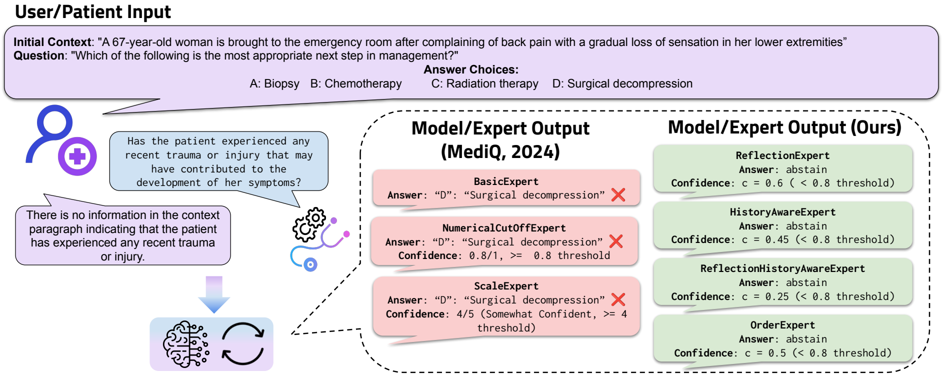
Med-CAM: Improving Medical Question Answering with Confidence-Aware Methods

Karina Halevy[*], **Kshitish Ghate**[*], **Jimin Mun, Mona Diab, Maarten Sap**
Carnegie Mellon University
msap2@andrew.cmu.edu

Abstract

Large language models (LLMs) show promise in medical question answering (QA) but often produce overconfident and incorrect responses due to a lack of calibrated uncertainty estimation. In this work, we explore confidence-aware methods (CAM) to improve LLM performance in multi-turn medical QA (Li et al., 2024c). We propose four techniques for eliciting confidence—three prompting-based and one log-probability-based—and integrate them into a medical QA framework. Experiments across four open-source LLMs from different families and sizes show that log-probability-based methods tend to outperform prompting-based approaches and existing baselines. Additionally, we find that LLMs frequently exhibit high confidence when prompted, even for incorrect answers, highlighting persistent miscalibration. These findings demonstrate that using output probabilities to elicit model confidence in a calibrated manner can significantly enhance their trustworthiness in medical QA.[*]

1 Introduction



Initial Context: "A 67-year-old woman is brought to the emergency room after complaining of back pain with a gradual loss of sensation in her lower extremities"
Question: "Which of the following is the most appropriate next step in management?"
Answer Choices:
A: Biopsy B: Chemotherapy C: Radiation therapy D: Surgical decompression

Has the patient experienced any recent trauma or injury that may have contributed to the development of her symptoms?

There is no information in the context paragraph indicating that the patient has experienced any recent trauma or injury.

BasicExpert
Answer: "D": "Surgical decompression" ❌

NumericalCutOffExpert
Answer: "D": "Surgical decompression" ❌
Confidence: 0.8/1, >= 0.8 threshold

ScaleExpert
Answer: "D": "Surgical decompression" ❌
Confidence: 4/5 (Somewhat Confident, >= 4 threshold)

ReflectionExpert
Answer: abstain
Confidence: c = 0.6 (< 0.8 threshold)

HistoryAwareExpert
Answer: abstain
Confidence: c = 0.45 (< 0.8 threshold)

ReflectionHistoryAwareExpert
Answer: abstain
Confidence: c = 0.25 (< 0.8 threshold)

OrderExpert
Answer: abstain
Confidence: c = 0.5 (< 0.8 threshold)

Figure 1: Example from MediQ dataset in which BasicExpert, ScaleExpert, and NumericalCutOffExpert all confidently answer a question incorrectly, while ReflectionExpert, HistoryAwareExpert, ReflectionHistoryAwareExpert, and OrderExpert appropriately abstain from answering.

Large Language Models (LLMs) are increasingly used for medical question answering (QA)—users include patients seeking medical advice and medical providers seeking to make diagnoses, recommend treatments, and plan patient care (Meng et al., 2024; Presiado et al., 2024; Spotnitz et al., 2024; Kim et al., 2025). Such QA tasks typically require multi-turn interactions in practice, as answers must only be generated if (1) there is sufficient

[*]Equal contribution
[*]Code: https://github.com/ENSCMA2/webmd

information available and (2) confidence is high enough for a proposed answer Bornstein & Emler (2001); Masic (2022); Trimble & Hamilton (2016). When these conditions are not met, a language model should abstain from answering and instead pose a follow-up question to get closer to criterion (1). To meet criterion (2), models should have relatively low confidence in any proposed answer if they have insufficient patient information, as reliable confidence signals are necessary for clinical decision-making Kell et al. (2024).

A major challenge is that current LLMs are often overconfident (Qin et al., 2024; Zhou et al., 2024) and trained to respond to nearly any query, regardless of uncertainty (Leng et al., 2024). Most medical QA systems also fail to communicate model confidence Kell et al. (2024), risking misleading recommendations when crucial information is missing. While confidence-based abstention has shown promise in interactive medical settings (Li et al., 2024c), existing approaches do not leverage models' internal confidence to guide abstention. Moreover, recent work on uncertainty has focused on user perception or post-hoc calibration (Xu et al., 2025) rather than integrating confidence into model decision-making. This leaves a critical gap: current medical QA systems lack mechanisms to self-assess and signal uncertainty Kell et al. (2024).

Our work addresses this gap by introducing and empirically evaluating the promise of medical QA with confidence-aware methods (Med-CAM) with LLMs. Focusing on realistic, interactive medical QA, where user queries may be incomplete, we use the iMedQA component of the MediQ dataset (Li et al., 2024c). As shown in Figure 1, iMedQA models the situation faced in many real-world medical encounters—it benchmarks iterative interactions with simulated patients that reveal important details only upon further questioning. We therefore measure model confidence throughout the process with increasing context and integrate this metric to model decisions on abstention not only at each turn but also at the end of an interaction.

To comprehensively evaluate confidence-aware medical QA, we introduce novel adaptations of four confidence-aware methods that span both prompting-based and log-probability-based strategies for confidence measurement. Inspired by previous works on prompting-based strategies for reasoning (Yao et al., 2023), self-refinement (Madaan et al., 2023), and confidence elicitation (Xiong et al., 2023; Tian et al., 2023), we propose novel adaptations using self-reflection of alternate answers (ReflectionExpert), a model's own history of confidence (HistoryAwareExpert), and both (ReflectionHistoryAwareExpert). While prompting is required for black-box models, a model's internal state should align with its outputs and reflect the current context rather than pretraining biases (Zhao et al., 2021). To this end, we propose a log-probability-based confidence measure that captures patient context-induced confidence shifts (OrderExpert).

