

CAST: Contrastive Adjudication via Selective Triggering for Clinical LLMs

Krishna Senthilkumar Archana Vaidheeswaran Shivank Garg

Algoverse AI Research
krishnasent6@gmail.com
archana@algoverseairesearch.org
shivank@algoverseairesearch.org

Abstract

Large language models (LLMs) often follow conflicting reasoning paths on difficult tasks, creating uncertainty that is especially dangerous in clinical settings where errors can inform incorrect diagnoses or treatments. Standard self-consistency (SC) improves accuracy by voting over multiple reasoning chains, but it does not resolve the contradictions that produced the uncertainty. We propose Contrastive Adjudication via Selective Triggering (CAST), a lightweight inference-time method that detects disagreement across chain-of-thought samples and triggers contrastive adjudication only when model confidence falls below a threshold. Instead of re-asking the same question, CAST extracts the strongest argument for each competing reasoning path, presents them side by side, and asks the model to resolve the conflict definitively. Across multiple model families and three medical benchmarks, including HealthBench, a physician-rubric clinical task benchmark, CAST outperforms standard SC, compute-matched SC, SC-Max (with the maximum possible number of calls, usually 13), and the pooled CASC variant (which votes across all pre- and post- adjudication votes) across all datasets. Gains are largest on uncertain questions, reaching up to 29% over SC. These results, coupled with selective use of model calls, make CAST practical for clinical deployment where models must be accurate, efficient, and uncertainty-aware.

1 Introduction

AI-assisted tools are increasingly being adopted across the medical industry, and have achieved impressive accuracy on medical question answering benchmarks (Singhal et al., 2023; Nori et al., 2023). Studies suggest, however, that a significant fraction of AI-generated clinical responses contain hallucinations (Ji et al., 2023; Huang et al., 2023) or factually incorrect answers, despite appearing confident (Kadavath et al., 2022; Guo et al., 2017).

These errors can have catastrophic consequences in high-stakes settings, especially when they directly inform incorrect diagnoses or treatments. Ensuring that LLM outputs are reliable and uncertainty-aware is therefore a prerequisite for safe medical deployment. Previous work has addressed this concern through self-consistency (SC) (Wang et al., 2023), where multiple reasoning chains are sampled and the majority answer is selected. While this approach outperforms greedy decoding, it does not resolve the conflicts that come with resampling. It only votes over them, thus preserving or amplifying uncertain reasoning. Additionally, previous work has primarily focused on quantifying uncertainty in medical contexts. Proposed paradigms such as two-step verification (Abbasi Yadkori et al., 2024) have been shown to improve uncertainty detection in these contexts. However, reducing errors and ultimately improving accuracy on these uncertain questions remains a major challenge.

To address this gap, we propose Contrastive Adjudication via Selective Triggering (CAST). At a high level, CAST works in two phases. First, it samples K independent reasoning chains and measures how much they agree. If the model is confident, meaning one answer dominates, it returns that answer immediately with no extra computation needed. If the model is uncertain, meaning a substantial minority disagrees, CAST enters a second phase. It extracts the strongest argument for each competing answer from Phase 1, constructs a contrastive prompt that explicitly presents both sides, and instructs the model to resolve the conflict definitively. Critically, CAST returns a majority vote over the Phase-2 adjudication responses only, discarding the conflicting Phase-1 votes that triggered uncertainty detection. We additionally evaluate CASC, a pooled variant that combines Phase-1 and Phase-2 votes subject to a do-no-harm guard, as an ablation to isolate the contribution of adjudication-only voting.

We evaluate CAST and its ablations on MedMCQA, PubMedQA, and HealthBench, an industry-grade clinical task benchmark scored with physician-written rubrics. Across all benchmarks, CAST outperforms greedy decoding, standard SC, compute-matched SC, SC Max, and the pooled CASC variant. CAST also achieves its largest gains on uncertain questions, strengthening the evidence that post-adjudication voting is a substantial driver of performance. These results show that adjudicating competing reasoning paths is a more effective use of inference-time compute than simply sampling more paths, with direct implications for latency-sensitive clinical deployment. Overall, CAST is a lightweight inference-time method that detects trajectory-level uncertainty, triggers contrastive adjudication only on uncertain questions, and resolves them by voting only over Phase-2 adjudication responses.

2 Related Work

2.1 Chain-of-Thought, Self-Consistency, and Uncertainty in LLMs

Wei et al. (2022) demonstrated that prompting LLMs to produce step-by-step reasoning chains significantly improves performance on complex reasoning tasks. Wang et al. (2023) extended this with self-consistency, showing that sampling multiple reasoning chains and taking a majority vote over final answers outperforms greedy decoding across a range of benchmarks. Iterative prompting strategies have also been explored in medical QA settings. Madaan et al. (2023) proposed Self-Refine, where a model iteratively critiques and revises its own outputs, and Pan et al. (2023) proposed a framework using external information to improve model responses.

