ConU: Conformal Uncertainty in Large Language Models with Correctness Coverage Guarantees

Zhiyuan Wang[1], Jinhao Duan[2], Lu Cheng[3], Yue Zhang[2], Qingni Wang[1], Xiaoshuang Shi[1]*, Kaidi Xu[2], Hengtao Shen[1], Xiaofeng Zhu[1]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China
[2]Department of Computer Science, Drexel University
[3]Department of Computer Science, University of Illinois Chicago

Abstract

Uncertainty quantification (UQ) in natural language generation (NLG) tasks remains an open challenge, exacerbated by the closed-source nature of the latest large language models (LLMs). This study investigates applying conformal prediction (CP), which can transform any heuristic uncertainty notion into rigorous prediction sets, to black-box LLMs in open-ended NLG tasks. We introduce a novel uncertainty measure based on self-consistency theory, and then develop a conformal uncertainty criterion by integrating the uncertainty condition aligned with correctness into the CP algorithm. Empirical evaluations indicate that our uncertainty measure outperforms prior state-of-the-art methods. Furthermore, we achieve strict control over the correctness coverage rate utilizing 7 popular LLMs on 4 free-form NLG datasets, spanning general-purpose and medical scenarios. Additionally, the calibrated prediction sets with small size further highlights the efficiency of our method in providing trustworthy guarantees for practical open-ended NLG applications.

1 Introduction

Despite advancements in various natural language generation (NLG) tasks (Katz et al., 2024; Touvron et al., 2023a; Chen et al., 2023; Duan et al., 2024b,c), large language models (LLMs) are proven to hallucinate facts and confidently generate textual information that is not correct or grounded in reality (Ji et al., 2023; Manakul et al., 2023). Factually incorrect answers can confuse and mislead users, resulting in erroneous conclusions and ultimately undermining the trustworthiness of LLMs-based high-stakes applications.

Uncertainty quantification (UQ) provides valuable insights into the reliability of model responses, facilitating risk assessment and hallucination detection (Kadavath et al., 2022; Lin et al., 2022a). However, it demands investigating black-box uncertainty measures with the proliferation of LLMs served via APIs (Achiam et al., 2023), which only allows textual inputs and outputs. Conformal prediction (CP) (Campos et al., 2024; Angelopoulos and Bates, 2021; Quach et al., 2024; Zhao et al., 2024) is known for providing a model-agnostic and statistically rigorous uncertainty estimation. CP was primarily employed in classification (Angelopoulos and Bates, 2021) and regression tasks (Wang et al., 2024a). For NLG tasks, CP is first adapted to the multiple-choice question-answering (MCQA) setting, where the acceptable response is selected from a fixed set of options (Kumar et al., 2023; Ye et al., 2024), limiting its applications in real-world open-ended NLG tasks. Conformal language modeling (Quach et al., 2024) relies on the model likelihoods and calibrates a stopping rule to sample prediction sets from the infinite output space until users are confident that the set covers at least one response satisfied. LofreeCP (Su et al., 2024) studies CP for API-only LLMs without logit access by leveraging uncertainty information from diverse sources.

Our study explores adapting CP for general NLG applications. The nonconformity score (NS) in CP serves as a criterion for calibrating prediction sets, which provide coverage guarantees by selecting a set of possible labels that satisfy the NS threshold (Angelopoulos and Bates, 2021). Since typical logits-based NS may encounter miscalibration, we aim to integrate black-box UQ into the definition of NS, by closely aligning it with the uncertainty condition of the correct answers and devising a conformal uncertainty criterion, while it is more reliable to analyze the uncertainty within LLMs' true output space. Then, we employ the uncertainty criterion, concluded from a small amount of independent and identically distributed (i.i.d.) calibration data, to construct prediction sets by selecting generations sharing a similar uncertainty condition

*Corresponding to: Xiaoshuang Shi <xsshi2013@gmail.com>

from the unbounded output space on test samples. Typically, there are two goals of CP: (1) the calibrated prediction set contains the correct answer with at least a user-specified probability; and (2) the average set size should be small, demonstrating the prediction efficiency of our method.