We evaluate our proposed methods against three baselines from Li et al. (2024c) using four open-source LLMs of varying families and sizes. OrderExpert, our log-probability-based method, outperforms prompt-based approaches by up to 27.84 accuracy percentage points on three of four models after filtering first-turn abstentions. ReflectionHistoryAwareExpert, our prompt-based method, performs best when filtering final-turn abstentions. Further analysis reveals that models are more accurate and confident when answering diagnostic questions than those about causality or treatment steps, and they are sensitive to how open-ended the initial query is. By exploring methods to improve LLM confidence calibration for response abstention, our study outlines a path towards safer triage, telehealth, and clinical decision support tools. The same calibration primitives can be grafted onto larger proprietary models, helping vendors satisfy AI regulatory requirements.[*]

2 Related Work

Our work aims to bridge two well-established areas of research: medical QA and factuality and confidence in general QA. This section reviews the most relevant prior work in each.

[*] https://artificialintelligenceact.eu/

2.1 Medical QA

Overall, work in medical question-answering has evolved from creating simple text-based retrieval methods to developing conversational and interactive AI agents, often based on LLMs. Some popular state-of-the-art LLMs fine-tuned to answer medical questions include Med-PaLM Singhal et al. (2023a), BioMedLM 2.7B Bolton et al. (2024), MedPaLM 2 Singhal et al. (2023b), MedAlpaca Han et al. (2025), and ChatDoctor Li et al. (2023). Along with models, researchers have produced various benchmark datasets for medical QA, including MultiMedQA Singhal et al. (2023a), PMC-VQA Zhang et al. (2024), and Medical Meadow Han et al. (2025). While these datasets mostly follow medical exam question formats in which all information is presented upfront in a single-turn format, MediQ Li et al. (2024c) breaks new ground by creating and demonstrating the importance of a multi-turn medical QA benchmark that simulates real-world doctor-patient interactions more closely.

Various innovative methods have also been used to optimize LLMs for medical QA at multiple stages, from pre-training to fine-tuning to inference-time retrieval and agentic collaboration. For example, Yasunaga et al. (2022) introduce DRAGON, a self-supervised approach to the pre-training of the joint language knowledge model that could be applied to medicine. Other approaches include instruction prompt tuning Singhal et al. (2023a), medical domain knowledge fine-tuning with ensemble refinement Singhal et al. (2023b), self-directed real-time info retrieval Li et al. (2023), and an LLM-collaboration approach called MDAgents Kim et al. (2024). However, these previous works focus only on answer factuality and lack assessment and integration of model confidence. This confidence assessment is important because it can better inform how an LLM proceeds in multi-turn interactions.

2.2 Factuality & Confidence in QA

Outside of the medical domain, several works have addressed trustworthiness through factuality and confidence benchmarking. Zhou et al. (2024) highlight that LLMs often avoid expressing uncertainty, even when producing incorrect outputs. To improve factuality, prior methods include clustering semantically equivalent answers by repeated sampling Li et al. (2024b), fact-checking answers against authoritative sources Lopez-Martinez (2024), prompting models to justify and integrate candidate answers Li et al. (2024a), reweighting answers based on explanation support and logical consistency Becker & Soatto (2024), using verification questions derived from explanations to assess answer consistency Wu et al. (2024), and benchmarking model confidence with a calibration prompt that replaces input context with a null phrase Zhao et al. (2021). When none of the above methods work to improve answer factuality, works such as Feng et al. (2024) propose having an LLM abstain from answering altogether to avoid false answers. However, there still lacks extensive integration and experimentation of above factuality, confidence estimation, and abstention decision methods in the medical context.

3 Methodology

3.1 Task Description

We frame our study in the context of interactive medical question answering (iMEDQA, as developed by Li et al. (2024c)), a variant of the MediQ benchmark in which an LLM plays the roles of (1) a PATIENT who has access to their own records but reveals only their initial question and context at first and (2) an EXPERT clinician who interacts with the simulated patient, potentially asking follow-up questions to elicit further patient records, to answer the patient's initial question. At the start of every case, the PATIENT only gives the EXPERT minimal patient demographics and consequential case information. The EXPERT must then decide, at each turn, whether to (i) ask the PATIENT for additional information, (ii) commit to one of four multiple choice answers, or (iii) abstain when evidence is insufficient based on a confidence score. Whenever the EXPERT answers with the best-choice label, it is additionally required to output a free-text rationale. Since each follow-up question reveals new clinical facts, the setting allows us to evaluate both the factual accuracy and the model's ability to track and calibrate its own uncertainty over time. Our goal is to design well-calibrated confidence estimation strategies that can guide a model in deciding when to abstain from responding in high-stakes clinical decision-making scenarios.

3.2 Proposed Methods

We implement three baseline methods from Li et al. (2024c) (one basic baseline, two reported as higher-performing) and four novel methods. For each method, we describe the prompt, the process of calculating a confidence score $\tau \in [0, 1]$, and the rule to determine whether the model would abstain from answering the question. Prompt examples for novel methods are in Appendix B.

Basic Baseline: BASICEXPERT. As implemented by Li et al. (2024c), the model is asked to either generate a follow-up question or produce an answer. For our study, we consider this the baseline approach where no abstention policy is applied post-hoc.

High-Performing Baseline: SCALEEXPERT. As implemented by Li et al. (2024c), the model is additionally given definitions of confidence levels on a 5-point Likert scale and is asked to express its confidence by selecting a rating R on the scale (1 is lowest, 5 is highest). τ is then calculated as $\frac{R}{5}$, and we abstain if $\tau < 0.8$.

Continuous Numerical Confidence Baseline: NUMERICALCUTOFFEXPERT. As implemented by Li et al. (2024c), the model prompt additionally includes an instruction to generate a numerical confidence score between 0 and 1 following the methodology of Tian et al. (2023), and τ is set to that score. We again consider $\tau < 0.8$ to be an abstention.

Novel Method, Prompt-Based: REFLECTIONEXPERT. For our first novel method, we adapt self-reflection in QA from Madaan et al. (2023), with our novelty lying in adapting this method to produce a calibrated confidence score at the end of each interaction turn. The method forces the EXPERT to produce an initial answer and then adopt a second opinion stance that actively searches for evidence and alternative diagnoses. The final step asks the model to reconcile its thought process and emit a calibrated confidence score to determine whether it should abstain. Multiple reasoning passes encourage deeper analysis and force the model to update confidence scores and subsequent answer choices based on considering alternative evidence and critiquing its earlier responses. Consistent with the baselines, we fix $\tau < 0.8$ as an abstention.