A closely related line of work studies uncertainty in LLMs. Kadavath et al. (2022) showed that LLMs can be calibrated to estimate their own confidence and that verbalized confidence correlates with accuracy. Xiong et al. (2024) provide a comprehensive survey of uncertainty quantification methods for LLMs, identifying entropy-based measures as among the most reliable signals. In medical contexts, Abbasi Yadkori et al. (2024) propose an iterative prompting method to estimate epistemic uncertainty from responses, showing that prompting can improve uncertainty detection. Unlike these studies, CAST does not only quantify or flag uncertainty. It uses trajectory-level disagreement to

detect uncertain predictions and then resolves them through contrastive adjudication, discarding conflicting Phase-1 votes entirely rather than pooling them.

2.2 Medical Question Answering

Jin et al. (2021) introduced MedQA, establishing a benchmark for clinical reasoning with LLMs. Pal et al. (2022) introduced MedMCQA, a large-scale multiple-choice dataset covering a broad range of medical subjects. Prior work has shown that standard prompting and chain-of-thought methods improve performance on these benchmarks, but uncertainty-aware inference strategies have received limited attention in the medical QA setting. Singhal et al. (2023) introduced Med-PaLM, showing that instruction-tuned LLMs can achieve physician-level performance on medical licensing exam questions. Nori et al. (2023) evaluated GPT-4 on USMLE-style questions, demonstrating that frontier models can pass medical board examinations without any domain-specific fine-tuning. Despite these advances, miscalibration remains a persistent concern because models that achieve high accuracy may still assign high confidence to incorrect answers (Kadavath et al., 2022). CAST directly addresses this by identifying uncertain predictions at inference time and resolving them through contrastive adjudication, without any additional training or fine-tuning.

2.3 Clinical Deployment of LLMs

Recent work has highlighted the gap between benchmark performance and safe clinical deployment of LLMs (Ji et al., 2023; Huang et al., 2023). A core challenge is that high aggregate accuracy does not preclude dangerous miscalibration on individual cases, precisely the failure mode that motivates uncertainty-aware inference. Proposed solutions have included retrieval augmentation, ensemble methods, and post-hoc calibration, but these approaches either require external knowledge sources, additional training, or significantly increased compute. CAST addresses this gap with a training-free, inference-time method that applies additional compute selectively and only where model confidence is lowest, making it directly compatible with existing clinical LLM deployment pipelines.

