Do Large Language Models Know When Not to Answer in Medical QA?

Sravanthi Machcha[*] [1], **Sushrita Yerra**[*] [1], **Sharmin Sultana** [2,3]
Hong Yu[†] [1,2,3], **Zonghai Yao**[†] [1,3]
[1]Manning College of Information and Computer Sciences, UMass Amherst, MA, USA
[2]Center for Healthcare Organization and Implementation Research, VA Bedford Health Care
[3]Miner School of Computer and Information Sciences, UMass Lowell, MA, USA
smachcha@umass.edu, sushrithay@gmail.com, zonghaiyao@umass.edu

Abstract

Uncertainty awareness is essential for large language models (LLMs), particularly in safety-critical domains such as medicine where erroneous or hallucinatory outputs can cause harm. Yet most evaluations remain centered on accuracy, offering limited insight into model confidence and its relation to abstention. In this work, we present preliminary experiments that combine conformal prediction with abstention-augmented and perturbed variants of medical QA datasets. Our early results suggest a positive link between uncertainty estimates and abstention decisions, with this effect amplified under higher difficulty and adversarial perturbations. These findings highlight abstention as a practical handle for probing model reliability in medical QA. Our codes will be released.

1 Introduction

Uncertainty is a defining feature of human language: ambiguity, underspecification, and incomplete information are the rule rather than the exception. Nevertheless, most NLP evaluation continues to assume that such ambiguities must be resolved, with accuracy as the dominant metric. This assumption becomes especially problematic in high-stakes domains such as medicine, law, and finance (Thirunavukarasu et al., 2023; Guha et al., 2023; Wu et al., 2023; Achiam et al., 2023; Chang et al., 2024), where overconfident but incorrect answers can cause harm.

Recent advances show that large language models (LLMs) can achieve near-expert performance on many tasks (Achiam et al., 2023), but their reliability hinges not only on being right when confident, but also on knowing when not to answer. In medical QA, for instance, users frequently pose ambiguous or even unanswerable queries (Thirunavukarasu et al., 2023), where calibrated abstention could prevent hallucinations and unsafe recommendations (Kirichenko et al., 2025). Existing benchmarks such as MedQA, MedQA-CS, and MedMCQA (Jin et al., 2021; Yao et al., 2024; Pal et al., 2022) mainly measure accuracy, leaving open the question of whether models can represent and act on their own uncertainty.

At the same time, broader efforts in uncertainty quantification (UQ) for LLMs, such as LM-Polygraph (Fadeeva et al., 2023; Vashurin et al., 2025), have begun to systematize estimation methods and provide unified implementations, while work in medical text analysis (Vazhentsev et al., 2025) highlights selective prediction as a practical approach to safety in diagnosis. These directions reinforce the importance of studying abstention-aware evaluation in medical QA, where ambiguity and incomplete context are unavoidable. We use medical multiple-choice QA as a controlled proxy for clinical decision making: its finite option space yields precise uncertainty sets and abstention rules, and the resulting signals about when to answer or defer carry over to broader medical NLP tasks.

In this work, we present ongoing work on abstention and uncertainty in medical multiple-choice QA. We combine conformal prediction (Angelopoulos et al., 2020) with adversarial perturbations and abstention-augmented questions to probe how models behave under ambiguity. Our preliminary findings suggest a consistent positive association between uncertainty and abstention: when given the explicit option to abstain, models tend to signal higher uncertainty, with effects amplified on more difficult and perturbed questions. We take these results as tentative evidence that abstention can serve as a conservative and responsible mechanism for handling uncertainty in medical QA with LLMs.

[*]Equal contribution, alphabetical order
[†]Co-corresponding authors

Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025), pages 27–35

2 Related Work

Uncertainty Quantification and Conformal Prediction Estimating uncertainty is critical for trustworthy AI, yet common tools such as entropy, calibration, Bayesian inference, and ensembling often miscalibrate or are impractical for black-box LLMs (Fomicheva et al., 2020; Gawlikowski et al., 2023; Abdar et al., 2021; Hu et al., 2023; Wimmer et al., 2023; Kwon et al., 2020; Rahaman et al., 2021). Conformal prediction (CP) offers model-agnostic, statistically grounded guarantees and has shown strong results in NLP and MCQA (Angelopoulos and Bates, 2021; Kumar et al., 2023; Kapoor et al., 2024; Deutschmann et al., 2024; Ye et al., 2024). We extend CP-based evaluation to both open and closed models, linking uncertainty to abstention in real-world MCQA, and situating verbalized confidence and aggregation baselines for black-box LLMs (Tian et al., 2023; Xiong et al., 2023). Beyond CP, frameworks such as LM-Polygraph (Fadeeva et al., 2023; Vashurin et al., 2025) systematize estimation methods and provide extensible evaluation environments, underscoring the growing demand for unified UQ infrastructure.

