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

Anonymous ACL submission

Abstract

Uncertainty quantification (UQ) in natural language generation (NLG) tasks remains an open challenge, exacerbated by the intricate nature of the recent large language models (LLMs). This study investigates adapting conformal prediction (CP), which can convert any heuristic measure of uncertainty into rigorous theoretical guarantees by constructing prediction sets, for black-box LLMs in open-ended NLG tasks. We propose a sampling-based uncertainty measure leveraging self-consistency, and develop a *conformal uncertainty* criterion by integrating the uncertainty condition aligned with correctness into the design of the CP algorithm. Experimental results indicate that our uncertainty measure generally surpasses prior state-of-the-art methods. Furthermore, we calibrate the prediction sets within the model's unfixed answer distribution and achieve strict control over the correctness coverage rate across 6 LLMs on 4 free-form NLG datasets, spanning general-purpose and medical domains, while the small average set 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 like question answering (QA) (Katz et al., 2024; Touvron et al., 2023a; Chen et al., 2023), 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., 2023) 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 correct 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., 2023) 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 trustworthy to analyze the uncertainty within LLMs' true output space. Then, we can leverage 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 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 NS as the uncertainty score 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 query-answering 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 free-form NLG tasks, facilitating trustworthy decision-making.
- We devise a conformal uncertainty criterion by strictly aligning the NS with the uncertainty condition of correct 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 (Kuhn et al., 2023; Duan et al., 2024; Wang et al., 2024b), internal state or activations (Yin et al., 2024; Chen et al., 2024), model fine-tuning (Lin et al., 2022a; 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 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., 2023; 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 independent and identically distributed (i.i.d.) samples for calibration. CP in classification defines NS that is correlated with the ground-truth label, obtains the quantile $\hat{q}$ of NSs for all calibration samples based on a user-specified error rate α, and utilizes $\hat{q}$ as a threshold to select possible labels on test samples, thereby establishing prediction sets that achieve 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., 2023) devises a stopping rule to sample generations until users are confident that a correct answer is covered in QA tasks. 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 of correct answers 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 quantify the uncertainty in black-box LLMs when we can only access the generated 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 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 prior utility of black-box LLMs (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 sample, 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

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

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 overall uncertainty, we randomly select one generation in the largest semantic cluster to be the most trustworthy generation in the M sampled generations and denote it as $\hat{y}_{mst}^{i}$:

$$\hat{y}_{mst}^{i} = \operatorname{argmax}_{\hat{y}_k^{(i)} \in \left\{\hat{y}_m^{(i)}\right\}_{m=1}^{M}} V_k. \quad (2)$$

Then, we define the uncertainty score of the i-th

query-response as

$$\mathcal{U}\left(\left\{\hat{y}_m^{(i)}\right\}_{m=1}^{M}|x_i\right) = 1-\lambda\cdot\mathcal{F}\left(\hat{y}_{mst}^{i}\right) - (1-\lambda)\cdot\frac{1}{K}\sum_{k=1}^{K}\mathcal{S}\left(\hat{y}_{mst}^{i},\hat{y}_k^{(i)}\right)\mathcal{F}\left(\hat{y}_k^{(i)}\right). \tag{3}$$

Intuitively, the most frequent semantic within the candidate generations represents the model's most confident answer to the current problem. Even though the reference semantic may not necessarily be the correct one, we can measure the degree of the model's uncertainty by calculating the confidence level of that semantic as well as the deviation between it and other semantics.

Since Eq. (1) can quantify the uncertainty of each candidate generation, we attempt to develop an uncertainty criterion to search for the correct answers within the unfixed output space of the LLM.

3.3 Conformal Correctness Coverage

Following the fundamental requirement in CP (Angelopoulos and Bates, 2021), we randomly employ N samples to construct the calibration data set $\{(x_i, y_i^*)\}_{i=1}^{N}$, and for each calibration sample we demand that at least one sampled generation $\hat{y}_j^{(i)}$ in $\left\{\hat{y}_m^{(i)}\right\}_{m=1}^{M}$ meets the correctness criterion. Our objective of **conformal correctness coverage** is by concluding the uncertainty criterion that is closely linked with correctness on $\{(x_i, y_i^*)\}_{i=1}^{N}$, we can calibrate an uncertainty (prediction) set $\mathcal{P}(x_{test})$ for the test prompt x_{test} by selecting generations that meet the common uncertainty condition, and the set can guarantee correctness coverage under various user-specificed error rates. Here, we approximate the prediction region of x_{test} to the M candidate generations $\left\{\hat{y}_m^{(test)}\right\}_{m=1}^{M}$.

Assumptions: (1) There is at least one candidate generation in $\left\{\hat{y}_m^{(test)}\right\}_{m=1}^{M}$ meeting the correctness criterion; (2) Samples in the calibration and test data sets are exchangeable.

As the sampled set $\left\{\hat{y}_m^{(test)}\right\}_{m=1}^{M}$ is a subset of the prediction region, which is impossible to enumerate, we can simplify it by stating that there is at least one correct answer in $\left\{\hat{y}_m^{(test)}\right\}_{m=1}^{M}$. Exchangeability is the fundamental assumption of CP techniques (Angelopoulos and Bates, 2021).

