Table 1: AUROC results for various uncertainty estimation methods across multiple datasets and model sizes. Methods include Lexical Similarity (LS), Semantic Entropy (SE), Predictive Entropy (PE), Length-normalized Entropy (LE), Step Verification (Step), Chain-of-Verification (CoVe) and Two-phase Verification (Two-phase). Results are shown for two model sizes: Llama 2 Chat (7b) and Llama 2 Chat (13b), evaluated on PubMedQA, MedQA, and MedMCQA datasets. The highest AUROC score for each model-dataset combination and the overall best results for averages and standard deviations (SDs) across datasets are highlighted in bold. For entropy-based methods, 5 answers are generated for each question, and the temperature is set to 0.5, which optimized Semantic Entropy (SE) and Length-normalized Entropy (LE) (Kuhn et al., 2023).

	LS[1]	SE[2]	PE[3]	LE[4]	Step	CoVe[5]	Two-phase (Ours)
Llama 2 Chat (7b)							
PubMedQA	0.5277	0.6320	0.6322	0.6028	0.6288	**0.6866**	0.6132
MedQA	0.4871	0.5154	0.5189	0.5224	0.4170	0.4861	**0.5553**
MedMCQA	0.3837	0.4676	0.5028	**0.6013**	0.5178	0.5509	0.5304
Average	0.4662	0.5383	0.5513	**0.5755**	0.5212	0.5745	0.5663
SD	0.0742	0.0846	0.0705	0.0460	0.1059	0.1023	**0.0425**
Llama 2 Chat (13b)							
PubMedQA	0.5551	0.5689	0.5681	0.4503	0.5085	0.5352	**0.5906**
MedQA	0.4860	0.4898	0.4010	0.5077	0.5934	0.5408	**0.6460**
MedMCQA	0.5142	0.5247	0.5708	0.5933	0.4895	**0.6026**	0.5793
Average	0.5184	0.5278	0.5133	0.5171	0.5305	0.5595	**0.6053**
SD	**0.0347**	0.0396	0.0973	0.0720	0.0553	0.0374	0.0357
Overall average	0.4923	0.5331	0.5323	0.5463	0.5258	0.5670	**0.5858**
Overall SD	0.0592	0.0593	0.0788	0.0628	0.0758	0.0694	**0.0411**

of a sequence by its length, handling the issue of disproportionate contribution to the total entropy due to variable sentence length.

Metrics Following Kuhn et al. (2023), we evaluate the performance of our uncertainty estimation approach using the area under the receiver operating characteristic curve (AUROC) as our metric. The metric measures the probability that a randomly chosen correct answer has a lower uncertainty level compared to a randomly chosen incorrect answer.

4.2 Results

We compare Two-phase Verification with several baseline methods on three medical datasets using two Llama 2 Chat models. The results are summarized in Table 1 and Figure 3.

Lexical Similarity (LS), which assesses uncertainty based on the overlap among sample responses, shows the lowest overall average AUROC. This suggests that lexical resemblances are insufficient indicators of certainty in the generated text, where semantic meaning is crucial. Semantic Entropy (SE) and Predictive Entropy (PE) demonstrate moderate improvements over LS. The two methods achieve similar AUROC scores, as they both estimate uncertainty from the entropy of sample responses. SE has slightly better overall results than PE, indicating that semantic clustering is an effective strategy in entropy-based methods. Length-normalized Entropy (LE) achieves the highest average AUROC for the Llama 2 Chat

[1]Lexical Similarity (Fomicheva et al., 2020)
[2]Semantic Entropy (Kuhn et al., 2023)
[3]Predictive Entropy (Kadavath et al., 2022)
[4]Length-normalized Entropy (Malinin & Gales, 2021)
[5]Chain-of-Verification (Dhuliawala et al., 2023)

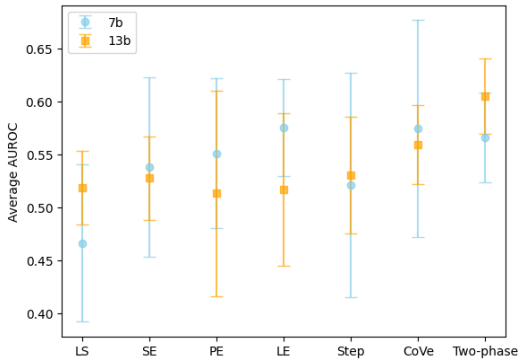

Figure 3: Performance comparison of UE methods on different model sizes

(7b) model, suggesting that normalizing entropy by answer length provides a more reliable uncertainty signal for smaller models. However, LE's performance does not consistently hold across larger model sizes or all datasets, indicating that while length normalization is beneficial, it is not a comprehensive solution for UE.