The first challenge is UQ for black-box LLMs. Our solution is inspired by an intuitive observation: If a language model generates more semantically diverse outputs for the same prompt, the uncertainty is likely higher (Su et al., 2024; Lin et al., 2023; Xiong et al., 2023). Regardless of the model's capability to tackle the current problem, the confidence score that the model assigns to a generation can be represented by its frequency within the output space. We approximate the model's output distribution by sampling multiple answers to the same question. Then, we perform semantic clustering on the sampled generations, and propose to measure the uncertainty of each generation by combining two factors: the frequency of occurrence of the semantic meaning it conveys, and the consistency between its semantic and other semantic clusters augmented by their individual frequency.

Based on the measure, we define the NS as the uncertainty of the generation. To this end, the generation meets the correctness criterion and is semantically most similar to the reference answer in the calibration set. We then calculate the quantile $\hat{q}$ of NSs for all calibration samples, based on the user-specified upper bound of error rate α. Next, we utilize the conformal uncertainty criterion (i.e., the uncertainty threshold $\hat{q}$) to construct a prediction set for each test sample by selecting generations that satisfy the uncertainty conditions strictly associated with correctness from the candidate generations. Additionally, for black-box UQ, we propose employing the most frequent generation or semantic (i.e., the model's most confident answer) as a more trustworthy reference object for the query and leveraging it to measure the overall uncertainty of the current UQ process. We term this measure *ConU*, as it employs the same approach as the conformal uncertainty criterion.

Extensive experimental results exhibit that *ConU* generally outperforms prior state-of-the-art methods and verify the strict correctness coverage guarantees. Specifically, the prediction sets calibrated by the conformal uncertainty criterion always encompass the correct answers under various user-specified error rates. Furthermore, the average prediction set size is small, highlighting the prediction efficiency of our approach. To our knowledge, this is the first method in the literature to strictly link the NS with the uncertainty condition aligned with correctness via black-box UQ, thereby developing a more robust conformal uncertainty criterion, which provides rigorous correctness coverage guarantees in practical open-ended NLG tasks, and its unique inspiration in benchmarking UQ in LLMs through CP generates independent interest*.

In summary, our major contributions are listed as follows:

- We propose a sampling-based black-box uncertainty measure, termed as *ConU*, utilizing self-consistency in open-ended NLG tasks, facilitating trustworthy decision-making.
- We devise a conformal uncertainty criterion by strictly aligning the NS with the uncertainty condition of acceptable answers, and achieve rigorous correctness coverage with at least a user-specified probability, thereby providing robust guarantees under various error rates in practical open-ended NLG applications.
- We conduct selective prediction leveraging the calibrated prediction sets and obtain promising improvements in model accuracy without requiring additional task-specific fine-tuning or architectural modifications.

2 Related Work

2.1 Uncertainty Quantification in LLMs

Prior work on UQ in LLMs predominantly focuses on white-box information like token-likelihoods or embeddings (Da et al., 2024; Kuhn et al., 2023; Duan et al., 2024a; Wang et al., 2024b), internal state or activations (Yin et al., 2024; Chen et al., 2024), model fine-tuning (Tian et al., 2023). These methods can encounter poor calibration and require substantial computational resources. Additionally, researchers lack white-box access to the internal information of LLMs served via APIs. These restrictions demand black-box measures for general UQ in LLMs generations.

Recent work (Lin et al., 2023) develops several sampling-based uncertainty measures, which can be applied to black-box LLMs by leveraging semantic similarity along with dispersion. Our study follows the sampling setting and proposes to employ the most frequent generation as the reference

*Our code is available at https://github.com/Zhiyuan-GG/Conformal-Uncertainty-Criterion/tree/main

object to measure the overall uncertainty based on the self-consistency theory (Wang et al., 2022).

2.2 Conformal Prediction in LLMs

CP (Angelopoulos and Bates, 2021; Quach et al., 2024; Campos et al., 2024) has emerged as a theoretically sound and practically useful way to guarantee ground-truth coverage with the aid of a small amount of exchangeable samples for calibration. CP in classification tasks defines the NS, which is correlated with the ground-truth label, obtains the quantile, $\hat{q}$, of NSs for all calibration samples based on a user-specified upper bound of the error rate α, and utilizes $\hat{q}$ as a threshold to select possible labels on test samples, thereby establishing prediction sets that guarantee ground truth coverage with at least the probability of $1-\alpha$.

