process as

$$\mathcal{U}\left(\left\{\hat{y}_m^{(i)}\right\}_{m=1}^{M}|x_i\right) = 1 - \lambda \cdot \mathcal{F}\left(\hat{y}_{mst}^{i}\right) - (1-\lambda)\cdot\frac{1}{K}\sum_{k=1}^{K}\mathcal{S}\left(\hat{y}_{mst}^{i}, \hat{y}_k^{(i)}\right)\mathcal{F}\left(\hat{y}_k^{(i)}\right). \tag{2}$$

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 split 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 (Angelopoulos and Bates, 2021). We provide the explanation for Assumption (1) in Appendix B.

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

$$r_i = r\left(x_i, y_i^*\right) = \mathcal{U}\left(\arg\max_{\hat{y}_j^{(i)}} \mathcal{S}\left(\hat{y}_j^{(i)}, y_i^*\right)\mathcal{E}\left(\hat{y}_j^{(i)}, y_i^*\right)\right), \tag{3}$$

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 $\hat{y}_j^{(i)}$ is semantically equivalent to y_i^*, and $\mathcal{E}\left(\hat{y}_j^{(i)}, y_i^*\right) = 0$ denotes it does not. That 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 emphasizing that we strictly align the NSs with the uncertainty conditions of correct answers within the fresh calibration set, concluding an honest insight into the model's performance, which is crucial for robust correctness coverage guarantees in new test samples.

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

$$\hat{q} = \inf\left\{q : \frac{|\{i : r_i \leq q\}|}{N} \geq \frac{\lceil (N+1)(1-\alpha)\rceil}{N}\right\} = r_{\lceil (N+1)(1-\alpha)\rceil}, \tag{4}$$

where α is the upper bound of the error rate.

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

$$\mathcal{P}\left(x_{test}\right) = \left\{\hat{y}_j^{(test)} : r\left(x_{test}, \hat{y}_j^{(test)}\right) \leq \hat{q}\right\}. \tag{5}$$

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. The best UQ methods are in **bold** and the second-best one is underscored.

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

$$\begin{aligned} P\left(y^*_{test} \in \mathcal{P}(x_{test})\right) &= P\left(r_{test} \leq r_{\lceil (N+1)(1-\alpha) \rceil}\right) \\ &= \frac{\lceil (N+1)(1-\alpha) \rceil}{N+1} \\ &\geq 1-\alpha, \end{aligned} \tag{6}$$

and obtain the user-specified lower bound (i.e., $1-\alpha$) of the correctness coverage rate guaranteed by these 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., 2024a), 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 C.1.

Base LLMs. We conduct empirical evaluations on 7 LLMs encompassing various sizes and architectures for comprehensive analysis, including GPT-3.5-turbo served by OpenAI(OpenAI, 2021), 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

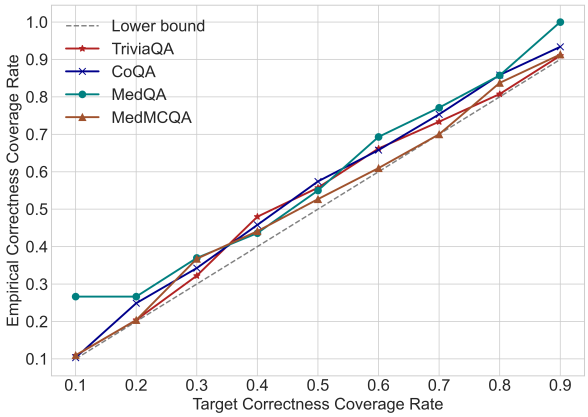

Figure 1: Target vs. empirical correctness coverage rate. We test the 4 datasets utilizing the LLaMA-2-7B-Chat model as the generator. Empirically, we achieve strict control over the coverage of correct answers by calibrating prediction sets on 4 free-form QA datasets.

generation configs and checkpoints provided by the HuggingFace platform[†] for all open-source LLMs.

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 C.2.

Evaluation Metric. Following prior work (Duan et al., 2024a; 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., 2024a) 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.

[†]https://huggingface.co/models

Table 2: The results of correctness coverage rate (%) on 7 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
GPT-3.5-turbo	93.14	91.66	91.78	90.36

Table 3: The average prediction set size on 7 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
GPT-3.5-turbo	1.96	2.13	2.49	2.02

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., 2024a). 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