Based on the uncertainty measure described as Eq. (1), we define the NS of the i-th calibration sample as

$$r_i = r\left(x_i, y_i^*\right) = \mathcal{U}\left(\underset{\hat{y}_j^{(i)}\in\left\{\hat{y}_m^{(i)}\right\}_{m=1}^{M}}{\operatorname{argmax}} \mathcal{S}\left(\hat{y}_j^{(i)}, y_i^*\right)\mathcal{E}\left(\hat{y}_j^{(i)}, y_i^*\right)\right), \tag{4}$$

where $\mathcal{E}(\cdot,\cdot)$ is the indicator function determining whether the two sentences share equivalent semantics, i.e., $\mathcal{E}\left(\hat{y}_j^{(i)}, y_i^*\right) = 1$ indicates that the sampled generation $\hat{y}_j^{(i)}$ is semantically equivalent to y_i^*, and $\mathcal{E}\left(\hat{y}_j^{(i)}, y_i^*\right) = 0$ denotes it does not. This is, the NS $r\left(x_i, y_i^*\right)$ represents the uncertainty condition of the candidate generation $\hat{y}_j^{(i)}$, which has the highest similarity score with the reference answer y_i^* in generations that are semantically equivalent to y_i^*. The criterion for determining semantic equivalence here is the same as that for correctness evaluation (i.e., $\hat{y}_j^{(i)}$ is correct according to y_i^* if $\mathcal{E}\left(\hat{y}_j^{(i)}, y_i^*\right) = 1$).

It is worth noting that we strictly align NS with the uncertainty condition of the correct answer in sampled generations, which is crucial for robust correctness coverage guarantees in test samples.

Following prior work (Angelopoulos and Bates, 2021; Quach et al., 2023; Campos et al., 2024), we sort $\{r_i\}_{i=1}^{N}$ ($\{r_1 \le \cdots \le r_N\}$) and calculate the $\frac{\lceil (N+1)(1-\alpha)\rceil}{N}$ quantile of NSs for all calibration data to develop the conformal uncertainty criterion

$$\hat{q} = \inf\left\{q : \frac{|\{i : r_i \le q\}|}{N} \ge \frac{\lceil (N+1)(1-\alpha)\rceil}{N}\right\} = r_{\lceil (N+1)(1-\alpha)\rceil}, \tag{5}$$

where α is the upper bound of the error rate.

As for each test sample, we construct the prediction set following

$$\mathcal{P}(x_{test}) = \left\{\hat{y}_j^{(test)} \in \left\{\hat{y}_m^{(test)}\right\}_{m=1}^{M} : r\left(x_{test}, \hat{y}_j^{(test)}\right) \le \hat{q}\right\}. \tag{6}$$

It is evident that the most semantically similar generation to $\hat{y}_j^{(test)}$ in $\left\{\hat{y}_m^{(test)}\right\}_{m=1}^{M}$ is itself, and we obtain $r\left(x_{test}, \hat{y}_j^{(test)}\right) = \mathcal{U}\left(\hat{y}_j^{(test)}\right)$. Recall the assumption that $\left\{\hat{y}_m^{(test)}\right\}_{m=1}^{M}$ contains at least

Table 1: Performance comparison (AUROC) of uncertainty quantification across our proposed method and 8 baseline approaches, evaluated on 5 instruction-tuned LLMs over 4 open-ended NLG datasets. The correctness criterion is based on the sentence similarity measured by the DistillRoBERTa model with a threshold of 0.7.

Dataset	LLMs	White-box				Black-box				
		PE	*LNPE*	*SE*	*SAR*	*LS*	*NumSet*	*Ecc*	*Deg*	*ConU*
TriviaQA	LLaMA-2-7B-Chat	0.6587	0.6459	0.7495	0.7876	0.5571	0.7763	0.7839	0.8103	**0.8198**
	Mistral-7B-Instruct-v0.3	0.6620	0.5968	0.7845	0.8306	0.5969	0.8491	0.8596	0.8596	**0.8671**
	LLaMA-3-8B-Instruct	0.7247	0.6465	0.7934	0.8271	0.4661	0.8201	0.7404	0.8246	**0.8275**
	Vicuna-13B-v1.5	0.5553	0.5543	0.7568	0.7207	0.5734	0.7629	0.6578	0.7858	**0.7926**
	LLaMA-2-13B-Chat	0.6065	0.5614	0.7624	0.7757	0.6121	0.7885	0.8035	0.8035	**0.8048**
	Average	0.6414	0.6010	0.7693	0.7883	0.5611	0.7994	0.7690	0.8167	**0.8224**
CoQA	LLaMA-2-7B-Chat	0.6236	0.5618	0.7120	0.7372	0.5403	0.7309	0.6769	**0.7613**	0.7600
	Mistral-7B-Instruct-v0.3	0.6746	0.5795	0.7062	0.7551	0.5799	0.7481	0.6931	0.7645	**0.7652**
	LLaMA-3-8B-Instruct	0.7495	0.6531	0.7652	**0.7902**	0.4532	0.7400	0.7288	0.7763	0.7702
	Vicuna-13B-v1.5	0.5928	0.5565	0.7110	0.6984	0.4965	0.6832	0.6679	**0.7191**	0.7106
	LLaMA-2-13B-Chat	0.6203	0.5634	0.7039	0.7427	0.5534	0.7230	0.6805	0.7546	**0.7591**
	Average	0.6522	0.5829	0.7197	0.7472	0.5247	0.7250	0.6894	**0.7552**	0.7530
MedQA	LLaMA-2-7B-Chat	0.4888	0.4925	0.5341	0.5862	0.5599	0.5933	0.5511	0.6064	**0.6120**
	Mistral-7B-Instruct-v0.3	0.4613	0.4639	0.5091	0.6397	0.5520	0.6282	0.6562	0.6660	**0.6789**
	LLaMA-3-8B-Instruct	0.5854	0.5781	0.6508	0.7167	0.4522	0.7093	0.6142	0.7159	**0.7196**
	Vicuna-13B-v1.5	0.4970	0.4922	0.5523	0.5854	0.5479	0.5926	0.5383	0.6261	**0.6360**
	LLaMA-2-13B-Chat	0.4618	0.4647	0.5277	0.5792	0.5734	0.6041	0.5743	0.6070	**0.6153**
	Average	0.4989	0.4983	0.5548	0.6214	0.5371	0.6255	0.5868	0.6443	**0.6524**
MedMCQA	LLaMA-2-7B-Chat	0.4774	0.4848	0.5221	0.5883	0.5531	0.6171	0.5165	0.5983	**0.6330**
	Mistral-7B-Instruct-v0.3	0.4971	0.4989	0.5491	0.6944	0.5103	0.7084	0.7170	0.7173	**0.7413**
	LLaMA-3-8B-Instruct	0.5414	0.5395	0.6244	0.6940	0.4817	0.6992	0.5952	0.6993	**0.7098**
	Vicuna-13B-v1.5	0.4614	0.4815	0.5550	0.5509	0.5377	0.5891	0.5135	0.6221	**0.6448**
	LLaMA-2-13B-Chat	0.4547	0.4712	0.5385	0.5701	0.5711	0.6378	0.6188	0.6188	**0.6414**
	Average	0.4864	0.4952	0.5578	0.6195	0.5308	0.6503	0.5922	0.6511	**0.6741**