Abstention, Refusal, and Calibration in LLMs Abstention, understood as deferring under uncertainty, spans from classic classification to modern LLMs (Yin et al., 2023; Wimmer et al., 2023; Amayuelas et al., 2023). Although some benchmarks add explicit abstain or "cannot answer" options, standardized MCQA evaluation, especially for proprietary models, remains scarce (Brahman et al., 2024; Madhusudhan et al., 2024). Existing approaches such as verbalized uncertainty, prompting, finetuning, and post-hoc rejection often show limited calibration or generalization (Lin et al., 2022; Xiong et al., 2023; Chen et al., 2024; Varshney and Baral, 2023; Vashurin et al., 2025). In medicine, selective prediction has been studied as a practical strategy for low-confidence cases, with recent work introducing HUQ-2, a hybrid method that combines aleatoric and epistemic uncertainty across tasks like mortality prediction, ICD coding, and mental health detection (Vazhentsev et al., 2025; Ashfaq et al., 2023; Peluso et al., 2024). These studies show abstention reduces overconfident errors and even supports label-level abstention in multi-label settings. Yet applications to medical QA remain limited, motivating our study. Our QA focus complements classification-centric selective prediction by converting uncertainty into explicit answer-or-abstain decisions that generalize to defer or retrieve policies in clinical NLP.

Reasoning, Prompting, and Hallucination in LLMs Reasoning-tuned models and chain-of-thought (CoT) prompting improve accuracy in math, science, and clinical QA (Zelikman et al., 2022; Luo et al., 2023; Muennighoff et al., 2025; Guo et al., 2025; Cobbe et al., 2021). Yet accuracy-centric evaluation neglects overconfidence and answer-at-all-costs behavior, compounding hallucination risks (Kadavath et al., 2022; Yin et al., 2024; Wen et al., 2025; Huang et al., 2025). Current benchmarks such as AbstentionBench, CONOT, and Abstain-QA mainly emphasize open-domain settings, seldom probing abstention under adversarial or perturbed MCQA or scaling effects (Kirichenko et al., 2025; Brahman et al., 2024; Madhusudhan et al., 2024; Ma et al., 2024; Rahman et al., 2024; Shi et al., 2023). We analyze how prompting and scale interact with abstention reliability in clinical MCQA.

3 Methodology

Our approach focuses on medical multiple-choice question answering (MCQA) tasks, consistent with the evaluation structure of the Open Medical-LLM Leaderboard.[1] The MCQ format is especially suitable for uncertainty analysis via conformal prediction, which requires a well-defined output label space $\mathcal{Y}$ (for more details, see Appendix A).

3.1 Datasets

We select the following medical MCQA datasets for evaluation: **MedQA (USMLE)** (Jin et al., 2021): The MedQA dataset is a large-scale, multiple-choice QA benchmark derived from professional medical licensing exams, typically 4–5 answer options per question. **AMBOSS** (Gilson et al., 2023) [2]: The AMBOSS dataset consists of clinical reasoning items for assessing medical decision-making and—through stratified difficulty annotations—supports systematic study of abstention strategies across difficulty levels; the dataset is private and used for research on medical QA and reasoning.

Dataset variants To evaluate the model's confidence, abstention behavior, and the correlation between the two, we construct multiple dataset vari-

[1] https://huggingface.co/blog/leaderboard-medicalllm

[2] https://www.amboss.com/us

ants. These variants are designed to probe how different conditions—such as missing information or the presence of an abstention option—affect model predictions, combined with the difficulty stratification of the questions.

Abstention This variant, also henceforth referred to as **A** (Abstention Variant), introduces an explicit abstention option to each question, allowing the model to refrain from answering when uncertain.

Perturbing This variant, also henceforth referred to as **NAP** (No-Abstention + Perturbed Variant), aims to assess the model's confidence when essential information is missing.

Abstention + Perturbing This variant, also henceforth referred to as **AP**, combines both abstention and perturbation. The model is presented with questions where some necessary information has been removed, along with the option to abstain from answering.