Recently, researchers have attempted to apply CP to LLMs for principled UQ. The work (Mohri and Hashimoto, 2024) achieves conformal factuality guarantees by progressively making generations less specific and establishing their corresponding entailment sets until correct answers are encompassed. For correctness coverage, two studies (Kumar et al., 2023; Ye et al., 2024) follow CP in classification tasks and convert NLG tasks into MCQA settings. For open-ended NLG, based on the output token sequence logits, the study (Quach et al., 2024) develops a stopping rule to sample generations until users are confident that a correct answer is covered in QA tasks, which can be impractical for API-only LLMs. LofreeCP (Su et al., 2024) leverages uncertainty information to construct prediction sets that achieve correctness coverage.

This paper focuses on more practical scenarios of black-box LLMs in open-ended NLG tasks. Differing from LofreeCP, we strictly connect the NS with the uncertainty condition aligned with correctness via black-box UQ, which concludes a more robust conformal uncertainty criterion to calibrate prediction sets with rigorous correctness coverage guarantees under various error rates despite the complexity of the model or datasets.

3 Method

Our method investigates two key issues: (1) how to estimate the uncertainty in black-box LLMs when we can only access the output texts; and (2) how to provide rigorous guarantees on the error rate in open-ended NLG tasks. We first devise a black-box uncertainty measure grounded in self-consistency to provide the trustworthiness notion of model responses. Furthermore, we utilize the split CP technique to convert the heuristic approximation into a statistically rigorous one, thereby ensuring a more robust and systematic assessment of uncertainty.

3.1 Preliminaries

Following the analysis of black-box LLMs in prior work (Xiong et al., 2023; Lin et al., 2023; Manakul et al., 2023), conditioned on each prompt (or question) x_i, we employ the most likely generation $\hat{y}_i$ for correctness evaluation. Additionally, we sample a set of M candidate generations $\left\{\hat{y}_m^{(i)}\right\}_{m=1}^{M}$ from the model's output space for black-box UQ and the derivation of conformal uncertainty criterion. We denote the reference answer to x_i as y_i^*.

3.2 Uncertainty Quantification

For each data point, we first cluster semantics in the M sampled generations and obtain K non-repeated semantics. We denote the number of generations sharing the k-th semantic as V_k (i.e., $\sum_{k=1}^{K} V_k = M$) and any one generation in this cluster as $\hat{y}_k^{(i)}$.

Building on earlier approaches that utilize self-consistency (Wang et al., 2022; Su et al., 2024; Yadkori et al., 2024) as a reliable measure of confidence, we employ the frequency of the k-th semantic as its proxy for reliability: $\mathcal{F}\left(\hat{y}_k^{(i)}\right) = \frac{V_k}{M}$. Then, we define the uncertainty score of each candidate generation in $\left\{\hat{y}_m^{(i)}\right\}_{m=1}^{M}$ as

$$\begin{aligned}\mathcal{U}\left(\hat{y}_m^{(i)}\right) =& 1 - \lambda \cdot \mathcal{F}\left(\hat{y}_m^{(i)}\right) - (1-\lambda) \cdot \\ & \frac{1}{K}\sum_{k=1}^{K} \mathcal{S}\left(\hat{y}_m^{(i)}, \hat{y}_k^{(i)}\right) \mathcal{F}\left(\hat{y}_k^{(i)}\right),\end{aligned} \tag{1}$$

where $\mathcal{F}\left(\hat{y}_m^{(i)}\right)$ refers to the frequency of the semantic that $\hat{y}_m^{(i)}$ conveys, and $\mathcal{S}(\cdot,\cdot)$ measures the semantic similarity between two generations utilizing a *cross-encoder* model (Reimers and Gurevych, 2019). $\mathcal{F}\left(\hat{y}_k^{(i)}\right)$ is to augment the persuasiveness of the similarity score associated with $\hat{y}_k^{(i)}$.

To measure the model uncertainty, we select any one generation in the largest semantic cluster to be the most trustworthy generation in the M sampled generations and denote it as $\hat{y}_{mst}^{i}$. Then, we define the uncertainty score of the i-th query-response