one correct generation (i.e., $y^*_{test} \in \left\{\hat{y}_m^{(test)}\right\}_{m=1}^{M}$), then the event $\{y^*_{test} \in \mathcal{P}(x_{test})\}$ is equivalent to $\{r_{test} = r(x_{test}, y^*_{test}) \leq \hat{q}\}$.

Since the calibration and test samples (x_1, y^*_1), ..., (x_N, y^*_N), (x_{test}, y^*_{test}) are exchangeable, we have $P(r_{test} \leq r_i) = \frac{i}{N+1}$. Then we conclude

$$\begin{aligned} P\left(y^*_{test} \in \mathcal{P}(x_{test})\right) &= P\left(r_{test} \leq r_{\lceil (N+1)(1-\alpha) \rceil}\right) \\ &= \frac{\lceil (N+1)(1-\alpha) \rceil}{N+1} \\ &\geq 1-\alpha, \end{aligned} \tag{7}$$

and obtain the user-specified lower bound (i.e., $1-\alpha$) of the correctness coverage rate guaranteed by the calibrated prediction sets.

4 Evaluations

4.1 Experimental Set-up

Baselines. We consider 8 baseline methods, including 4 white-box methods: Predictive Entropy (*PE*) (Kadavath et al., 2022), Length-normalized Predictive Entropy (*LNPE*) (Malinin and Gales, 2020), Semantic Entropy (*SE*) (Kuhn et al., 2023), and Shift Attention to Relevance (*SAR*) (Duan et al., 2024), and 4 black-box approaches: Lexical Similarity (*LS*) (Lin et al., 2022b) and Number of Semantic Sets (*NumSet*) (Kuhn et al., 2023; Lin et al., 2023). Moreover, we also include the most recent state-of-the-art uncertainty quantification methods, Degree Matrix (*Deg*) (Lin et al., 2023), and Eccentricity (*Ecc*) (Lin et al., 2023). More details of baseline methods can be found in Appendix B.1.

Base LLMs. We conduct experimental evaluations on 6 open-source LLMs encompassing various sizes and architectures for comprehensive analysis, including LLaMA-2-7B-Chat (Touvron et al., 2023b), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Llama-3-8B-Instruct (AI@Meta, 2024), Vicuna-13B-v1.5 (Zheng et al., 2023), LLaMA-2-13B-Chat (Touvron et al., 2023b), LLaMA-3-70B-Instruct (AI@Meta, 2024). We utilize the default generation configs and checkpoints provided by the

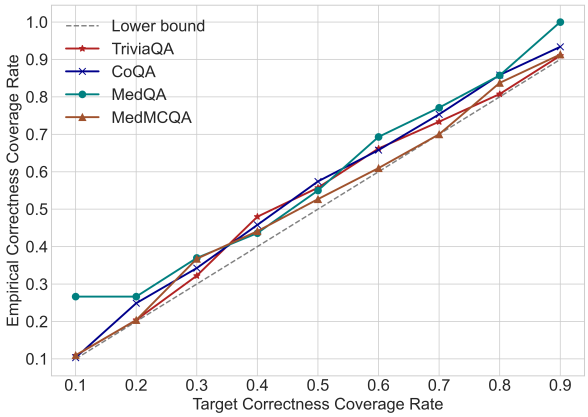

Figure 1: Target vs. empirical correctness coverage rate. We test the 4 datasets utilizing the LLaMA-2-7B-Chat model as the generator. Empirically, we achieve strict control over the coverage of correct answers by calibrating prediction sets on 4 free-form QA datasets.

Table 2: The results of correctness coverage rate (%) on 6 LLMs with various sizes across 4 open-ended NLG datasets. The user-specified error rate α is set to 0.1.