Contrastive Adjudication Via Selective Triggering

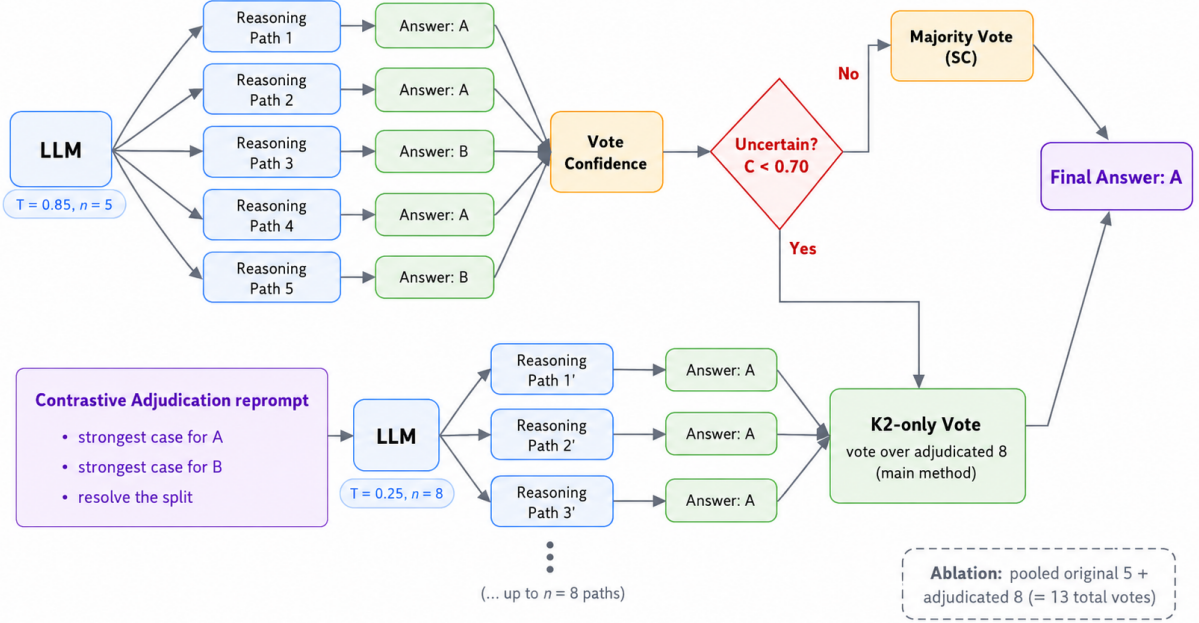


Figure 1: Overview of CAST. Phase 1 samples multiple reasoning paths and computes vote confidence. If the vote is uncertain, CAST extracts the strongest competing arguments, runs contrastive adjudication, and returns a Phase-2-only majority vote.

3 Methodology

3.1 Preliminaries

Given a question q with answer options $\mathcal{O}$, we aim to select the correct answer $a^* \in \mathcal{O}$. We assume access to a language model $\mathcal{M}$ that generates reasoning paths. Standard self-consistency (Wang et al., 2023) samples K reasoning chains $\{c_1, \dots, c_K\}$ at temperature $\tau > 0$, extracts a final answer $\hat{a}_k$ from each chain, and returns the majority vote:

$$\hat{a}_{SC} = \arg \max_a \sum_{k=1}^K \mathbf{1}[\hat{a}_k = a]. \quad (1)$$

3.2 Uncertainty Estimation via Vote Disagreement

CAST measures uncertainty over the answer distribution induced by the K sampled chains. Let $p(a) = \frac{1}{K} \sum_{k=1}^K \mathbf{1}[\hat{a}_k = a]$ be the empirical answer distribution. We compute vote confidence $C = \max_a p(a)$, the fraction of votes for the most common answer, and the minority fraction $m = p(a_{(2)})$, the share of votes received by the leading competing answer.

A question is flagged as uncertain when $C < \theta$ and $m \geq \mu$, where θ is a confidence threshold and μ is a minimum minority share. We chose

$\theta = 0.70$, $\mu = 0.20$, $K = 5$, and $K_2 = 8$ using a 200-question held-out MedMCQA validation subset. Appendix C reports the hyperparameter and threshold sensitivity analysis.

3.3 Contrastive Adjudication

When a question is flagged as uncertain, CAST clusters the Phase 1 reasoning trajectories by their final answer. Within each answer cluster, CAST selects the single most representative trajectory as the argument for that answer. Specifically, it selects the trajectory with the highest mean pairwise unigram Jaccard similarity to all other trajectories in that cluster, computed over the reasoning text and excluding the final answer token. For singleton clusters, the only chain is selected. This passes the strongest, most consensus-aligned reasoning chain to Phase 2, rather than an outlier chain.

The contrastive prompt presents the original question, the strongest argument for the majority answer, and the strongest argument for the minority answer. The model is instructed to identify the evidence that settles the disagreement and resolve it definitively rather than simply re-answering. Phase 2 is sampled $K_2 = 8$ times at temperature $\tau_2 = 0.25$. Full prompt templates and pseudocode are provided in Appendix A.

3.4 CAST and CASC Voting

The key design decision that distinguishes CAST from pooling-based methods is the treatment of Phase-1 votes on uncertain questions. When the model is uncertain, the Phase-1 votes are by definition conflicted. Including them in the final vote can dilute the signal from the contrastive adjudication phase. CAST instead discards Phase-1 votes entirely on triggered questions and returns the majority vote over Phase-2 adjudication responses only:

$$\hat{a}_{\text{CAST}} = \begin{cases} \hat{a}_{\text{SC}}, & \text{if not triggered,} \\ \arg \max_a \sum_{k=K+1}^{K+K_2} \mathbf{1}[\hat{a}_k = a], & \text{if triggered.} \end{cases} \quad (2)$$

We additionally evaluate CASC, a pooled variant that combines Phase-1 and Phase-2 votes subject to a do-no-harm guard. CASC takes the majority vote over all $K + K_2$ answers only when pooled confidence exceeds Phase-1 confidence by at least $\delta = 0.10$. Otherwise, CASC keeps the Phase-1 majority vote. CASC serves as an ablation to isolate the contribution of adjudication-only voting.