Novel Method, Prompt-Based: HISTORYAWAREEXPERT. While REFLECTIONEXPERT focuses on horizontal breadth, HISTORYAWAREEXPERT adds a temporal dimension such that at every turn the model recaps how its confidence has evolved, identifies which patient facts drove large swings, and justifies any remaining uncertainty. By having the expert explicitly consider its own meta-reasoning and confidence history, the EXPERT can down-weight stale or spurious early estimates of confidence. This method is conceptually similar to methods introduced by Li et al. (2025) and Wang et al. (2024), as it leverages the evolving confidence signal across turns. Uniquely, our method also requires the model to explicitly verbalize the causes of its confidence shifts within each turn, enabling temporally aware self-recalibration.

Novel Method, Prompt-Based: REFLECTIONHISTORYAWAREEXPERT. This method combines reflection-based reasoning with historical awareness. We prompt the model perform multi-step reflection to analyze evidence and alternatives for a given patient scenario and contextualize that analysis within its longitudinal historical confidence trajectory to recalibrate its assessment. Finally, a unifying reasoning pass makes decisions based on both reflection and historical performance to produce a “calibrated” τ, for which $\tau < 0.8$ is an abstention.

Novel Method, Logprob-Based: ORDEREXPERT. This method builds on the idea from Zhao et al. (2021) that a high classification probability is not meaningful if a model gives the same probability on an input with a nonsensical or null context. An analysis of the confusion matrices for BASICEXPERT on the Llama models, shown in Figure 4, also shows that there may be some bias toward predicting choice "A" even when the true distribution does not lean toward choice "A." This motivates a calibration method to adjust for such ordinal choice bias. While Zhao et al. (2021) create a calibration method for binary classification, we adapt their method to compare probabilities of *entire generated token sequences* given substantive vs. null input context. In particular, given question Q and substantive patient context C, we use the same prompt template as BASICEXPERT to verbalize Q and C into an input prompt P_0. Then, we create a second prompt P_1 that is identical to P_0, except that C is replaced by the string "N/A." We obtain LLM outputs $[t_1, ... t_{n_0}]$ from P_0 and $[s_1, ..., s_{n_1}]$ from P_1, where each t_i and s_i is an output token. Let $L(t_i)$ be the log probability of generated token t_i, and let $m = \min(n_0, n_1)$. We then compute an unnormalized confidence score

$$r = \frac{\sum_{i=1}^{m} \exp(L(t_i) - L(s_i))}{m}, \tag{1}$$

which represents the average scalar ratio r between probabilities of output tokens in the substantive answer vs. the calibration answer. A ratio r of 1 means that the model is equally confident about the actual answer versus the answer without input context. Hence, we empirically seek a minimum threshold $r_0 > 1$ for each model and make the decision to abstain if $r \leq r_0$ for a given entry. We use the methodology in Li et al. (2024c) to force an intermediate answer in any case so that we can measure hypothetical performance given any r_0. We also compute and report a normalized score

$$\tau = \frac{r}{1 + r}. \tag{2}$$

τ is always in the range $[0, 1)$ when $L(t_i)$ values are log probabilities, and as with r, a higher value of τ corresponds to higher confidence. Our abstention rule is dynamically adjusted for each model based on its dev set performance: we set τ to be the value that produces the maximum dev set accuracy, within the constraints that at least three questions are answered. For certain deeper analyses, we adjust τ to secondary or tertiary peaks if we require a larger coverage quantity for analysis.

4 Experiments

4.1 Datasets

We use the MediQ dataset Li et al. (2024c), built on MED-QA Jin et al. (2020), with 10,178 training and 1,272 development samples. To better mimic clinical workflows, we use its interactive variant, iMEDQA, where the EXPERT initially receives only partial patient information (age, gender, chief concern). Additional details (e.g., symptoms, history, exam findings) are hidden unless explicitly requested from the PATIENT.

4.2 Models

We experiment with four open-source models that allow us to (a) access model internals for calibration and (b) assess both model size and model family. In particular, we experiment on Meta's `Llama-3.2-3B-Instruct`[*] and `Llama-3.1-8B-Instruct-Turbo`,[*] as well as Alibaba's `Qwen-2.5-3B-Instruct`[*] and `Qwen-2.5-7B-Instruct-Turbo`.[**]

4.3 Experimental Setup

We fix INSTRUCTPATIENT by Li et al. (2024c) as the PATIENT—a prompt template that asks a patient to truthfully respond to the doctor's questions using only the context given in the

[*] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[*] https://www.together.ai/models/llama-3-1

[*] https://huggingface.co/Qwen/Qwen2.5-3B-Instruct

[*] https://www.together.ai/models/qwen2-5-7b-instruct-turbo

[*] The `Turbo` models are quantized to FP8 via Together AI.

patient record. We use INSTRUCTPATIENT because we find that it has the highest accuracy of all PATIENTs from Li et al. (2024c), based on a preliminary analysis on LLAMA-3.1-8B-INSTRUCT. Thus, our experiments only vary LLMs and EXPERT prompting methods.

Baseline. We aim to quantify how confidence calibration affects performance on iMEDQA. We track whether instructing a model to show its uncertainty leads to more cautious and accurate answers, and whether an abstention feature can prevent overconfident mistakes. We therefore compare our proposed methods with the baselines described in §3.2. We also note that the previous best performance was achieved by GPT-4 as reported in Li et al. (2024c). When given all patient details upfront, GPT-3.5 and GPT-4 reach about 55.8% and 80% accuracy, respectively. If the model does not see patient information, performance drops substantially (GPT-3.5 to 36.7%, GPT-4 to 42.2%). These scenarios provide key reference points for gauging how interactive or confidence-based methods compare.

Metrics. Our primary set of metrics for the multi-class classification task of MediQ is the **abstention-excluded performance** of each model and EXPERT. Given the entire iMEDQA dataset D, we compute the abstention-excluded dataset $D_{A,M,E}$ as the entries for which model M, following the decision rule from EXPERT E, *is sufficiently confident to answer the question* rather than abstaining. We report accuracy on $D_{A,M,E}$. The motivation for these metrics is that in the real world, we would implement a decision rule for a model to abstain from answering a question if it is not sufficiently confident—thus, the performance on $D_{A,M,E}$ better reflects how a system would perform in real-world use cases than performance on D, which would include questions that the model would not answer when deployed.