LLMs	TriviaQA	CoQA	MedQA	MedMCQA
LLaMA-2-7B-Chat	91.00	93.37	100.00	91.32
Mistral-7B-Instruct-v0.3	90.83	91.87	90.70	90.39
LLaMA-3-8B-Instruct	94.27	90.73	90.46	93.17
LLaMA-2-13B-Chat	91.68	91.63	91.72	92.45
Vicuna-13B-v1.5	90.19	92.68	90.25	92.13
LLaMA-3-70B-Instruct	92.18	90.95	93.70	92.48

Table 3: The average prediction set size on 6 LLMs with various sizes across 4 open-ended NLG datasets. The user-specified error rate α is set to 0.1.

LLMs	TriviaQA	CoQA	MedQA	MedMCQA
LLaMA-2-7B-Chat	2.28	2.26	4.28	3.07
Mistral-7B-Instruct-v0.3	2.24	2.49	4.20	3.26
LLaMA-3-8B-Instruct	2.34	2.45	2.68	2.60
LLaMA-2-13B-Chat	2.19	2.28	3.40	2.73
Vicuna-13B-v1.5	2.26	2.47	3.29	2.98
LLaMA-3-70B-Instruct	1.03	1.71	2.15	1.60

HuggingFace platform[1] for all models.

Datasets. We evaluate the performance of *ConU* and verify the correctness coverage guarantees on 4 free-form NLG datasets, including CoQA (Reddy et al., 2019) for conversational QA task, TriviaQA (Joshi et al., 2017) for reading comprehension, MedQA (Jin et al., 2021) for solving medical problems, and MedMCQA (Pal et al., 2022) for medical entrance exam questions. More details of datasets can be found in Appendix B.2.

Evaluation Metric. Following prior work (Duan et al., 2024; Wang et al., 2024b), we evaluate the performance of UQ by treating it as the problem of predicting whether to trust a generation given the prompt, and utilize the Area Under the Receiver Operating Characteristic Curve (AUROC) which gauges if the uncertainty scores can effectively distinguish between correct and incorrect generations. To verify if the correctness coverage is strictly guaranteed, we evaluate the coverage rate under various user-specified error rates. We also report the average prediction set size to evaluate the prediction efficiency and practicality of our approach.

Correctness and Equivalence Metric. We utilize sentence similarity (Duan et al., 2024) as the metric for correctness and equivalence evaluation. We employ the *cross-encoder* model (Reimers and Gurevych, 2019) with DistillRoBERTa (Sanh et al., 2019) serving as the backbone to measure the semantic similarity score between the most likely generation and reference answer and set a strict correctness threshold of 0.7.

Hyperparameters. We randomly sample 5 answers to each question for UQ and 10 candidate generations for verification of correctness coverage guarantees. We leverage beam search for the most likely generations for correctness evaluation and multinominal sampling for candidate generations (Duan et al., 2024). The max length of each generation is set to 128 tokens. The temperature of generation is set to 1.0. The coefficient λ introduced in Eq. (1) is set to 0.5. The ratio of calibration and test set is set to 1:10 by default.

4.2 UQ in Black-Box LLMs

As defined in failure prediction (Xiong et al., 2023) which evaluates whether the uncertainty score can effectively distinguish between correct and incorrect generations, an effective measure should assign higher uncertainty to incorrect generations and lower to correct ones. We compare our approach with state-of-the-art methods utilizing AUROC. Experimental results are summarized in Table 1. Generally, our method outperforms baseline methods in most of the settings. For instance, our method consistently beat 8 baseline methods on the TriviaQA datasets. It is worth noting that our method outperforms other methods by at most 2.4% AUROC on the MedMCQA dataset and 1.29% AUROC on the MedQA, which indicates the potential impacts of our methods on real-world high-stakes NLG appli-

[1] https://huggingface.co/models

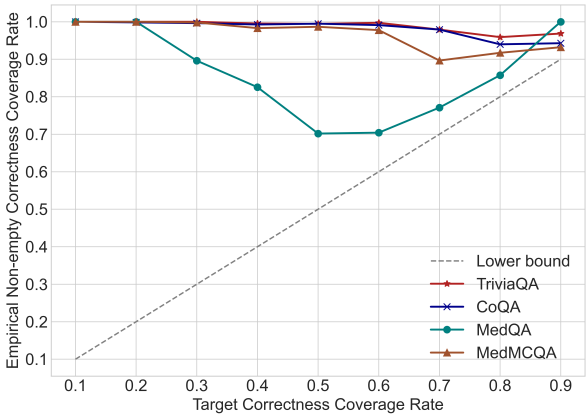

Figure 2: Target correctness coverage rate vs. empirical correctness coverage rate on non-empty prediction sets. We test the 4 datasets utilizing the LLaMA-2-7B-Chat model. We can almost obtain absolute coverage of correct answers in non-empty calibrated prediction sets even at a strict user-accepted error rate.

Table 4: The enhancement of model accuracy (%) after conducting selective prediction within the calibrated prediction sets based on the black-box uncertainty measure, utilizing sentence similarity as the criterion for correctness evaluation under the threshold of 0.7.