3.5 Application to Free-Response Clinical Conversations

For HealthBench, answers are free-response clinical outputs rather than answer letters. We keep the same Phase 1, trigger, and Phase 2 structure, but change how response paths are formed and how the final response is selected. Phase 1 samples $K = 5$ candidate clinical responses to the same conversation. Because free-response paths do not map cleanly to answer clusters, a selector with no access to rubrics, ideal completions, gold information, answer keys, or grader feedback chooses the best Phase-1 response, estimates confidence, and identifies distinct response paths using only the original conversation and model-generated candidates.

A HealthBench example is treated as uncertain when the selector indicates that adjudication is needed or when its confidence falls below $\theta = 0.70$. Phase 2 then receives the original conversation and the two strongest or most distinct Phase-1 response paths, and generates $K_2 = 8$ adjudicated responses at temperature $\tau_2 = 0.25$. CAST selects the final response from the Phase-2 responses only on triggered examples, and otherwise keeps the Phase-1 selected response. CASC selects between the Phase-1 selected response and the Phase-2 candidate pool. Rubrics are used only after all method outputs are fixed, for post-hoc grading. We discuss the free-response selector, why we use it instead of

answer clustering, and the held-out trigger check in Appendix B.

4 Experimental Setup

4.1 Datasets

We evaluate CAST on three biomedical benchmarks spanning multiple-choice question answering and free-response clinical tasks.

MedMCQA - MedMCQA (Pal et al., 2022) is a large-scale medical examination benchmark containing clinically oriented multiple-choice questions across subjects including pathology, pharmacology, anatomy, and physiology.

PubMedQA - PubMedQA (Jin et al., 2019) consists of biomedical research questions paired with PubMed abstracts and yes/no/maybe labels designed to evaluate scientific reasoning and evidence interpretation.

HealthBench - HealthBench (Arora et al., 2025) contains realistic free-response clinical conversations evaluated with physician-written rubrics. This benchmark is especially useful for clinical deployment because the model must produce a patient-facing clinical response rather than choose from a fixed answer set.

To ensure consistent computational cost across methods, we evaluate on fixed, randomly sampled benchmark subsets consisting of 500 examples from MedMCQA, 500 examples from PubMedQA, and 500 examples from HealthBench, using the same subset for all methods.

4.2 Model Configuration

Experiments use three model families: GPT-4o-mini, DeepSeek V3, and Llama-4-Maverick. For standard self-consistency, we sample $K = 5$ reasoning trajectories at temperature $\tau = 0.85$. CAST first performs the same sampling procedure and computes vote confidence over sampled outputs. When the trigger condition is met, CAST samples $K_2 = 8$ contrastive adjudication responses at temperature $\tau_2 = 0.25$ and returns the majority vote over Phase-2 responses only.

For HealthBench, outputs are free-response clinical answers rather than answer letters. We therefore use a selector to choose among generated clinical responses and estimate whether adjudication is needed. The selector has no access to rubrics, ideal completions, gold information, answer keys, or grader feedback. Rubrics are used only after all method outputs are fixed, for post-hoc grading.

We include the selector prompt and the held-out HealthBench trigger comparison in Appendix B.

4.3 Baselines

We compare CAST against greedy decoding, standard self-consistency (SC), compute-matched self-consistency (SC-CM), SC Max, and CASC. Greedy decoding uses a single response. SC uses $K = 5$ sampled responses and selects the majority answer for MedMCQA and PubMedQA, or the selected best response for HealthBench. SC-CM uses the same average compute budget as CAST for each dataset and model. SC Max uses the maximum sampling budget available for that setting. This is usually 13 samples on the multiple-choice benchmarks, matching the full $K + K_2$ budget used on triggered examples, but is adjusted for HealthBench because free-response selection and rubric-call accounting change the total evaluation budget. CASC is the pooled variant that combines Phase-1 and Phase-2 votes subject to the do-no-harm guard. CAST differs from CASC by voting only over Phase-2 adjudication responses on triggered questions.

4.4 Evaluation Metrics

For MedMCQA and PubMedQA, we report exact-match accuracy. For HealthBench, we report physician-rubric score on free-response clinical conversations.

To evaluate uncertainty-aware behavior, we also report performance on the triggered subset: questions where the Phase 1 responses exhibit enough disagreement to activate CAST. We report Δ_{unc} as the improvement of CAST over SC on this triggered subset. This is the only subset where CAST can change the output, since confident questions return the same output as SC by design.

5 Results

Table 1 reports overall performance on MedMCQA, PubMedQA, and HealthBench. For MedMCQA and PubMedQA, we report exact-match accuracy. For HealthBench, we report physician-rubric score on free-response clinical conversations. Across all three datasets and all three model families, CAST is the strongest method. CAST outperforms standard SC, compute-matched SC, SC Max, and CASC in every model-dataset setting.

