Uncertainty Estimation of Large Language Models in Medical Question Answering

Jiaxin Wu
The University of Hong Kong
lisa24@connect.hku.hk

Yizhou Yu
The University of Hong Kong
yizhouy@acm.org

Hong-Yu Zhou
The University of Hong Kong & Harvard Medical School
whuzhouhongyu@gmail.com

Abstract

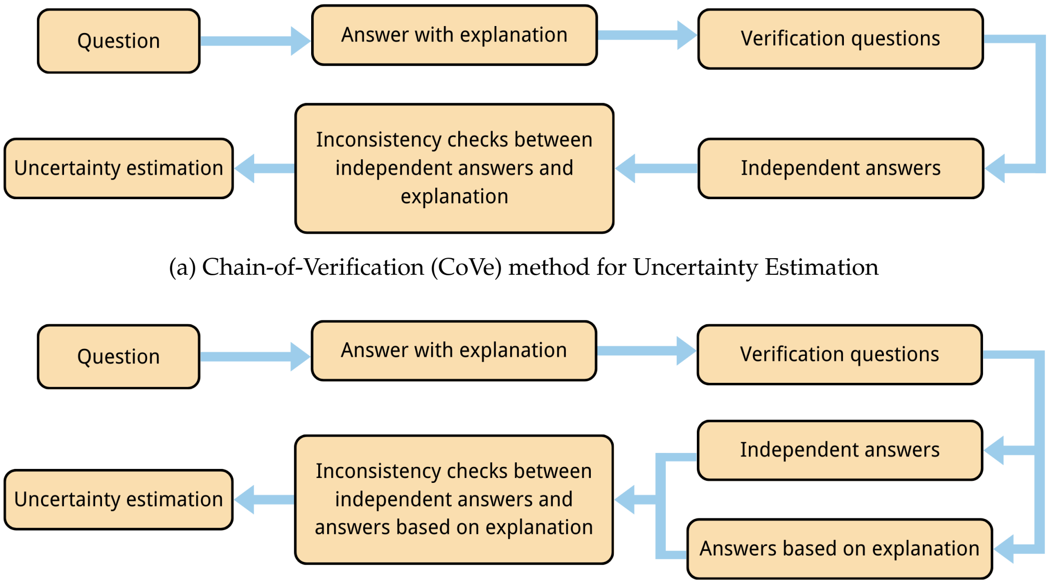
Large Language Models (LLMs) show promise for natural language generation in healthcare, but risk hallucinating factually incorrect information. Deploying LLMs for medical question answering necessitates reliable uncertainty estimation (UE) methods to detect hallucinations. In this work, we benchmark popular UE methods with different model sizes on medical question-answering datasets. Our results show that current approaches generally perform poorly in this domain, highlighting the challenge of UE for medical applications. We also observe that larger models tend to yield better results, suggesting a correlation between model size and the reliability of UE. To address these challenges, we propose Two-phase Verification, a probability-free Uncertainty Estimation approach. First, an LLM generates a step-by-step explanation alongside its initial answer, followed by formulating verification questions to check the factual claims in the explanation. The model then answers these questions twice: first independently, and then referencing the explanation. Inconsistencies between the two sets of answers measure the uncertainty in the original response. We evaluate our approach on three biomedical question-answering datasets using Llama 2 Chat models and compare it against the benchmarked baseline methods. The results show that our Two-phase Verification method achieves the best overall accuracy and stability across various datasets and model sizes, and its performance scales as the model size increases.

1 Introduction

Large Language Models (LLMs), such as GPT-4 and Llama 2, have demonstrated considerable potential in generating human-like text across a broad spectrum of fields, without additional domain-specific training. Their capabilities can be harnessed to provide assistance in the healthcare sector for a wide range of applications, including but not limited to disease diagnosis, clinical decision-making, and patient communication (Cascella et al., 2023). Despite the potential, the deployment of LLMs faces challenges. A prevalent concern is the tendency of LLMs to 'hallucinate', a term used to describe circumstances where the model generates plausible yet incorrect information, particularly when they are not able to provide an accurate response (Ji et al., 2023). In high-risk scenarios such as healthcare, where decisions can have direct impact on human lives, ensuring the reliability of LLMs becomes critical. This underscores the need for effective approaches to accurately estimate the uncertainty of generated responses and detect instances of hallucination.

In medical settings, existing methods for quantifying uncertainty, including entropy-based methods (Kadavath et al., 2022; Kuhn et al., 2023) and fact-checking (Guo et al., 2022; Shuster et al., 2021), have demonstrated certain limitations. Entropy-based methods operate on the assumption that a model, when confident in its answer, generates a distribution of responses with a small entropy. On the contrary, if the model is unsure, it might hallucinate and produce a diverse range of responses, thus increasing the entropy (Kadavath et al., 2022). However, within the complexity of the medical domain, the model can often fabricate untruthful information with a high level of confidence. This results in a misleadingly low entropy, which fails to accurately represent the uncertainty embedded in the generated response. Fact-checking, another common approach for uncertainty estimation, validates the generated responses by comparing them with relevant truth retrieved from an external knowledge database. However, this method encounters limitations due to the scarcity of comprehensive and professional medical knowledge bases.