Dataset	LLMs	Original	Calibrated
TriviaQA	LLaMA-2-7B-Chat	68.43	70.77
	Mistral-7B-Instruct-v0.3	79.04	81.45
	LLaMA-3-8B-Instruct	79.36	80.00
	Vicuna-13B-v1.5	78.40	78.80
	LLaMA-2-13B-Chat	76.70	78.13
CoQA	LLaMA-2-7B-Chat	73.00	75.53
	Mistral-7B-Instruct-v0.3	78.25	80.80
	LLaMA-3-8B-Instruct	72.93	74.67
	Vicuna-13B-v1.5	76.17	78.43
	LLaMA-2-13B-Chat	80.00	81.23
MedQA	LLaMA-2-7B-Chat	37.88	40.80
	Mistral-7B-Instruct-v0.3	38.65	43.90
	LLaMA-3-8B-Instruct	66.29	70.59
	Vicuna-13B-v1.5	44.42	46.78
	LLaMA-2-13B-Chat	42.07	46.15

cations. We discuss the impact of the number of sampled generations on UQ in Section 4.4.

4.3 Conformal Correctness Coverage

In this section, we verify that the calibrated prediction sets constructed following Eq. (6) indeed achieve rigorous correctness coverage guarantees under various user-specified error rates as described in Eq. (7). Then we explore the utility of prediction sets and conduct selective prediction based on our proposed uncertainty measure.

Empirical Coverage Guarantees. To guarantee the derived lower bound of correctness coverage rate in practice, we randomly split the four datasets at a ratio of 1:10, employing the respective portions as the calibration and test set. We utilize the calibration set to derive the conformal uncertainty criterion specified by the upper bound of the error rate. Then, we measure the correctness coverage rate on the test set and plot the results on four datasets in Figure 1. It is evident that we achieve strict control of the correctness coverage rate under various error rates. The verification on other models can be found in Appendix C.

Following (Ye et al., 2024), we set the error rate α to 0.1 and test the coverage rate on 4 datasets utilizing 7 LLMs with multiple scales. As is exhibited in Table 2, the coverage rate is at least 90%, indicating that the requirement of correctness coverage guarantees is satisfied. It is worth noting that prior work (Ye et al., 2024; Kumar et al., 2023) selects the possible option from the fixed choices while we characterize the unbound answer distribution by sampling and utilize our devised conformal uncertainty criterion to search for the correct answer, which is more practical.

We also evaluate the prediction efficiency of the conformal uncertainty criterion utilizing the average size of these calibrated prediction sets, which is the primary metric for CP (Angelopoulos and Bates, 2021). Table 3 demonstrates that the average size of prediction sets calibrated by our method remains very small across the 4 datasets. For instance, the average set size is 1.03 on the LLaMa-3-70B-Instruct model in the TriviaQA task, indicating that we can almost directly identify the correct answers through these calibrated prediction sets.

We boldly expect that as long as the language model has the capability to solve the current problem, despite the unfixed answer distribution, we can always find the correct generation by performing black-box UQ on each sampled answer and searching for answers meeting the conformal uncertainty criterion, and then limit the selection region to the calibrated prediction set for post-processing.

Utility of Calibrated Prediction Sets. Since for some test samples, all the candidate generations can be filtered out by the conformal uncertainty criterion, we explore the utility of non-empty prediction sets in practice. Figure 2 exhibits that the prediction sets achieve promising correctness coverage rate, raising to 100% as the accepted error rate increases. In the MedQA dataset, while the

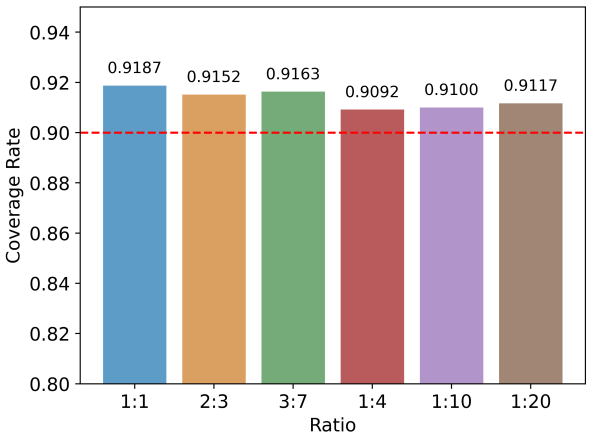

Figure 3: The average correctness coverage rate across 4 datasets at different ratios of the size between the calibration and test set utilizing the LLaMA-3-8B-Instruct model. The accepted error rate is set to 0.1, and the red dashed line in the figure indicates the lower bound of accuracy at 0.9 (i.e., $\alpha = 0.1$).

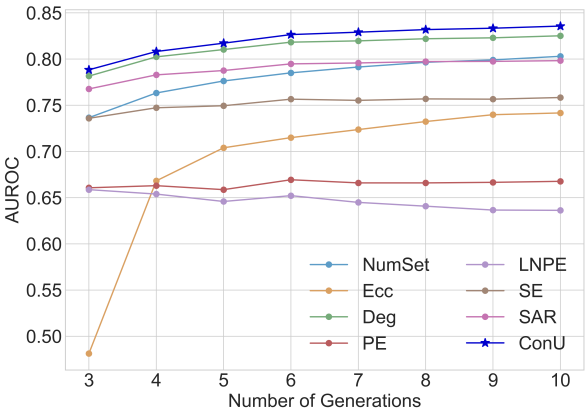

Figure 4: The performance of UQ over various numbers of generations. Results are obtained from the LLaMA-3-8B-Instruct model on the TriviaQA dataset. Our method consistently surpasses 7 baseline methods.

error rate is set to 0.1, we almost achieve absolute correctness coverage guarantees, indicating that, without reference answers provided in real-world high-stakes situations, we can ensure that the small reference range we have established contains the correct answer for posterior selection, and then high-uncertainty problems will be handed over to experts, which aligns with the selective prediction and abstention criterion.