The gains are not explained by simply sampling more reasoning paths. On MedMCQA and PubMedQA, CAST improves over SC by up to +5.8

percentage points overall, and it also outperforms SC Max, which uses 13 samples. On HealthBench, CAST improves over SC by +1.0 to +2.4 rubric-score points, despite the harder free-response setting. This pattern supports the main claim: adjudicating competing reasoning paths is more effective than only adding more samples.

The largest gains appear exactly where CAST is designed to help. On triggered questions, where Phase 1 responses disagree, CAST improves over SC by +4.7% to +29.0% across the multiple-choice benchmarks. The largest gain appears on PubMedQA with GPT-4o-mini, where CAST improves triggered-question accuracy from 28.0% to 57.0%. On HealthBench, CAST also improves triggered examples, with gains of +1.8 to +4.0 rubric-score points. This confirms that the trigger identifies questions where extra computation is most useful.

CAST also outperforms CASC in every setting. CASC often improves over SC, but pooling Phase-1 and Phase-2 votes still underperforms adjudication-only voting. This ablation demonstrates that once a question triggers, the Phase-1 votes can dilute the adjudication signal rather than strengthen it.

On the multiple-choice benchmarks, CAST remains incredibly compute-efficient. It triggers on only 8.0% to 20.0% of questions and averages roughly 1.2 additional generation calls per question. HealthBench requires additional selector as described in Appendix B and rubric-call accounting because the outputs are free-response clinical answers rather than answer letters. Still, however, the selective trigger makes the method meaningfully more computationally efficient. To further test generalization, we evaluated CAST on Claude Sonnet 5, a substantially stronger frontier reasoning model, across MedMCQA, PubMedQA, and the non-clinical StrategyQA benchmark using the same fixed threshold $\theta = 0.70$ with no model- or dataset-specific retuning, where CAST again outperformed every baseline, improving over SC on the triggered subset (Δ_{unc}) by +8.0% to +25.0%.

6 Discussion

A central finding of this work is that CAST outperforms both compute-matched SC and SC Max. This means the gains are not just a function of using more model calls. Standard SC asks the same question again and counts votes. CAST uses the disagreement from those votes to build a stronger second prompt, where the model must compare

Dataset	Model	SC	SC-CM	SC Max	CASC	CAST	Trigger	Δ_{unc}
MedMCQA	GPT-4o-mini	68.0%	67.4%	67.6%	68.6%	68.8%	17.0%	+4.7%
	DeepSeek V3	74.8%	75.4%	76.2%	76.0%	77.2%	10.4%	+23.1%
	Llama-4-Maverick	80.0%	80.0%	79.8%	80.0%	80.4%	8.0%	+4.8%
PubMedQA	GPT-4o-mini	35.2%	35.0%	35.4%	38.8%	41.0%	20.0%	+29.0%
	DeepSeek V3	46.8%	47.4%	48.2%	48.8%	49.2%	17.4%	+13.8%
	Llama-4-Maverick	58.0%	58.8%	59.2%	58.6%	59.8%	13.6%	+13.2%
HealthBench	GPT-4o-mini	14.0%	13.6%	14.2%	15.3%	15.8%	55.0%	+3.3%
	DeepSeek V3	24.4%	24.7%	24.6%	24.9%	25.4%	55.6%	+1.8%
	Llama-4-Maverick	14.6%	15.8%	15.1%	16.3%	17.0%	60.0%	+4.0%

Table 1: Overall performance across MedMCQA, PubMedQA, and HealthBench. MedMCQA and PubMedQA report exact-match accuracy; HealthBench reports normalized physician-rubric score. Δ_{unc} is CAST’s gain over SC on triggered questions. CAST outperforms CASC, the pooled variant, and all baselines in every setting.