In this report, we benchmark several popular methods using different model sizes and datasets to establish a comparative understanding of their performance. These benchmarks reveal the challenges of uncertainty estimation in medical question-answering. We also propose Two-phase Verification, a probability-free approach based on the *Chain-of-Verification* (CoVe) concept (Dhuliawala et al., 2023). This approach operates independently of token-level probabilities and thus can be applied to black-box models. First, the model generates an explanation alongside its initial answer. Next, it formulates verification questions targeting the explanation, to which it provides independent answers. Two-phase Verification refines CoVe's inconsistency check process by prompting the model to answer the verification questions again, using the statement in question as a reference. The inconsistencies between the two sets of answers serve as a measure of uncertainty in the answer. The workflows of CoVe and Two-phase Verification are visualized in Figures 1a and 1b, respectively.



(a) Chain-of-Verification (CoVe) method for Uncertainty Estimation

(b) Two-phase Verification method for Uncertainty Estimation

Figure 1: Comparison of CoVe and Two-phase Verification Methods

2 Related Work

2.1 Entropy-based methods

Large Language Models generate output on a token-by-token basis based on the sequence so far. Token probabilities are a direct measure of a model's confidence in its next-token prediction. Xiao & Wang (2021) explore the link between hallucination in conditional language generation tasks and predictive uncertainty, which is quantified using entropy measures of the token probability distributions. They find that higher levels of predictive uncertainty, especially epistemic uncertainty originating from the model's knowledge gaps, correlate with an increased propensity for hallucinatory outputs. Duan et al. (2023) addresses the challenge of token-level generative inequality by a method termed Shifting Attention to Relevance (SAR), which reassigns attention to more semantically relevant components when estimating uncertainty. Malinin & Gales (2021) introduce information-theoretic uncertainty measures at both the token and sequence levels and propose a novel metric, reverse mutual information, for structured uncertainty assessment, utilizing ensemble methods like Monte-Carlo Dropout and Deep Ensembles. Kuhn et al. (2023) propose semantic entropy, which measures uncertainty over meanings rather than just sequences of words. Their unsupervised method clusters semantically equivalent generations and calculates the predictive entropy of the resulting probability distribution over these clusters. A concurrent work of Wang et al. (2024) calibrates entropy-based uncertainty at word and sequence levels according to their semantic relevance, aiming to address the generative inequality challenge.

2.2 Self-assessment methods

LLMs possess the inherent potential to reflect on their outputs; however, the self-evaluation may not be robust as LLMs are inclined to find their own content credible (Kadavath et al., 2022). To enhance the calibration and confidence estimation of LLMs, researchers have developed techniques including fine-tuning and prompting. Lin et al. (2022) finetune GPT-3 by supervised learning to express its uncertainty in natural language. Their experiment demonstrates that GPT-3 can be trained to provide answers along with a corresponding confidence level. Similarly, Kadavath et al. (2022) investigate the self-awareness of LLMs by training them to estimate the likelihood that their generated responses are correct. Their research reveals that the effectiveness of self-evaluation improves with model size and few-shot prompting. Additionally, the process benefits from presenting the model with various answer samples before asking it to assess the validity of a single proposed response. Kojima et al. (2023) explore the zero-shot capabilities of LLMs, finding that chain-of-thought prompting boosts the reasoning abilities of LLMs, especially in arithmetic tasks. By instructing LLMs to generate intermediate steps explicitly before answering the questions, a simple prompt template provides performance gain. Manakul et al. (2023) introduce a zero-resource methodology for LLMs to self-check hallucinated generations based on the hypothesis that hallucinations tend to diverge. It operates by generating multiple responses to a prompt and then assessing the factual consistency between these responses.

2.3 External tools

Since knowledge gaps are a common cause of hallucinations, external knowledge retrieval is often utilized to mitigate hallucinations and produce more faithful generations (He et al., 2022; Shuster et al., 2021). For example, Chern et al. (2023) collect external evidence to validate the factuality of claims extracted from the LLM output. While prompting strategies enhance LLM performance in certain tasks, plausible explanations are often provided even when the final answer is wrong (Kojima et al., 2023). To overcome this limitation, Chen et al. (2023) propose a multi-turn conversation framework to integrate prompting and external tools including calculators and search engines, which reduces the mistakes made by LLMs and enhances the accuracy in complex reasoning tasks.

3 Methodology

In this section, we elaborate on our approach to estimating generation uncertainty, which leverages the idea from the Chain-of-Verification (CoVe) framework. Our approach is inspired by the foundational work of Dhuliawala et al. (2023) and extends it by integrating a measure for confidence level based on discovered inconsistencies. The primary goal is to identify the occurrence of possible hallucinations by incorporating a robust, unsupervised verification mechanism that operates independently of the model's initial outputs.