Based on the proposed uncertainty measure, we conduct post-processing to select the generation with the lowest uncertainty score from each calibrated prediction set and evaluate the total selective accuracy. It is worth noting that the performance depends on the quality of the uncertainty measure. Results are summarized in Table 4. Through posterior selection, we obtain promising accuracy improvement despite several empty prediction sets.

4.4 Ablation Studies

Considering that these sampling-based methods integrate multiple generations within the candidate set, We investigate the effects of the number of sampled generations (i.e., M) on the performance of UQ. As illustrated in Figure 4, our uncertainty measure consistently outperforms the baseline approaches, and its performance can be further boosted by incorporating more generations. While employing just 4 generations, our method is able to achieve the highest AUROC of 0.8082, demonstrating its generation-efficient nature.

As described in Section 3.3, conformal prediction assumes a calibration set for the threshold $\hat{q}$. In our prior analysis, We divide the dataset into the calibration and test set at a fixed ratio of 1:10. Here, we investigate the correctness coverage rate at different ratios of size between the calibration and test set, and present the results in Figure 3. Despite various ratios of set size, we can always obtain a strict lower bound of correctness coverage by constructing prediction sets based on our devised conformal uncertainty criterion. This indicates the potential impacts of our method for robust guarantees in real-world open-ended NLG applications.

5 Conclusion

In this work, we introduce *ConU* tailored for black-box UQ in open-ended NLG tasks. Relying on CP which can transform any heuristic approximation into a statistically rigorous uncertainty notion, we develop a robust conformal uncertainty criterion to provide reliable guarantees of correctness coverage under various user-specified error rates. We achieve strict control of the coverage rate across 6 practical LLMs on 4 free-from NLG datasets. Furthermore, the small average uncertainty set size underscores the efficiency of our methods. Utilizing these calibrated prediction sets, we perform selective prediction and obtain remarkable improvements in model accuracy. We envisage that our conformal uncertainty criterion can provide new strategies for principled UQ in open-ended NLG tasks.

Limitations

Our approach has some limitations. In our study, we assume that at least one correct answer exists

in the candidate generations. However, we need to develop a criterion to verify whether the correct answer has been sampled from the unbound output space in real-world applications. Secondly, our findings are limited to the four datasets and future works will extend to other typical NLG tasks like document summarization. Finally, we will attempt to expand our conformal uncertainty criterion to non-exchangeability scenarios, aiming to establish a general criterion across different NLG tasks.

References

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*.

AI@Meta. 2024. Llama 3 model card.

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

Margarida M Campos, António Farinhas, Chrysoula Zerva, Mário AT Figueiredo, and André FT Martins. 2024. Conformal prediction for natural language processing: A survey. *arXiv preprint arXiv:2405.01976*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

Zhipeng Chen, Kun Zhou, Beichen Zhang, Zheng Gong, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Chatcot: Tool-augmented chain-of-thought reasoning on chat-based large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14777–14790.

Jinhao Duan, Hao Cheng, Shiqi Wang, Chenan Wang, Alex Zavalny, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the uncertainty estimation of large language models. In *The 62nd Annual Meeting of the Association for Computational Linguistics*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

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*.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

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*.

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

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. 2022b. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. *Findings of the Association for Computational Linguistics: ACL 2022*.

Andrey Malinin and Mark Gales. 2020. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees. *arXiv preprint arXiv:2402.10978*.

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.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. 2023. Conformal language modeling. *arXiv preprint arXiv:2306.10193*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. 2024. Api is enough: Conformal prediction for large language models without logit-access. *arXiv preprint arXiv:2403.01216*.

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*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Fangxin Wang, Lu Cheng, Ruocheng Guo, Kay Liu, and Philip S Yu. 2024a. Equal opportunity of coverage in fair regression. *Advances in Neural Information Processing Systems*, 36.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zhiyuan Wang, Jinhao Duan, Chenxi Yuan, Qingyu Chen, Tianlong Chen, Huaxiu Yao, Yue Zhang, Ren Wang, Kaidi Xu, and Xiaoshuang Shi. 2024b. Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond. *arXiv preprint arXiv:2402.14259*.

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*.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, et al. 2024. Mitigating llm hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *arXiv preprint arXiv:2401.12794*.

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. 2024. Characterizing truthfulness in large language model generations with local intrinsic dimension. *arXiv preprint arXiv:2402.18048*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Proof of the Coverage Property

This is the explanation of validity for the conformal uncertainty criterion introduced in Section 3.3. We reproduce the derivation here for completeness. Let us break down the overall implementation into the following five steps:

Black-box Uncertainty Measure. We first conduct semantic clustering within the M candidate generations and obtain K non-repeated semantics for each sample. Since generations in the k-th cluster share the equivalent meaning, we denote any one generation in the k-th cluster as $\hat{y}_k^{(i)}$. Then we rely on self-consistency and define the uncertainty score of each candidate generation as $\mathcal{U}\left(\hat{y}_m^{(i)}\right)$ as described in Eq. (1).