3.2 Evaluation Metrics

The models are evaluated on the following metrics for each of the datasets and their variants. More details in Appendix A. **Accuracy:** Accuracy measures how often the model's top prediction matches the correct label. **Conformal Prediction:** We compute conformal scores using both the Least Ambiguous Classifier (LAC) and Adaptive Prediction Set (APS) scoring functions. **Abstention Rate:** Abstention rate is the percentage of test instances where the model outputs the abstention option. We report this value for the Abstention and Perturbed Abstention dataset variants.

4 Experiments

We evaluate a broad set of both open-source and closed-source LLMs, spanning multiple architectural families and model scales. This diverse selection allows us to assess the generality of abstention and uncertainty behaviors across different LLM paradigms. Section B provides a comprehensive list of the models used for the study.

Under each experimental condition, models are prompted to output a single answer token (the selected option), and accuracy is computed by comparing this token with the gold label. The logit corresponding to the emitted token, together with the logits for the remaining candidate choices, is then extracted to compute conformal-prediction scores. For closed-source GPT-family models, these scores are derived from the API-exposed `top-logprobs`.

5 Results and Discussion

Comparison of APS and LAC Distributions As shown in Fig.2(a), APS produces tighter, lower-variance set-size distributions than LAC across both datasets, suggesting more stable thresholding. Under AP conditions, APS distributions also crowd near the upper limit, indicating that prediction sets frequently expand to include most options. This compactness carries over when conditioning on correctness (Fig.2(b)), where APS remains less variable, though still skewed toward larger sets for incorrect answers. Together, these patterns suggest that APS offers more consistency in how uncertainty maps to abstention. By contrast, LAC produces broader set-size distributions (Fig.2(a)), with a wider gap between correct and incorrect cases and heavier right tails (Fig.2(b)). This separation is particularly visible in MedQA, where LAC more distinctly highlights error-prone instances. While less stable for thresholding, LAC may therefore be more useful in contexts where surfacing likely mistakes – for example, for human review or triage.

Effect of difficulty across different settings Across settings (Fig. 1), APS behaves like an uncertainty signal: APS–abstention is consistently positive and APS–accuracy consistently negative. Across difficulty levels, the trend is modestly upward but non-uniform, see appendix: C. With difficulty, APS–abstention strengthens in NoCoT, weakens under CoT, is roughly flat in few-shot, and ticks up under perturbations (and mildly in zero-shot/not-perturbed). APS–accuracy grows more negative with difficulty for NoCoT/zero-shot/perturbed runs, but becomes less negative under CoT and is flatter when not perturbed.

LAC exhibits a nuanced profile: LAC–accuracy is consistently negative, whereas LAC–abstention is prompt-dependent—positive under NoCoT but declining with difficulty; under CoT it is slightly negative at d1–d2, 0 at d3, and only mildly positive by d4–d5. Few-shot mainly increases error risk (more negative APS–/LAC–accuracy) and has small, non-monotonic effects on APS–abstention; for LAC–abstention it remains positive in NoCoT but 0/negative on easy CoT. Overall, CoT weakens the set-size–abstention link, though both metrics still flag accuracy risk. Perturbations heighten uncertainty:they raise APS–abstention, make APS–accuracy more negative, and shift LAC similarly.