We work with two variants of $D_{A,M,E}$: $D_{A,M,E,0}$ and $D_{A,M,E,-1}$. For all prompt-based methods, we compute performance on $D_{A,M,E,-1}$, which is the subset of D that would have been answered after the *last* turn of the multi-turn PATIENT-EXPERT interaction, with the number of turns capped at 10 follow-up questions, following Li et al. (2024c). This multi-turn analysis aligns with that of Li et al. (2024c), extending their assessment of abstention as a tool to trigger further information seeking from the EXPERT. In contrast, for ORDEREXPERT, we only implement one conversational turn, thus computing performance on $D_{A,M,E,0}$, the subset of D that would have been answered after the *first* turn of the interaction. We consider $D_{A,M,E,0}$ because we also aim to assess abstention as a tool to indicate a need for human intervention in its own right, not necessarily as a tool for an LLM to continue asking questions and potentially compound errors in simulated questions and responses throughout further interaction turns. This analysis is useful for end-users with resource constraints—especially since ORDEREXPERT requires two model generations per entry and requires extra vector computations to obtain τ, an end-user has a substantial chance of simply using the system once to see if their question is easily answerable and then conducting their own further research if not. For comparison with ORDEREXPERT, we also report the performance on $D_{A,M,E,0}$ for the method with the best performance on $D_{A,M,E,-1}$.

In addition to the abstention-excluded performance, we also report accuracy on all of D, with answers forced out of each entry regardless of abstention decision as used in Li et al. (2024c). We present these metrics alongside abstention-excluded performance to assess the quality of each type of confidence calibration—higher abstention-excluded vs. overall performance indicates a method that does well in reflecting lack of knowledge through lack of confidence, and the converse is true if abstention-excluded performance is lower.

Calibration Metrics. To quantify how faithfully an EXPERT's confidence scores reflect its true correctness likelihood, we report the Expected Calibration Error (ECE). We divide predictions into 10 equal-width confidence bins, compute the absolute gap between average confidence and empirical accuracy within each bin, and weight each gap by the bin's relative frequency. The resulting scalar ranges from 0 (perfect calibration) to 1 (maximal mis-calibration). Lower values indicate that the model's probabilities are more reliable for downstream decisions such as abstaining to answer. Intuitively, ECE measures how far predicted probabilities deviate from the observed success rates.

5 Results

We present the performance of our CAM on iMedQA. We show abstention-excluded performance on the last interaction turn for prompt-based methods ($D_{A,M,E,-1}$) and on the first interaction turn ($D_{A,M,E,0}$) for the best prompt-based method alongside OrderExpert. We also analyze our results by various semantic and syntactic features of input questions. While we present overall results on all models, we focus our in-depth analysis on the best-performing model, `Llama-3.1-8B-Instruct-Turbo`.

5.1 Overall Results

Expert	`Llama-3B`		`Llama-8B`		`Qwen-3B`		`Qwen-7B`	
	Pre	**Post**	**Pre**	**Post**	**Pre**	**Post**	**Pre**	**Post**
`Basic`	37.42	–	40.33	–	37.81	–	**44.18**	–
`NumericalCutOff`	41.77*	44.72*	**50.63***	53.85*	39.15	**44.3**	43.63	46.74
`Scale`	**43.19***	43.91*	47.01*	54.07	**40.98***	43.62	43.72	46.66
`Reflection`	41.32*	46.4*	50.08*	56.1*	38.97	41.03	44.16	46.88
`HistoryAware`	41.42*	47.4	49.92*	51.62	39.51	40.67	44.09	**47.71**
`ReflectionHistoryAware`	41.03*	**49.44**	50.55*	**64.29**	**40.98***	42.43*	42.99	43.49

Table 1: Summary of **last-turn accuracy** on Expert methods on all models. The **Pre** column indicates the accuracy on the entire dataset D, while the **Post** column indicates the accuracy on $D_{A,M,E,-1}$. Our fixed rule for computing $D_{A,M,E,-1}$ is to abstain if $\tau < 0.8$. Bolded entries are the highest accuracies in each column. Starred items pass McNemar's test with $p < 0.05$.

Expert	`Llama-3B`		`Llama-8B`		`Qwen-3B`		`Qwen-7B`	
	Pre	**Post**	**Pre**	**Post**	**Pre**	**Post**	**Pre**	**Post**
`ReflectionHistoryAware`	41.46*	60.87	**50.16***	62.16	**40.98***	45.53	42.99	**51.35**
`Order`	**44.58***	**66.67**	47.48*	**90**	40.02*	**52.94**	**43.32**	50.0

Table 2: Summary of **first-turn accuracy** on ReflectionHistoryAwareExpert and Order expert systems on all four models. The **Pre** column indicates the accuracy on the entire dataset D, while the **Post** column indicates the accuracy on $D_{A,M,E,0}$. The τ cutoff is fixed at 0.8 for ReflectionHistoryAwareExpert. For OrderExpert, the τ cutoff for each model is the cutoff that produces the best accuracy while still maintaining $>$ 3 samples. Namely, the cutoffs are 0.84 for `Llama-3B`, 0.85 for `Llama-8B`, 0.4 for `Qwen-3B`, and 0.82 for `Qwen-7B`. Starred items pass McNemar's test with $p < 0.05$.

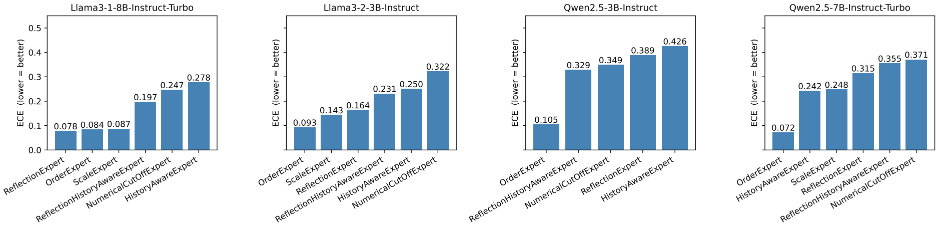


Figure 2: Bars show how well each expert is calibrated in terms of Expected Calibration Error (ECE).

Tables 1 and 2 contrast the performance of seven Experts across models on the dev split of iMedQA. Figure 2 shows the aggregate ECE across confidence-expressing Experts.

Excepting `Qwen-2.5-7B-Instruct-Turbo`, **all pre-abstention non-Basic methods outperform BasicExpert**, suggesting that some form of confidence elicitation is useful. The **best-performing prompt-based method on the Llama family on $D_{A,M,E,-1}$ is ReflectionHistoryAwareExpert** (with up to 64.29% accuracy), indicating that confidence elicitation

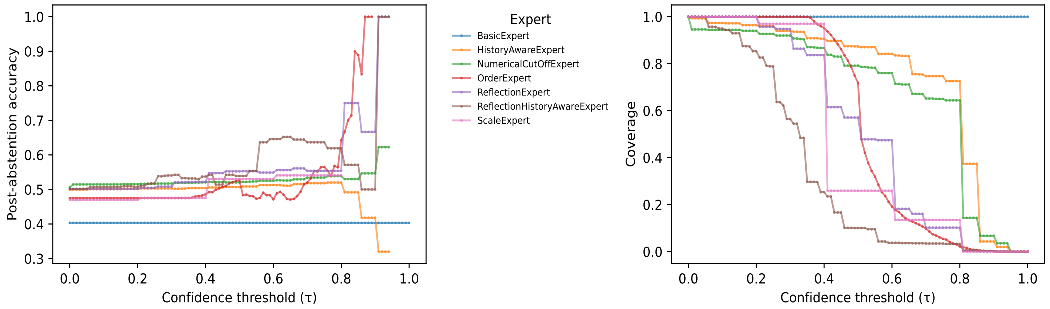


Figure 3: Impact of confidence threshold (τ) on expert performance for `Llama 3.1-8B-Instruct-Turbo`. Left: Post-abstention accuracy rises modestly for most experts as τ increases, except OrderExpert and ReflectionExpert, which show sharp gains once $\tau \simeq 0.75$. Right: Coverage drops monotonically with higher τ; ReflectionHistoryAwareExpert sacrifices coverage earliest.