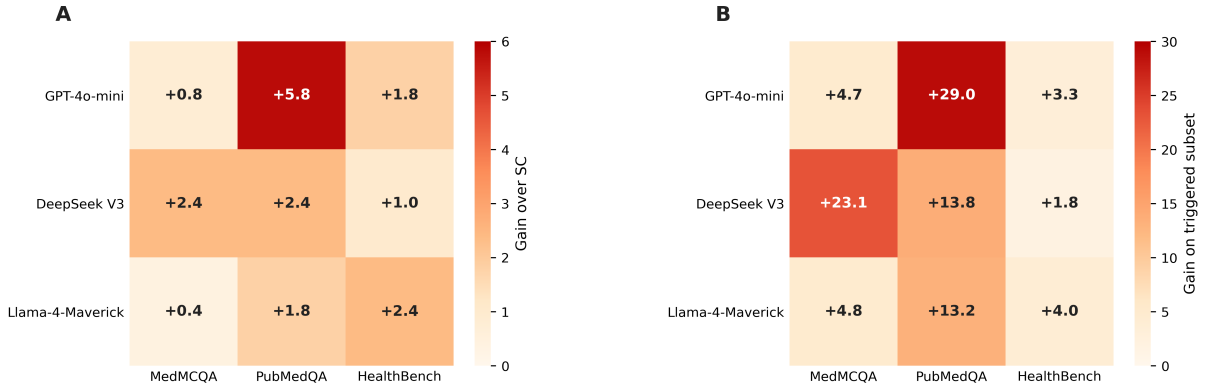


Figure 2: CAST improves over SC across datasets and model families. Panel A shows overall gain; Panel B shows gain on triggered questions, where CAST can change the output.

the competing reasoning paths and resolve the conflict directly. This suggests that CAST’s gains are not explainable by the number of reasoning paths sampled alone (Lightman et al., 2024; Cobbe et al., 2021).

CAST also consistently outperforms CASC, the pooled variant with a do-no-harm guard. This shows that the final voting rule matters. When the model is uncertain, the Phase-1 votes are already conflicted. Pooling those votes with adjudication responses can preserve the same uncertainty that Phase 2 was meant to resolve. CAST avoids this by voting only over the Phase-2 adjudication responses on triggered questions.

The overall gains come from the uncertain subset, where the model initially follows competing paths. CAST leaves confident questions unchanged and spends extra computation where the model is least stable, rather than increasing compute uniformly across all questions, which is promising for future clinical evaluation. Threshold and hyperparameter sensitivity analyses are provided in Appendix C.

7 Clinical Deployment Considerations

A key motivation for this work is the practical deployment of LLMs in clinical settings, where inference-time reliability directly affects patient safety. CAST is useful here for three reasons. First, it applies additional computation only when the model is uncertain, triggering on only a subset of questions, and even in contexts where uncertainty is high, CAST still significantly reduces necessary computation. Second, CAST is a purely inference-time method and requires no modification to model weights, labeled training data, or domain-specific fine-tuning. Third, the trigger itself provides a natural uncertainty flag: questions that activate Phase 2 are questions where the model’s initial reasoning was conflicted, and can be routed to human review instead of returning an automated answer.

8 Conclusion

We presented CAST (Contrastive Adjudication via Selective Triggering), a lightweight inference-time method that detects model uncertainty through disagreement across sampled reasoning paths and

triggers contrastive adjudication only when confidence falls below a threshold. Across MedMCQA, PubMedQA, and HealthBench, CAST consistently outperforms standard self-consistency, compute-matched self-consistency, SC Max, and CASC, with the largest gains on uncertain questions while leaving confident questions unchanged. These results suggest that resolving disagreement directly is more effective than simply sampling more answers, making CAST practical for uncertainty-aware clinical LLM deployment.

9 Limitations

While CAST is more computationally efficient than uniformly running the contrastive stage on every question, owing to its selective triggering mechanism, it still requires multiple API calls per question. In clinical settings, the additional calls required for resampling and contrastive adjudication may not be available. We continue to pursue future work aimed at achieving uncertainty aware clinical deployment in further resource-constrained settings.

Additionally, while HealthBench provides a benchmark aimed to replicate the clinical industry, true deployment in healthcare settings comes with additional safety and privacy concerns. Future work is aimed at integrating this paradigm into live clinical workflows. Our work isolates CAST as a promising candidate for testing with a medical team, and we are pursuing this direction. Ultimately this future work will inform us of additional real-world deployment challenges.

Finally, our evaluation on HealthBench and other free response benchmarks requires additional computation due to its more complex mechanisms. Future work should consider strategies to alleviate this additional computational burden. Future work may also test CAST on additional samples, datasets, and benchmarks to understand strategies for real world deployment of CAST in the clinical industry and beyond.

References

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. To believe or not to believe your LLM. *arXiv preprint arXiv:2402.13769*.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and

Karan Singhal. 2025. Healthbench: Evaluating large language models towards improved human health. *Preprint*, arXiv:2505.08775.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

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 and the 9th International Joint Conference on Natural Language Processing*, pages 2567–2577.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Mujumdar, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative

refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 6981–7004.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguerre y Arcas, Dale Webster, and 11 others. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed 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*, volume 35, pages 24824–24837.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

A Prompts and Algorithm

A.1 Multiple-Choice Prompts

Phase 1 prompt. You are a senior medical consultant. Think step by step: mechanisms, differentials, eliminate wrong options. Show your reasoning chain for each option considered. End with exactly: FINAL ANSWER: X, where X is A, B, C, or D.