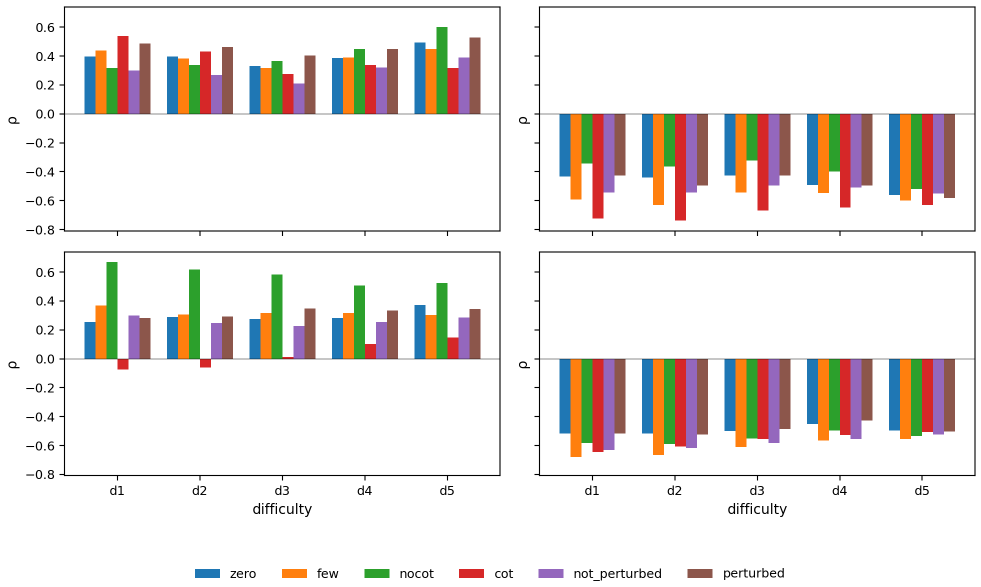

Figure 1: Grouped-bar correlations (Spearman ρ) across difficulty (d1–d5) and settings. Panels: (TL) APS - abstention, (TR) APS - accuracy, (BL) LAC - abstention, (BR) LAC - accuracy. Desired pattern: (TL,BL) positive and (TR,BR) negative.

6 Conclusion

In this study, we asked how item difficulty shapes model uncertainty and abstention, and how two set-size signals: LAC and APS, serving as uncertainty proxies across prompting style (CoT vs. No-CoT), demonstration count (zero- vs. few-shot), and input perturbations. We find a strong, positive uncertainty–abstention relationship and a consistently negative association between both APS/LAC and accuracy. Averaged over datasets, increasing difficulty does not materially change aggregate uncertainty or abstention. Practically, APS is a reliable gate for abstention across conditions, while LAC is a robust indicator of accuracy risk whose coupling to abstention weakens with CoT: especially on easier items. APS produces tighter, more stable distributions, whereas LAC yields clearer separation between correct and incorrect answers, suggesting complementary strengths in threshold-

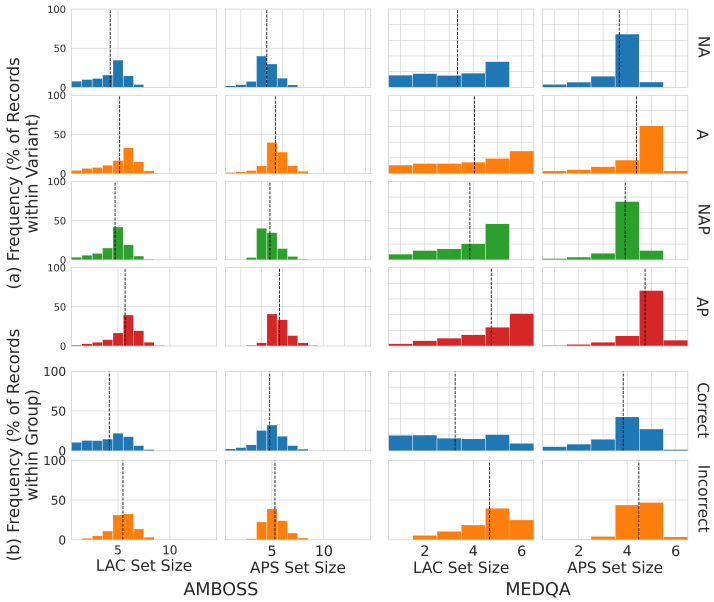

Figure 2: Distribution of conformal-prediction set sizes: (a) across variants NA, A, NAP, and AP, (b) by correctness for both datasets; mean shown as dashed line

ing versus error triage.

7 Limitations

First, the study is confined to English-language datasets, limiting its applicability to multilingual or non-English medical contexts. Expanding the benchmark to additional languages and healthcare systems is essential for broader relevance.

Second, while both open- and closed-source LLMs across diverse architectures and scales are included, the coverage is inherently finite. Given the rapid evolution of model capabilities and training paradigms, the reported findings may not generalize to future or unreleased models.

Finally, the evaluation centers on multiple-choice QA, whose structured label space facilitates conformal prediction and abstention analysis. However, this focus overlooks the complexity of real-world clinical reasoning and open-ended tasks, where uncertainty manifests differently. Extending abstention-aware evaluation to generative, free-form, and multi-modal settings remains a key direction for future work.

8 Ethics Statement

In this study, we examine large language models for medical question answering with a particular focus on abstention and uncertainty. Evaluation is carried out on two datasets: the publicly available MedQA benchmark and a proprietary clinical QA set provided by AMBOSS. MedQA is openly distributed for research, whereas access to AMBOSS is restricted by license and the dataset is used exclusively for internal evaluation under the terms of a research agreement.