Category	**Acc. (%) Within**	**Acc. Outside**	**Conf. Within**	**Conf. Outside**
Causal	25	64.52	0.8316	0.8255
Diagnosis	62.5	54.84	0.8218	0.8280
Next Steps	50	56.76	0.8177	0.8273
Which of the Following	66.67	22.22	0.8289	0.8197

Table 3: Confidence and accuracy by semantic and syntactic question categories for OrderExpert on `Llama-3.1-8B-Instruct-Turbo`, $\tau \geq 0.8$.

based on considering interaction history and reflecting on responses is more helpful than simply asking the model to assess its confidence based only on one pass of information from one turn. In the Qwen family, we see stronger performance from ScaleExpert, and ReflectionExpert, and HistoryAwareExpert than ReflectionHistoryAwareExpert, suggesting that for Qwen, all of these forms of confidence elicitation may be helpful in their own right but may not work well together. Within these three high-performing methods on the Qwen models, we see that **ScaleExpert tends to have the lowest ECE**, which may mean that its confidence estimate is most reflective of the final performance.

Examining first-turn post-abstention performance, **OrderExpert outperforms ReflectionHistoryAwareExpert by a large margin** once thresholding is applied on three of four models, achieving an accuracy of 90% on `Llama-3.1-8B-Instruct-Turbo` on $D_{A,M,E,0}$. However, `Qwen-2.5-7B-Instruct-Turbo` shows slightly higher post-abstention accuracy on ReflectionHistoryAware, though both methods still see performance improvements of 6+ percentage points on $D_{A,M,E,0}$ vs. D.

Examining confidence more closely through Figure 2, we see that **OrderExpert has the lowest ECE** on three of four models and the second lowest on the fourth model, suggesting that it is overall the best reflection of model performance compared to prompting-based methods. The ECE of the prompting-based methods seems to be highly family- and size-dependent, as results are mixed across the four models. Thus, another strength of OrderExpert is that it seems to not only have the overall lowest ECE but also have the **most robust ECE** with respect to different models.

5.2 Analysis of confidence threshold, accuracy and coverage

Figure 3 shows how post–abstention accuracy (left panel) and coverage (right panel) vary as we raise the confidence threshold τ for experiments with `Llama-3.1-8B-Instruct-Turbo`. Across Experts, we see a **safety–utility trade-off**. Higher thresholds filter out low–confidence answers, reducing coverage (number of questions answered), but usually also improving accuracy. BasicExpert, which never abstains, provides a useful lower bound with a flat curve that shows constant accuracy at 100% coverage. REFLECTIONHISTORYAWAREEXPERT abstains the earliest with coverage falling below 50% once $\tau > 0.3$, yet the corresponding accuracy plateau remains below 55%, indicating a limited benefit from aggressive refusal when the confidence signal is weak.

Two experts stand out. ORDEREXPERT's accuracy improves sharply at $\tau \simeq 0.75$, climbing from $\sim 50\%$ to $> 90\%$ while still answering 20% of questions. REFLECTIONEXPERT shows a similar but less pronounced jump, reaching 78% accuracy at comparable coverage. These steeper curves suggest the calibrated probability ratio used by ORDEREXPERT, and to a lesser extent the deliberative reasoning in REFLECTIONEXPERT, produce reliable uncertainty estimates in the high-confidence regime.

When taken together, the curves confirm that (i) confidence-based abstention can materially improve performance only when the expert's scores are themselves well calibrated and (ii) ORDEREXPERT provides the most favorable Pareto frontier on this model in terms of accuracy and coverage.

We include plots of confidence threshold, accuracy, and coverage over the remaining models—-`Llama-3.2-3B-Instruct` and the two Qwen models—in Appendix C, where we similarly see ORDEREXPERT showing better tradeoffs between accuracy and coverage.

5.3 Analysis of ORDEREXPERT

We further conduct an in-depth examination of ORDEREXPERT at its tertiary peak of $\tau > 0.79$, with the 39 most confidently answered questions. In this subset, we find through Table 3 that `Llama-3.1-8B-Instruct-Turbo` is relatively less confident in answering questions about adverse effects or treatment recommendation and less accurate in open-ended questions and questions about causes of diseases and in recommending next steps in patient care. Appendix D presents more detailed analyses on various syntactic and semantic axes.

6 Discussion and Future Directions

Our results highlight the advantages of instilling calibrated confidence-awareness in LLMs for medical QA. Naive methods of confidence estimation often lead to miscalibration. Novel prompt-based and internal confidence calibration and elicitation methods partially correct this. When we used these confidence scores to allow models to abstain below a confidence threshold, accuracy on answered items rose to SoTA levels on iMEDQA. This shows that today's LLMs encode useful internal uncertainty signals, but the challenge is to integrate them in a form that can be reliably used for downstream decision making.

However, our gains come at the cost of coverage. The abstention method improves precision by skipping uncertain cases, but the system answers fewer questions overall. In clinical workflows, this trade-off is often acceptable as it would entail erring on the side of caution as opposed to confidently recommending harmful interventions. The acceptability of such workflows will depend on the task (triage vs. patient education) and the availability of a human clinician to fill gaps. Future deployments will therefore need dynamic thresholding or cost–benefit optimization tuned to the specific clinical context (e.g., emergency medicine may favor higher coverage than tele-triage chatbots). We discuss limitations in-depth in Appendix E.

7 Conclusion

This paper benchmarks four LLMs on the task of medical QA in terms of confidence and objective multiple choice performance. It introduces four novel methods of eliciting model confidence, of which both ORDEREXPERT and REFLECTIONHISTORYAWAREEXPERT performance improvements. We further analyze performance and confidence in various semantic and syntactic categories, arguing that more work needs to be done to improve QA about next steps in patient care and model robustness to question phrasing.