Contrastive adjudication prompt. Original question and answer options: [question and options]. Two independent medical consultants disagreed. Argument for [majority answer]: [strongest

Algorithm 1: Contrastive Adjudication via Selective Triggering (CAST)

Input : Question q , options $\mathcal{O}$, model $\mathcal{M}$, sample sizes K, K_2 , threshold θ , minority share μ .

Output : Final answer $\hat{a}$.

- 1 Sample K Phase-1 reasoning chains from $\mathcal{M}(q, \mathcal{O})$;
 - 2 Extract final answers $\{\hat{a}_1, \dots, \hat{a}_K\}$;
 - 3 Compute vote confidence C and minority share m ;
 - 4 If $C < \theta$ and $m \geq \mu$, cluster Phase-1 trajectories by final answer;
 - 5 Select the strongest argument for each competing answer;
 - 6 Build the contrastive adjudication prompt;
 - 7 Sample K_2 Phase-2 adjudication responses;
 - 8 **return** the majority vote over Phase-2 answers only;
 - 9 If the trigger condition is not met, **return** the majority vote over Phase-1 answers;
-

majority reasoning trace]. Argument for [minority answer]: [strongest minority reasoning trace]. Identify the specific pharmacological, physiological, or anatomical error in the weaker argument. Then commit to the answer with correct first-principles reasoning. End with exactly: FINAL ANSWER: X, where X is A, B, C, or D.

A.2 CAST Algorithm

B HealthBench Free-Response Setup

For HealthBench, the multiple-choice clustering approach does not transfer directly because outputs are free-response clinical answers rather than answer letters. Two responses can use different wording while giving the same clinical advice, and two responses can look similar while missing different safety points. We therefore use a selector to choose the best Phase-1 response, estimate confidence, and identify distinct response paths.

The selector has no access to rubrics, ideal completions, gold information, answer keys, scoring rules, or grader feedback. It sees only the original clinical conversation and the model-generated candidate responses. It returns the best candidate, a confidence score from 0.0 to 1.0, whether adjudication is needed, and up to two distinct response paths. CAST triggers Phase 2 when the selector

says adjudication is needed or confidence is below $\theta = 0.70$.

B.1 HealthBench Prompts

Phase 1 prompt. You are a careful medical assistant responding to the user’s latest message in a health conversation. Use only the conversation. Be medically safe, accurate, concise, and helpful. Give urgent-care advice when appropriate. Acknowledge uncertainty when information is incomplete. Return only valid JSON with exactly two keys:

- **rationale:** a concise high-level clinical justification, not hidden chain-of-thought;
- **final_response:** the exact assistant response to send to the user.

Selector prompt. Original conversation: [conversation]. Candidate responses: [candidate responses with rationales]. Task: Pick the single best candidate response for the user. Estimate confidence from 0.0 to 1.0 that this candidate is clearly best. Say whether adjudication is needed because top candidates contain clinically meaningful disagreements, safety or triage differences, or important omissions. Identify up to two strongest distinct response paths. Each path should cite candidate indices and summarize the path’s strengths and risks. Return only JSON with `best_index`, `confidence`, `needs_adjudication`, `ranking`, and `distinct_paths`.

Contrastive adjudication prompt. Original conversation: [conversation]. Two strongest or most distinct response paths from Phase 1: Path A: [path summary, strengths, risks, candidate rationale, final response]. Path B: [path summary, strengths, risks, candidate rationale, final response]. Selector confidence for Phase-1 best: [confidence]. Selector said adjudication needed: [true/false]. Now perform contrastive adjudication between these paths and produce the best final assistant response for the original conversation. Check whether either path misses a clinically important safety point, triage or red-flag point, medication or contraindication point, or follow-up and uncertainty point. Use only the original conversation and the candidate paths above. Do not use or infer any benchmark rubric, answer key, ideal completion, scoring rule, or grader feedback.

Post-hoc grader prompt. You are grading a HealthBench-style medical answer against

physician-written rubric criteria. Use only the conversation, the candidate response, and the rubric criteria shown here. For each criterion, decide whether it is satisfied. Return only valid JSON.

B.2 Free-Response Trigger Check

We also tested a clustering-style trigger for HealthBench on a 200-question held-out validation set. The selector version had higher average accuracy across the three models. The clustering-style trigger was also much less selective; for GPT-4o-mini, it triggered Phase 2 on more than 94% of questions. These results led us to use the selector method for HealthBench, because it gave a better uncertainty test and avoided triggering Phase 2 on nearly every example.