The experiments rely solely on de-identified or synthetic exam-style material; no patient-identifiable data are involved. All procedures follow established ethical standards for research using such resources. Our goal is to advance the safe and reliable use of LLMs in high-stakes medical contexts, emphasizing mechanisms to counter overconfidence and hallucination. The datasets, benchmark variants, and analyses are intended strictly for research purposes and are not designed for direct integration into clinical workflows.

Although abstention mechanisms can reduce the risk of severe errors, they do not eliminate bias or inaccuracy, as models may still reproduce artifacts from their training data or benchmarks. Accordingly, abstention should be seen as a supplement to—not a replacement for—clinical validation and human oversight.

All models and APIs are employed in their original, unmodified form, and any subsequent use of the benchmark must respect the corresponding licenses and terms of service. To support transparency and reproducibility, we release the benchmark and code under a CC-BY-NC 4.0 license. The AMBOSS dataset itself is not part of this release.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Wang. 2023. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *arXiv preprint arXiv:2305.13712*.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. 2020. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Anastasios N Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Awais Ashfaq, Markus Lingman, Murat Sensoy, and Sławomir Nowaczyk. 2023. Deed: Deep evidential doctor. *Artificial Intelligence*, 325:104019.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, et al. 2024. The art of saying no: Contextual noncompliance in language models. *Advances in Neural Information Processing Systems*, 37:49706–49748.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Lida Chen, Zujie Liang, Xintao Wang, Jiaqing Liang, Yanghua Xiao, Feng Wei, Jinglei Chen, Zhenghong Hao, Bing Han, and Wei Wang. 2024. Teaching large language models to express knowledge boundary from their own signals. *arXiv preprint arXiv:2406.10881*.

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

Nicolas Deutschmann, Marvin Alberts, and María Rodríguez Martínez. 2024. Conformal autoregressive generation: Beam search with coverage guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11775–11783.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. 2023. Lm-polygraph: Uncertainty estimation for language models. *arXiv preprint arXiv:2311.07383*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR medical education*, 9(1):e45312.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in natural language processing: Sources, quantification, and applications. *arXiv preprint arXiv:2306.04459*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

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.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katie Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew G Wilson. 2024. Large language models must be taught to know what they don't know. *Advances in Neural Information Processing Systems*, 37:85932–85972.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J Bell. 2025. Abstentionbench: Reasoning llms fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*.

Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. 2020. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Jingyuan Ma, Damai Dai, Zihang Yuan, Weilin Luo, Bin Wang, Qun Liu, Lei Sha, Zhifang Sui, et al. 2024. Large language models struggle with unreasonability in math problems. *arXiv preprint arXiv:2403.19346*.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2024. Do llms know when to not answer? investigating abstention abilities of large language models. *arXiv preprint arXiv:2407.16221*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Alina Peluso, Ioana Danciu, Hong-Jun Yoon, Jamaludin Mohd Yusof, Tanmoy Bhattacharya, Adam Spannaus, Noah Schaefferkoetter, Eric B Durbin, Xiao-Cheng Wu, Antoinette Stroup, et al. 2024. Deep learning uncertainty quantification for clinical text classification. *Journal of biomedical informatics*, 149:104576.

Rahul Rahaman et al. 2021. Uncertainty quantification and deep ensembles. *Advances in neural information processing systems*, 34:20063–20075.

AM Rahman, Junyi Ye, Wei Yao, Sierra S Liu, Jesse Yu, Jonathan Yu, Wenpeng Yin, and Guiling Wang. 2024. From blind solvers to logical thinkers: Benchmarking llms' logical integrity on faulty mathematical problems. *arXiv preprint arXiv:2410.18921*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Neeraj Varshney and Chitta Baral. 2023. Post-abstention: Towards reliably re-attempting the abstained instances in qa. *arXiv preprint arXiv:2305.01812*.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, et al. 2025. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Artem Vazhentsev, Ivan Sviridov, Alvard Barseghyan, Gleb Kuzmin, Alexander Panchenko, Aleksandr Nesterov, Artem Shelmanov, and Maxim Panov. 2025. Uncertainty-aware abstention in medical diagnosis based on medical texts. *arXiv preprint arXiv:2502.18050*.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. 2023. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in artificial intelligence*, pages 2282–2292. PMLR.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, et al. 2024. Medqa-cs: Benchmarking large language models clinical skills using an ai-sce framework. *arXiv preprint arXiv:2410.01553*.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. In *Advances in Neural Information Processing Systems*, volume 37, pages 15356–15385. Curran Associates, Inc.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Xiaonan Li, Junqi Dai, Qinyuan Cheng, Xuan-Jing Huang, and Xipeng Qiu. 2024. Reasoning in flux: Enhancing large language models reasoning through uncertainty-aware adaptive guidance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2401–2416.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