NS Definition. For each calibration sample, we select the generation that first shares the equivalent semantics with the reference answer and then exhibits the highest semantic similarity to the reference answer, and then define the NS as its uncertainty score calculated following Eq. (1). The first condition is to tightly couple the NS with correctness and the second is to select generations in test samples. The NS of the i-th calibration data r_i is described as Eq. (4).

Conformal Uncertainty Criterion. We calculate the $\frac{\lceil (N+1)(1-\alpha) \rceil}{N}$ quantile of the NSs for all calibration data to develop our conformal uncertainty criterion (i.e., the uncertainty threshold $\hat{q}$) based on the user-specified error rate α. As described in Eq. 5, $\hat{q} = r_{\lceil (N+1)(1-\alpha) \rceil}$.

Construction of Prediction Sets. For each test data, we construct the prediction set following Eq. (6). Since the generation that is semantically equivalent to $\hat{y}_i^{(test)}$ and shares the highest semantic similarity to $\hat{y}_i^{(test)}$ in $\left\{\hat{y}_m^{(test)}\right\}_{m=1}^{M}$ is itself, we obtain $r\left(x_{test}, \hat{y}_j^{(test)}\right) = \mathcal{U}\left(\hat{y}_j^{(test)}\right)$. Then we calibrate the prediction set by selecting generations, of which the uncertainty satisfies the conformal uncertainty criterion closely linked with correctness.

Correctness Coverage Guarantees. Considering the assumption that there is at least one correct answer in $\left\{\hat{y}_m^{(test)}\right\}_{m=1}^{M}$, we can conclude that the event $\{y_{test}^* \in \mathcal{P}(x_{test})\}$ is equivalent to $\{r_{test} = r(x_{test}, y_{test}^*) \leq \hat{q}\}$. Since (x_1, y_1^*), ..., (x_N, y_N^*), (x_{test}, y_{test}^*) are exchangeable, we have $P(r_{test} \leq r_i) = \frac{i}{N+1}$. Ultimately, we achieve rigorous guarantees of the correctness coverage rate on test samples as described as Eq. (7).

B Implementation Details

B.1 Baselines

We compare *ConU* with 8 baseline measures. *PE* is defined as the entropy over the whole generation and *LNPE* is the length normalized *PE*. *SE* tackles the issue of semantic equivalence by gathering generations sharing the same meaning into semantic clusters and calculating cluster-wise entropy. *SAR* solves the issue of generative inequality and allocates more attention to key tokens and sentences. *LS* measures the average sentence similarity among sampled responses. *NumSet* employs the number of semantic sets (equivalence classes) as a reflection of uncertainty. *Deg* and *Ecc* treat each generation as one node, calculate the symmetric normalized graph Laplacian, and respectively utilize the degree matrix and the average distance from the center as the uncertainty measures.

We do not compare the two recent approaches that adapt CP for correctness coverage in open-ended NLG tasks for several reasons: (1) Conformal language modeling (Quach et al., 2023) relies on the white-box model likelihoods information, which is impractical for recent LLMs served via API without logit access; (2) LofreeCP (Su et al., 2024) is susceptible to different settings of datasets and models, and cannot consistently guarantee the correctness coverage rate; (3) Our conformal uncertainty criterion achieves strict control of the correctness coverage rate under various user-specified error rates, model settings, and datasets, first linking black-box UQ with rigorous guarantees of correctness coverage, which meets the requirement for general NLG applications.

B.2 Datasets

CoQA (Reddy et al., 2019) is a large-scale conversational QA dataset with more than 127k question-answer pairs equipped with contextual information. TriviaQA (Joshi et al., 2017) is a reading comprehension dataset with over 650k question-answer pairs. MedQA (Jin et al., 2021) is a medical MCQA dataset collected from professional medical board exams. MedMCQA (Pal et al., 2022) is a large-scale MCQA dataset for practical medical entrance exam questions. For the evaluation of UQ, we randomly select 3,000 samples from each dataset. For the verification of correctness coverage guarantees, we utilize the development set (7,983 questions) of CoQA and full validation sets of MedQA and MedMCQA. For TriviaQA, we utilize the same

3,000 samples in UQ evaluations.

For CoQA, we utilize the contextual information combined with the question as the prompt. For TriviaQA and MedMCQA, we randomly select 5 question-answer pairs as a fixed few-shot template and combine it with the current question. For MedQA, we employ 3 question-answer pairs.

C Robustness of Conformal Uncertainty Criterion

We verify the correctness coverage guarantees on other 5 LLMs across 4 datasets. As demonstrated in Figures 5 ~ 9, we achieve rigorous control of coverage rate under various user-specified error rates despite different model settings or datasets. We also report the results of the correctness coverage rate under two strict error rates of 0.05 and 0.01. Table 5 and Table 6 indicate the robustness of our conformal uncertainty criterion.

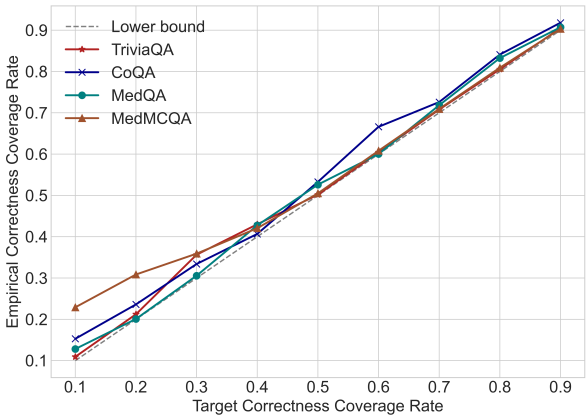

Figure 5: Target vs. empirical correctness coverage rate. We test the 4 datasets utilizing the Mistral-7B-Instruct-v0.3 model as the generator.