Ethics Statement

Overall, we caution that medical QA is a relatively high-stakes setting, and *r* should be set quite high. However, at such a high threshold, there are not many data points for which models are so highly confident, which leads to a utility/inaccuracy tradeoff. Being overconfident about an erroneous diagnosis or drug prescription is harmful (at times deadly) to patients, so any medical QA system should be used with great caution, and clinicians should not over-rely on such technology. These harms are also most likely to fall disproportionately on people with rarer or under-studied conditions who seek medical care, which may consist of marginalize people who have been historically excluded from medical and broader scientific research.

Another ethical risk of these systems is that their performance seems to be very sensitive to prompt formatting, which may privilege certain writing or speaking styles associated with certain cultures or neurotypes; likely the cultures and neurotypes reflected in the pool of developers and data labellers involved in model creation.

Next, there is some discrepancy between how the task in our paper is carried out and how medical QA would work in real-world clinical settings. Specifically, there is unlikely to be a constrained set of four answer choices that a clinician can choose from for each query, but this is the set up of MediQ, presumably for ease of objective evaluation. Further work should test and evaluate free-text QA systems.

Finally, we report the computational cost of our experiments. The `Llama-3.2-3B-Instruct`, `Qwen-2.5-3B-Instruct`, and `Ministral-3B` experiments took 32 CPUs and no GPUs. Methods took various lengths of time, with a minimum of approximately 1 hour and a maximum of approximately two weeks for the 1,272 MediQ entries for a single method.

References

Evan Becker and Stefano Soatto. Cycles of thought: Measuring llm confidence through stable explanations, 2024. URL https://arxiv.org/abs/2406.03441.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. Biomedlm: A 2.7b parameter language model trained on biomedical text, 2024. URL https://arxiv.org/abs/2403.18421.

Brian H. Bornstein and A. Christine Emler. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *Journal of Evaluation in Clinical Practice*, 7(2):97–107, 2001. doi: https://doi.org/10.1046/j.1365-2753.2001.00284.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2753.2001.00284.x.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*, 2024.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexei Figueroa, Alexander Löser, Daniel Truhn, and Keno K. Bressem. Medalpaca – an open-source collection of medical conversational ai models and training data, 2025. URL https://arxiv.org/abs/2304.08247.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

Gregory Kell, Angus Roberts, Serge Umansky, Linglong Qian, Davide Ferrari, Frank Soboczenski, Byron C Wallace, Nikhil Patel, and Iain J Marshall. Question answering systems for health professionals at the point of care-a systematic review. *J Am Med Inform Assoc*, 31(4):1009–1024, April 2024.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=EKdk4vxKO4.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, et al. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*, 2025.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in llms: Reward calibration in rlhf. *arXiv preprint arXiv:2410.09724*, 2024.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11858–11875, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.693. URL https://aclanthology.org/2024.findings-emnlp.693/.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. TRAQ: Trustworthy retrieval augmented question answering via conformal prediction. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3799–3821, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.210. URL https://aclanthology.org/2024.naacl-long.210/.

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 28858–28888. Curran Associates, Inc., 2024c. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/32b80425554e081204e5988ab1c97e9a-Paper-Conference.pdf.

Yubo Li, Yidi Miao, Xueying Ding, Ramayya Krishnan, and Rema Padman. Firm or fickle? evaluating large language models consistency in sequential interactions. *arXiv preprint arXiv:2503.22353*, 2025.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge, 2023. URL https://arxiv.org/abs/2303.14070.

Daniel Lopez-Martinez. Trustworthiness in medical product question answering by large language models. 2024. URL https://www.amazon.science/publications/trustworthiness-in-medical-product-question-answering-by-large-language-models.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36: 46534–46594, 2023.

Izet Masic. Medical decision making - an overview. *Acta Inform. Med.*, 30(3):230–235, September 2022.

Xiangbin Meng, Xiangyu Yan, Kuo Zhang, Da Liu, Xiaojuan Cui, Yaodong Yang, Muhan Zhang, Chunxia Cao, Jingjia Wang, Xuliang Wang, et al. The application of large language models in medicine: A scoping review. *Iscience*, 27(5), 2024.

M Presiado, A Montero, L Lopes, and L Hamel. Kff health misinformation tracking poll: artificial intelligence and health information. *Published August*, 15, 2024.

Jeremy Qin, Bang Liu, and Quoc Dinh Nguyen. Enhancing healthcare llm trust with atypical presentations recalibration. *arXiv preprint arXiv:2409.03225*, 2024.

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, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620 (7972):172–180, Aug 2023a. ISSN 1476-4687. doi: 10.1038/s41586-023-06291-2. URL https://doi.org/10.1038/s41586-023-06291-2.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023b. URL https://arxiv.org/abs/2305.09617.

Matthew Spotnitz, Betina Idnay, Emily R Gordon, Rebecca Shyu, Gongbo Zhang, Cong Liu, James J Cimino, and Chunhua Weng. A survey of clinicians' views of the utility of large language models. *Applied Clinical Informatics*, 15(02):306–312, 2024.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Michael Trimble and Paul Hamilton. The thinking doctor: clinical decision making in contemporary medicine. *Clin. Med.*, 16(4):343–346, August 2016.

Zezhong Wang, Xingshan Zeng, Weiwen Liu, Yufei Wang, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Chain-of-probe: Examing the necessity and accuracy of cot step-by-step. *arXiv preprint arXiv:2406.16144*, 2024.

Jiaxin Wu, Yizhou Yu, and Hong-Yu Zhou. Uncertainty estimation of large language models in medical question answering, 2024. URL https://arxiv.org/abs/2407.08662.

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

Zhengtao Xu, Tianqi Song, and Yi-Chieh Lee. Confronting verbalized uncertainty: Understanding how llm's verbalized uncertainty influences users in ai-assisted decision-making. *International Journal of Human-Computer Studies*, pp. 103455, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

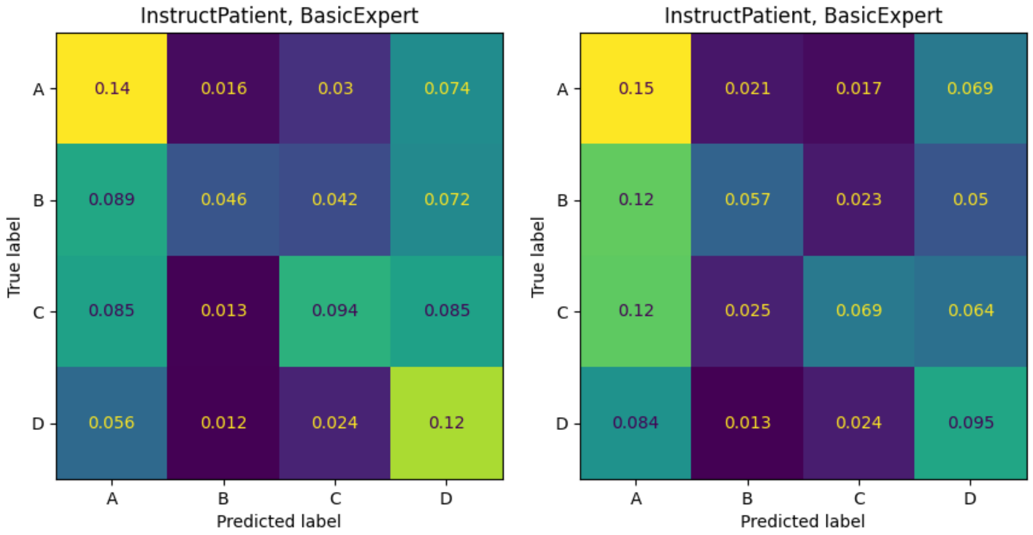
Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 37309–37323. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/f224f056694bcfe465c5d84579785761-Paper-Conference.pdf.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering, 2024. URL https://arxiv.org/abs/2305.10415.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3623–3643, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.198. URL https://aclanthology.org/2024.acl-long.198/.