C Hyperparameter and Threshold Sensitivity

We performed parameter selection on a 200-question held-out MedMCQA validation subset. Table 2 reports accuracy under different values of K and K_2 , with threshold fixed at $\theta = 0.70$. Performance is relatively stable across values of K_2 , suggesting the method is not sensitive to the exact number of contrastive samples. We use $K = 5$ and $K_2 = 8$ in the main experiments as the best performance.

K	K_2	Trigger	Accuracy
3	2	18.5%	69.5%
3	3	18.5%	70.5%
3	5	18.5%	72%
5	3	16%	71%
5	8	16%	73%

Table 2: Sensitivity to hyperparameters K and K_2 on the 200-question held-out MedMCQA validation subset. We fix $\theta = 0.70$ throughout all experiments.

The confidence threshold θ controls the trade-off between triggering rate and selectivity. Lower values trigger fewer questions and approach SC performance. Higher values trigger more questions but can resample confident questions unnecessarily. In our experiments, $\theta = 0.70$ gave the best balance on the same 200-question held-out MedMCQA validation subset.

D Additional Graphs and Tables

The following supplemental table and figures support the main trends reported in Table 1, includ-

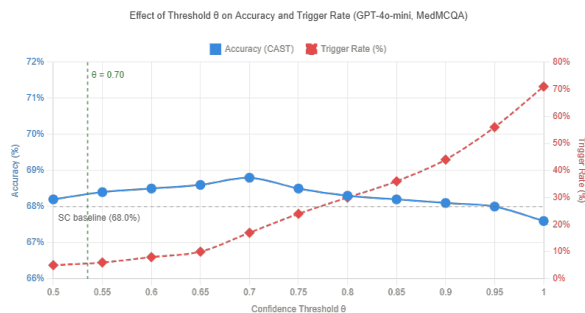


Figure 3: Effect of confidence threshold θ on CAST accuracy and trigger rate on the 200-question held-out MedMCQA validation subset with GPT-4o-mini. Accuracy peaks at $\theta = 0.70$.

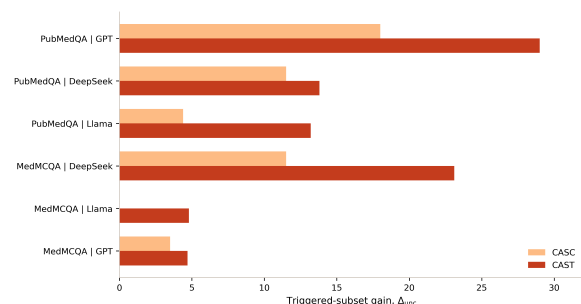


Figure 4: CAST outperforms CASC on triggered questions.

ing CAST’s gains over CASC and its behavior on HealthBench.

Dataset	Model	CASC	CAST
MedMCQA	GPT-4o-mini	+3.5%	+4.7%
MedMCQA	DeepSeek V3	+11.5%	+23.1%
MedMCQA	Llama-4-Maverick	+0.0%	+4.8%
PubMedQA	GPT-4o-mini	+18.0%	+29.0%
PubMedQA	DeepSeek V3	+11.5%	+13.8%
PubMedQA	Llama-4-Maverick	+4.4%	+13.2%

Table 3: Triggered-subset ablation comparing CASC against CAST. CASC pools Phase-1 and Phase-2 votes, while CAST votes only over Phase-2 adjudication responses.

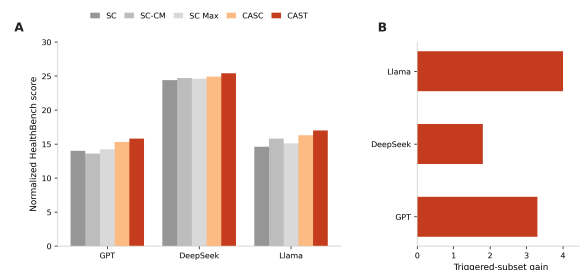


Figure 5: Healthbench results on free-response clinical conversations.

Dataset	Method	Accuracy	Trigger Rate
MedMCQA	CAST (fixed θ)	76.0%	5.5%
MedMCQA	CAST (adaptive θ)	75.5%	3.0%
PubMedQA	CAST (fixed θ)	66.5%	14.0%
PubMedQA	CAST (adaptive θ)	66.0%	14.0%

Table 4: Comparison of the fixed threshold $\theta = 0.70$ against an empirically calibrated, per-bin adaptive threshold on a held-out calibration split disjoint from evaluation. The adaptive threshold yields statistically indistinguishable performance (paired bootstrap), indicating that $\theta = 0.70$ is not a fragile choice.