A Evaluation Metrics

Conformal Prediction Conformal Prediction (CP) provides a statistically rigorous way to quantify uncertainty (Angelopoulos and Bates, 2021). Given a model f and a test instance x_t, we compute a *prediction set* $C(x_t) \subseteq \mathcal{Y}$ of plausible answers such that:

$$P(y_t \in C(x_t)) \geq 1 - \alpha$$

where α is a user-set error rate. The size of the prediction set, or **Set Size (SS)**, reflects the model's confidence: $|C(x_t)| = 1$ implies highest confidence, and larger sets reflect higher uncertainty.

We compute conformal scores using both the Least Ambiguous Classifier (LAC) and Adaptive Prediction Set (APS) scoring functions:

1) Adaptive Prediction Set (APS)

$$\text{APS: } s(x, y) = \sum_{y' : f(x)_{y'} \geq f(x)_y} f(x)_{y'}$$

2) Least Ambiguous Classifier (LAC)

$$\text{LAC: } s(x, y) = 1 - f(x)_y$$

where $f(x)_y$ is the probability assigned to label y. Using a calibration set, we compute a quantile threshold $\hat{q}_\alpha$ and define the conformal prediction set for each test instance x as:

$$C(x) = \{y \in \mathcal{Y} \mid s(x, y) \leq \hat{q}_\alpha\}$$

where $\hat{q}_\alpha$ is the $(1 - \alpha)$ quantile of calibration scores.

B Experiment Models

To evaluate performance across varying model scales and architectural families, we benchmark a diverse set of both open-source and closed-source models, listed below:

Open-source Models:

- **LLaMA Family:** [3] [4] Llama3.2-1B-Instruct, Llama3.2-3B-Instruct, Llama3.1-8B-Instruct
- **Phi Family:** Phi-4-mini[5], phi-4[6]
- **Qwen Family:** [7] [8] Qwen2.5-0.5B-Instruct, Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-32B-Instruct, Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen3-32B
- **Gemma Family:** gemma-3-4b[9], medgemma-4b-it[10]

Closed-source Models:

- **GPT Family:** gpt-4.1-nano-2025-04-14, gpt-4.1-mini-2025-04-14, gpt-4o-mini-2024-07-18, gpt-4o-2024-08-06, gpt-4.1-2025-04-14

C Additional Results Discussion

Across all comparisons, abstention has a slight uptrend with difficulty but not a consistent increase, and variance grows under perturbation and at the highest difficulty levels, consistent with greater heterogeneity or fewer items per stratum. Effects differ in magnitude across conditions: CoT exerts limited influence on abstention relative to perturbation (↑ abstention) and few-shot prompting (↓ abstention), but it meaningfully alters how set size relates to the decision to abstain—especially on easier items. For completeness, it is useful to pair these correlation patterns with risk–coverage or selective-accuracy summaries by stratum, to verify that improved error signaling from APS/LAC translates into better risk control at comparable coverage.

C.0.1 Zero-shot vs Few-shot

As can be seen from 3, Providing demonstrations amplifies error signaling more than abstention signaling. Few-shot runs systematically make both APS–accuracy and LAC–accuracy more negative across difficulty strata, indicating that larger prediction sets track error risk more faithfully when demonstrations are present. By contrast, the effect on APS–abstention is small and irregular with difficulty, suggesting that demonstrations primarily reshape confidence within the commit region rather than pushing the model to defer. For LAC–abstention, the NoCoT condition preserves a positive association across strata, whereas under

[3] https://huggingface.co/collections/meta-llama/llama-32-66f448ffc8c32f949b04c8cf