A Additional Figures



(a)

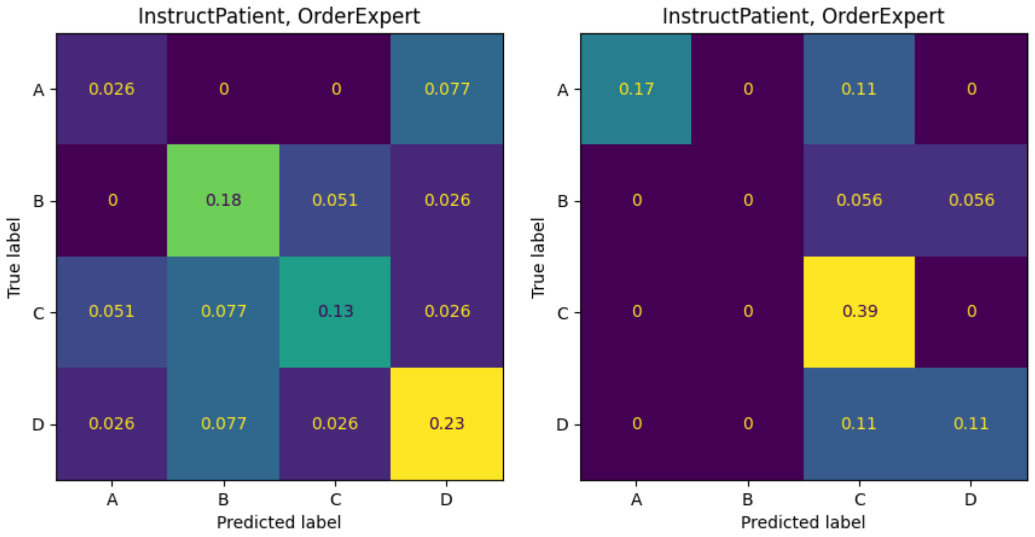
(b)

Figure 4: Confusion matrices for BasicExpert on `Llama-3.1-8B-Instruct-Turbo` (left) and `Llama-3.2-3B-Instruct` (right), normalized such that the grids sum to 1.

B Prompt Examples

Example prompt for ReflectionExpert:

```
"""|Based on the information provided so
far, which option do you think is the most
likely answer? Explain your  reasoning
step by step.
Now, challenge your initial diagnosis.
What alternative diagnoses might also fit
this patient's presentation? What evidence
might contradict your first choice? Consider
the evidence objectively as if you're a
```



(a)

(b)

Figure 5: Confusion matrix for ORDEREXPERT for `Llama-3.1-8B-Instruct-Turbo` (left, $\tau > 0.79$) and `Llama-3.2-3B-Instruct` (right, $\tau > 0.83$), normalized such that the grid sums to 1.

```
different doctor reviewing the case.
Based on your initial analysis and
consideration of alternatives, how confident
are you in your diagnosis? Provide a
numerical confidence score between 0.0 and
1.0, where:
- 0.0 means complete uncertainty
(random guess)
- 1.0 means absolute certainty"""
```

Example prompt for HISTORYAWAREEXPERT:

```
"""|Based on how your confidence has
changed throughout the conversation,
evaluate your current level of
certainty. Have you gained crucial
information? Has your confidence
plateaued? How do you explain any
significant changes in your confidence?
Provide a numerical confidence score
between 0.0 and 1.0, where:
- 0.0 means complete uncertainty
(random guess)
- 1.0 means absolute certainty"""
```

C Confidence threshold, accuracy and coverage for remaining models

Figures 6, 7, 8 show the plots that demonstrate the change in post-abstention accuracy with different confidence thresholds (left panels), and coverage with change in confidence thresholds (right panels).

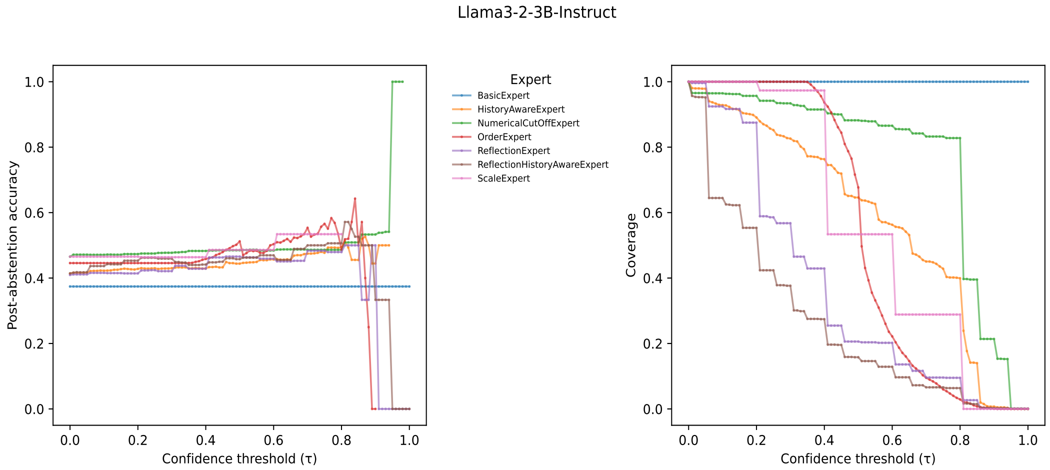


Figure 6: Impact of confidence threshold (τ) on expert performance for `Llama3-2-3B-Instruct`.