Step Verification, as the most straightforward verification method that relies solely on the model's self-validation ability, does not show a performance improvement compared to other baseline methods. Interestingly, for the 7b model, the performance of Step Verification appears to be correlated with the accuracy of the model's answers on each dataset. PubMedQA, which has the highest answer accuracy (0.65), also shows the best Step Verification result, while MedQA and MedMCQA, with lower accuracies (0.2991 and 0.3429, respectively), have poorer Step Verification performance. This observation suggests that smaller models may exhibit overconfidence in their generated answers and struggle to identify their own mistakes during self-verification. This limitation highlights the necessity of introducing a verification chain to help the model recognize hallucinations in its outputs.

CoVe slightly outperforms Two-phase Verification in some cases but has a high standard deviation, particularly with smaller model sizes. This may stem from the variability in the quality of independently generated answers, as smaller models might produce less faithful responses. Further research could explore ways to improve the reliability of answering verification questions, potentially by integrating external knowledge bases.

In general, Two-phase Verification demonstrates the best overall performance, achieving the highest AUROC in half of the model-dataset combinations and the highest average AUROC across all experiments. It also exhibits stable performance with the lowest overall standard deviation, unlike other methods such as CoVe, which show performance fluctuations in certain scenarios. Moreover, the scalability of Two-phase Verification is particularly noteworthy when comparing the AUROC results for Llama 2 Chat (7b) and Llama 2 Chat (13b). While most methods show only modest improvements or, in some cases, a decrease in performance with the larger model size, Two-phase Verification not only improves but also does so at a higher rate than its counterparts. These characteristics suggest the method's potential to provide reliable uncertainty estimation across various datasets and to scale with larger model sizes.

5 Discussion

5.1 Uncertainty Estimation in medical QA

Uncertainty Estimation is of paramount importance in the medical domain, where untruthful information can lead to severe consequences. In medical applications utilizing LLMs, such as AI medical chatbots, it is crucial to assess the trustworthiness of model outputs to ensure patient safety. In the cases where the model is less certain in its predictions, the user

should be alerted and advised to seek further verification or expert opinion before following the model's suggestions.

The findings of our empirical study contribute to the literature on UE of LLMs, particularly in the context of medical question answering, which has been less studied. Previous research has primarily focused on UE from a statistical perspective, hypothesizing that the model intrinsically knows when it is uncertain about an answer, leading to higher variability in its outputs. However, professional medical knowledge is often underrepresented in the training data, which can lead to the model generating responses confidently even when it is hallucinating. As a result, answers may exhibit low entropy, falsely suggesting high certainty.

Our Two-phase Verification method mitigates these issues by independently verifying responses, thus providing an effective measure of a model's certainty without needing token-level probabilities. This is especially useful for black-box models, where architectural details are inaccessible. Moreover, the scalability of the Two-phase Verification method is a critical aspect for future applications. As large-scale models continue to evolve, the ability to maintain and even enhance performance with increased model size is essential.

The concept of *Chain-of-Verification* (CoVe) has been previously proposed to reduce hallucinations and then self-correct them to generate more factual statements (Dhuliawala et al., 2023). However, to the best of our knowledge, this concept has not been explored for uncertainty estimation. Our work demonstrates the effectiveness of integrating a verification chain for uncertainty estimation in medical question-answering, opening up new possibilities for future research in this direction.

5.2 Limitations and future work

Verification question generation A critical stage of Two-phase Verification is to generate verification questions that effectively challenge the initial explanation. As explanation paragraphs are generated with linguistic coherence, sentences often use pronouns or references that rely on previous sentences. When verification questions are derived from discrete sentences, the model may miss essential context. Thus, the verification questions might not always incisively interrogate the key information presented. Although few-shot prompts aid question formulation, they can inadvertently inhibit the LLM's creativity, confining it to the patterns seen in these examples. In future work, it will be essential to enhance the generation of verification questions to be more context-aware and adaptable.

Domain knowledge constraints Another constraint for Two-phase Verification is the knowledge capacity of the language model, which directly affects the quality of the answers to verification questions. Llama 2 Chat, as a general-purpose language model, possesses only a broad understanding of medical knowledge, lacking the depth required for specialized areas. To improve the model's responses to verification questions, we integrate dense retrieval techniques to source relevant information from external databases like Wikipedia. However, this method falls short as the retrieved results frequently have low relevance scores to the verification queries and fail to provide the necessary knowledge. Future improvements should focus on retrieving relevant information from professional medical datasets, such as research papers, medical textbooks, and expert-curated knowledge bases. By leveraging these domain-specific resources, the model can generate more accurate and reliable independent answers to verification questions, enabling more effective detection of hallucinations and uncertainties in medical explanations.

6 Conclusion

In this paper, we conduct an empirical study on the Uncertainty Estimation of LLMs in medical question-answering tasks. We find Uncertainty Estimation challenging in the medical domain, with existing methods performing poorly, especially with smaller model sizes. To address this challenge, we propose Two-phase Verification, a novel approach that integrates the concept of CoVe to assess the reliability of language model outputs. We show that the model is capable of detecting its own hallucinations by answering verification