[4] https://huggingface.co/meta-llama/Llama-3.1-8B

[5] https://huggingface.co/microsoft/Phi-4-mini-instruct

[6] https://huggingface.co/microsoft/phi-4

[7] https://huggingface.co/collections/Qwen/qwen25-66e81a666513e518adb90d9e

[8] https://huggingface.co/collections/Qwen/qwen3-67dd247413f0e2e4f653967f

[9] https://huggingface.co/google/gemma-3-4b-it

[10] https://huggingface.co/google/medgemma-4b-it

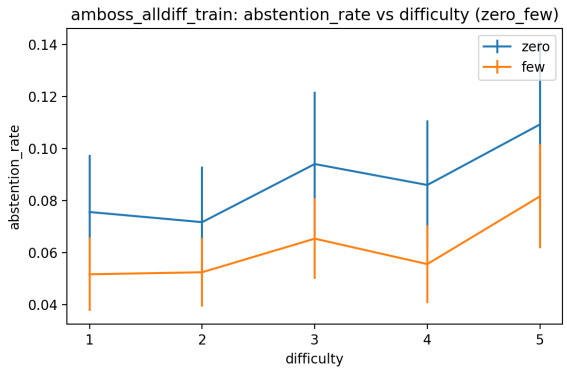

Figure 3: Amboss zero v few shot performance across difficulty settings d1-d5

CoT the easy-difficulty caveat persists (near-zero or slightly negative at d1–d2) before turning weakly positive by d4–d5. Together, these patterns imply that demonstrations improve the calibration of which answers are risky without uniformly increasing the tendency to abstain.

C.0.2 Cot vs No Cot

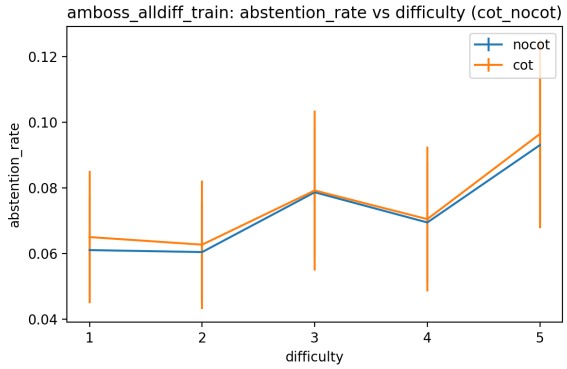

Figure 4: Amboss cot v nocot performance across difficulty settings d1-d5

Fig 4 demonstrates reasoning changes how prediction-set size maps to the abstain decision. APS remains positively associated with abstention with or without CoT, but its strength diminishes with difficulty under CoT while increasing in NoCoT. This indicates that generating rationales encourages commitment on harder items even when the prediction set is larger, possibly because intermediate reasoning consolidates probability mass on a preferred candidate. For LAC, CoT partially decouples set size from abstention at easy levels: the model may explore more candidates yet still commit, so larger LAC does not reliably imply greater deferral at d1–d2; only by d4–d5 does the LAC–abstention link re-emerge as mildly positive. Importantly, APS–accuracy and LAC–accuracy remain negative in all cases, so both set sizes continue to flag accuracy risk even when CoT reduces their influence on abstention behavior.

C.0.3 Perturbed vs Not Perturbed

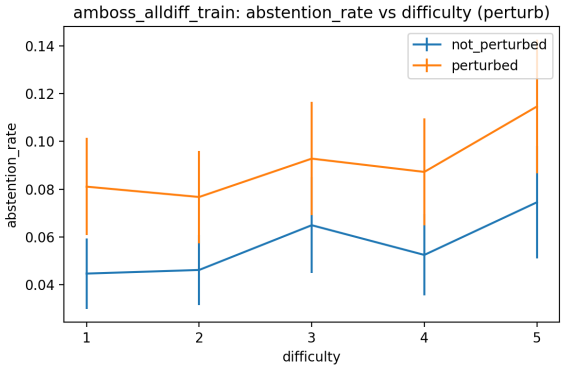

Figure 5: Amboss pert v nopert performance across difficulty settings d1-d5

Noise generally sharpens uncertainty signals and increases deferral as observed in 5. Perturbations raise APS–abstention at every difficulty level and make APS–accuracy more negative, with the largest shifts at higher difficulties. For LAC, perturbations push correlations in the same directions—more positive with abstention (especially in NoCoT) and more negative with accuracy overall—consistent with broader or less concentrated prediction sets under input shift. The CoT interaction holds: under CoT, LAC remains a weak abstention trigger at easier difficulties even as its negative relation to accuracy persists, indicating that reasoning can sustain commitment under mild noise while still reflecting error risk in set size.