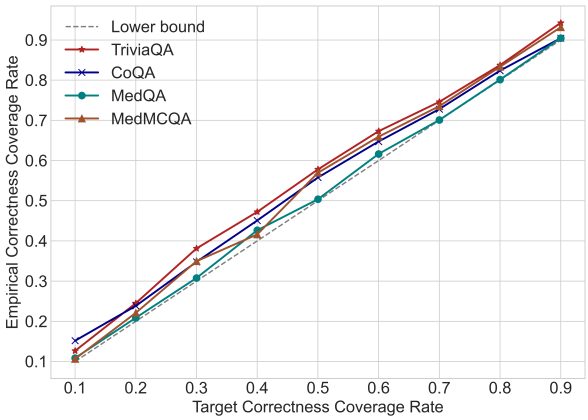

Figure 6: Target vs. empirical correctness coverage rate. We test the 4 datasets utilizing the LLaMA-3-8B-Instruct model as the generator.

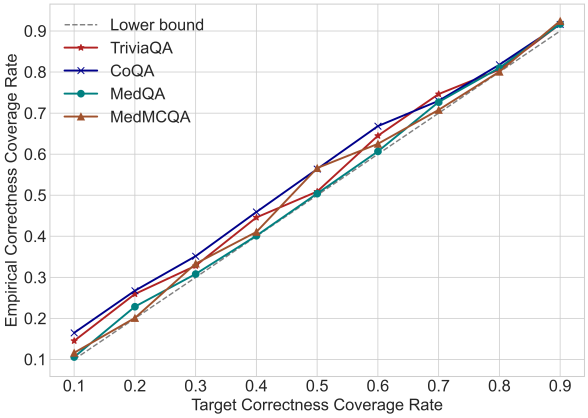

Figure 7: Target vs. empirical correctness coverage rate. We test the 4 datasets utilizing the LLaMA-2-13B-Chat model as the generator.

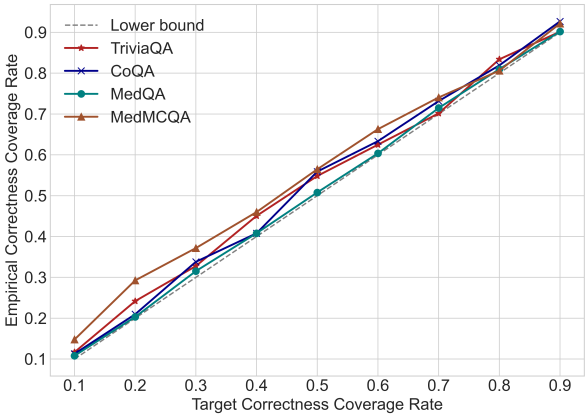

Figure 8: Target vs. empirical correctness coverage rate. We test the 4 datasets utilizing the Vicuna-13B-v1.5 model as the generator.

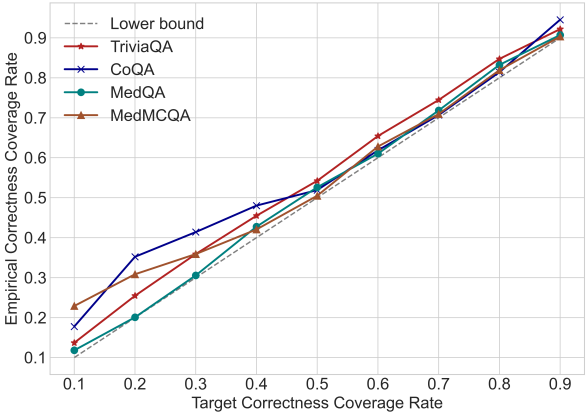

Figure 9: Target vs. empirical correctness coverage rate. We test the 4 datasets utilizing the LLaMA-3-70B-Instruct model as the generator.

Table 5: The results of correctness coverage rate (%) on 6 LLMs across 4 open-ended NLG datasets. The user-accepted error rate α is strictly set to 0.05.

LLMs	TriviaQA	CoQA	MedQA	MedMCQA
LLaMA-2-7B-Chat	95.26	96.45	100.00	95.99
Mistral-7B-Instruct-v0.3	95.01	95.72	95.79	95.12
LLaMA-3-8B-Instruct	98.17	95.23	95.78	98.38
LLaMA-2-13B-Chat	95.04	96.96	95.15	96.59
Vicuna-13B-v1.5	97.28	95.33	95.51	97.29
LLaMA-3-70B-Instruct	95.38	95.33	95.51	97.29

Table 6: The results of correctness coverage rate (%) on 6 LLMs across 4 open-ended NLG datasets. The user-accepted error rate α is strictly set to 0.01.

LLMs	TriviaQA	CoQA	MedQA	MedMCQA
LLaMA-2-7B-Chat	99.93	99.83	100.00	99.14
Mistral-7B-Instruct-v0.3	99.38	99.27	99.15	99.81
LLaMA-3-8B-Instruct	99.79	99.53	100.00	99.76
LLaMA-2-13B-Chat	99.06	99.13	99.51	99.48
Vicuna-13B-v1.5	99.52	100.00	99.94	100.00
LLaMA-3-70B-Instruct	99.84	99.75	99.15	99.82