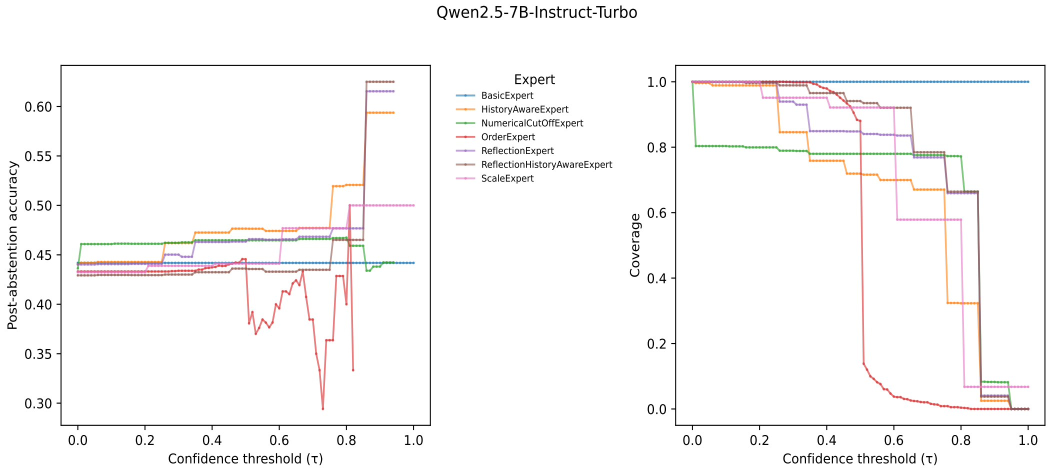


Figure 7: Impact of confidence threshold (τ) on expert performance for `Qwen2.5-7B-Instruct-Turbo`.

D Further Analysis of ORDEREXPERT

Figure 3 shows the accuracy and coverage of ORDEREXPERT when filtering for various exclusive lower bounds of τ on `Llama-3.1-8B-Instruct-Turbo`. While the peak accuracy of 100% is achieved at $\tau = 0.87$, there are very few items with such a high confidence (as exemplified in Figure 3, so we also note a secondary peak of 90% at $\tau > 0.84$ and a tertiary peak of 56.81% at $\tau > 0.79$.

We perform our category analysis at the tertiary peak of $\tau > 0.79$ (39 items), as this is the highest peak for which there is a large-enough set of samples. Semantically, there are no questions in this set that concern adverse effects or treatment, suggesting that the model is not quite as confident about these types of questions. As shown in Table 3, the model has the highest precision for questions about diagnoses, compared to other categories, and compared to all questions not related to diagnoses. In contrast, the model performs worse on the questions about recommending next steps and about ascertaining the cause of a patient's condition. These insights help us inform future work by informing us that more care is needed to better answer questions about the next steps.

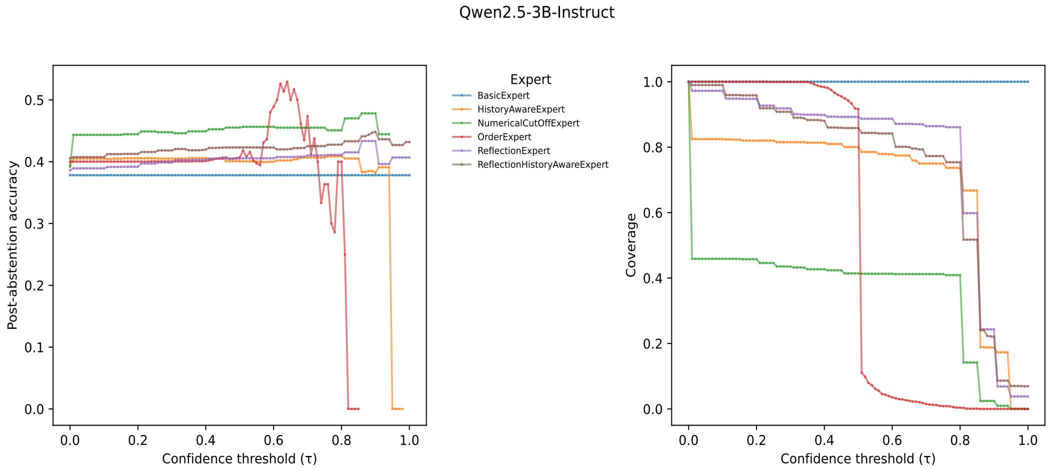


Figure 8: Impact of confidence threshold (τ) on expert performance for `Qwen2.5-3B-Instruct`.

Syntactically, Table 3 shows that the model performs better in answering questions formulated as "which of the following" than questions phrased otherwise (usually "what is" or "what are"). We conjecture that this may be because the phrase "which of the following" signals to the model that it should draw from its knowledge about the constrained set of four answer choices at hand, rather than retrieving more general information about the question as if it were open ended, allowing it to focus on the pertinent information that is most important for disambiguating the four answer choices. However, we caution that this result does not hold for all models—in fact, "which of the following" questions have lower accuracy for `Llama-3.2-3B-Instruct`. This may be because "which of the following" signals a cognitive shortcut by encouraging the model to look at the options provided and compare them without drawing on its general knowledge of the subject of the questions. A smaller model may have less answer-choice-specific knowledge, while a larger model may suffer more from too much irrelevant question-related knowledge.

In terms of other syntactic factors, we do not see significant correlations between input token count and context length, confidence, or accuracy. However, an interesting factor was the ordinal position (letter choice) of the correct answer. For BASICEXPERT, Fig. 4 shows many "A" predictions, even though the true distribution does not lean towards "A." This is improved in ORDEREXPERT (Fig. 5), suggesting that our method is effective in correcting for ordinal position bias. Appendix A shows additional result figures.

E Limitations

We acknowledge several limitations in our study. First, MediQ's multiple-choice format simplifies evaluation but does not reflect free-text diagnosis and management questions asked in practice. Second, we focus on English-language prompts and cross-lingual calibration may vary, especially given domain-specific terminology.

Methodologically, `OrderExpert` only considers a model's abstention decision at one interaction without accumulating confidence scores over an entire conversation history, so an augmentation to our method could be some method of incorporating confidence information throughout the entire interaction. Furthermore, having to compute a calibration output on a calibration prompt doubles the computational cost per query, making this method potentially financially costly for real-world implementers.

Additionally, accuracy may not be a fully representative performance metric for our task, especially given inconsistent sizes of $D_{A,M,E}$ across models and methods. However, we choose this metric because precision, recall, and F1 scores are quite arbitrary for such multiple-choice questions, and for some high τ thresholds, more robust metrics such as balanced accuracy and AUCROC are ill-defined due to high abstention rates and small sample coverage. Future work could build larger datasets such that more robust metrics such as AUCROC could be used for all parts of the analysis.

Future work can also further examine the second focus of the work in Li et al. (2024c)—asking helpful follow-up questions based on abstention decisions. Our work currently focuses solely on optimizing the abstention decision, but future work should extend ours by operationalizing this decision to better obtain the missing information causing an abstention.