework for real-time simultaneous human-ai collaboration. *arXiv preprint arXiv:2502.11882*.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Take a step back: Evoking reasoning via abstraction in large language models. In *International Conference on Learning Representations (ICLR)*.

Yucheng Zhou, Xiubo Geng, Tao Shen, Chongyang Tao, Guodong Long, Jian-Guang Lou, and Jianbing Shen. 2024. Thread of thought unraveling chaotic contexts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*.

Alireza S. Ziabari, Nona Ghazizadeh, Zhivar Sourati, Farzan Karimi-Malekabadi, Payam Piray, and Morteza Dehghani. 2025. Reasoning on a spectrum: Aligning LLMs to system 1 and system 2 thinking. *arXiv preprint arXiv:2502.12470*.

A Prompting techniques

Table 3 lists the twelve techniques and their dual-process buckets. Verbatim prompt templates are released with the code.

ID	Bucket	Technique
A1	Direct	Zero-Shot
A2	Direct	Few-Shot (Brown et al., 2020)
B1	1-pass S2	Zero-Shot CoT (Kojima et al., 2022)
B2	1-pass S2	Few-Shot CoT (Wei et al., 2022)
B3	1-pass S2	Step-Back (Zheng et al., 2024)
B4	1-pass S2	Thread-of-Thought (Zhou et al., 2024)
C1	Multi-pass	Self-Consistency (Wang et al., 2023)
C2	Multi-pass	Chain-of-Verification (Dhuliawala et al., 2024)
C3	Multi-pass	Self-Refine (Madaan et al., 2023)
C4	Multi-pass	Self-Verification (Weng et al., 2023)
D1	Integration	Fast-Draft then Verify
D2	Integration	Few-Shot CoT, Self-Consistency, CoVe

Table 3: The twelve prompting techniques and their dual-process buckets.

B Per-model and per-technique results

The full per-(model, technique, benchmark) table (mean and standard error over ten seeds) and the per-model detection AUROCs are generated deterministically from the released per-cell outputs and included in the supplementary material.

Table 4 reports each model's zero-shot accuracy on every benchmark, the System-1 baseline.

C Robustness without Mistral-Small

Mistral-Small is the one primary model that collapses under specific techniques (Figure 1). Recomputing the headline contrasts over the other three primary models gives bucket B−A $+2.0$, C−A -1.0, and D−A $+0.3$ points, and of 132 technique-versus-zero-shot contrasts 44 are significantly better than zero-shot and 40 significantly worse (versus 49 and 70 over all four models). The conclusion that deliberate prompting yields no reliable gain holds; the appearance of net harm does not.

D Pre-registered hypotheses

The registration states directional hypotheses on the bucket comparison, the substitution effect for reasoning-tuned models, an integration-versus-cost hypothesis, and a medical-tuning hypothesis, together with the analysis plan summarised in §3. The registered comparisons were phrased on detection AUROC by bucket; because only the resampling techniques yield a usable AUROC, we apply the registered statistical machinery to the accuracy contrasts the bucket comparison supports and report this deviation here.

E Reproducibility

The protocol, including the directional hypotheses and the analysis plan, was pre-registered before any aggregate statistic was computed, at `https://osf.io/xhk34` (locked at git SHA `019103e`). Upon publication we will release the repository, the run manifest with pinned model SHAs, the dependency lock hash, the prompt templates, and the grading and analysis scripts, with the per-cell model outputs as supplementary material. All accuracy and uncertainty numbers are reproduced from the saved per-cell outputs by those scripts without re-running any model.

F AI Assistant Acknowledgment

The authors acknowledge the use of AI assistants (such as large language models) during the preparation of this manuscript for tasks including code generation, data processing pipeline drafting, and text refinement. All AI-generated content was critically reviewed, heavily edited, and validated by the human authors. All authors have consented to this use and take full, final responsibility for the content, methodology, and conclusions presented in this paper.

Model	PubMedQA	MedHallu	MIMIC-IV-ED	LongHealth	BioASQ
Primary					
Qwen3-32B	.676	.874	.546	.850	.785
Qwen3-30B-A3B	.526	.643	.499	.808	.787
Llama-3.1-8B-UltraMedical	.739	.665	.542	.693	.774
Mistral-Small-3-24B	.579	.807	.530	.851	.799
Reasoning-tuned					
DeepSeek-R1-Distill-Llama-70B	.730	.736	.598	.876	.842
QwQ-32B	.637	.768	.404	.678	.794

Table 4: Zero-shot accuracy (mean over ten seeds), the System-1 baseline. For reference, the PubMedQA single-model state of the art at release was 0.681 with 0.780 human performance (Jin et al., 2019); on LongHealth, Mistral-Small ranks among the strongest open models (Adams et al., 2024), which our results reproduce; the best MedHallu detector reaches F1 0.625 on the hard subset (Pandit et al., 2025). Our accuracies are consistent with these, indicating a sound setup rather than weak baselines.