

- 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*.
- OpenAI. 2021. [Chatgpt](#).
- 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. 2024. Conformal language modeling. In *International Conference on Learning Representations*.
- 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*.
- Tianyi Zhao, Jian Kang, and Lu Cheng. 2024. Conformalized link prediction on graph neural networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4490–4499.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zuohan 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 $\mathcal{U}(\hat{y}_m^{(i)})$ as described in Eq. (1).

NS Definition. For each calibration sample, we select the generation that (1) first shares the equivalent semantics with the reference answer and (2) 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 facilitate generation selection in test samples. The NS of the i -th calibration data r_i is described as Eq. (3).

Conformal Uncertainty Criterion. We calculate the $\frac{\lceil(N+1)(1-\alpha)\rceil}{N}$ quantile of the NSs for all fresh 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. 4, $\hat{q} = r_{\lceil(N+1)(1-\alpha)\rceil}$.

Construction of Prediction Sets. For each test data, we construct a prediction set following Eq. (5). Since the generation that is semantically equivalent to $\hat{y}_i^{(test)}$ and shares the highest semantic similarity to $\hat{y}_i^{(test)}$ in $\{\hat{y}_m^{(test)}\}_{m=1}^M$ is itself, we can obtain $r(x_{test}, \hat{y}_j^{(test)}) = \mathcal{U}(\hat{y}_j^{(test)})$. 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 $\{\hat{y}_m^{(test)}\}_{m=1}^M$, 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 $(x_1, y_1^*), \dots, (x_N, y_N^*)$, (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. (6).

B Validity of Assumption (1)

We assume that at least one acceptable response is sampled into the candidate set for each test data point. For each calibration data point, we sample multiple generations from the output space, denoted as $\mathcal{C}_m(X_i) = \{\hat{Y}_j^{(i)}\}_{j=1}^m$. Then, we define the loss of miscoverage by the candidate set as

$$l(\mathcal{C}_m(X_i), Y_i^*) = \mathbf{1}\{Y_i^* \notin \mathcal{C}_m(X_i)\}, \quad (7)$$

and the loss is non-increasing in m .

We set $A_N(m) = \sum_{i=1}^N l(\mathcal{C}_m(X_i), Y_i^*)$. Given that $l(\mathcal{C}_m(X_{test}), Y_{test}^*) \in \{0, 1\}$, we obtain

$$\begin{aligned} & A_{N+1}(m) \\ &= \sum_{i=1}^{N+1} l(\mathcal{C}_m(X_i), Y_i^*) \\ &= A_N(m) + l(\mathcal{C}_m(X_{test}), Y_{test}^*) \\ &\in \{A_N(m), A_N(m) + 1\}. \end{aligned} \quad (8)$$

By the exchangeability of N calibration data points and the test data point, we have $l_{test} \sim \text{Uniform}(\{l_1, \dots, l_N, l_{test}\})$, where l_i is the abbreviation for $l(\mathcal{C}_m(X_i), Y_i^*)$ (Angelopoulos et al., 2024). Then, we have

$$\begin{aligned} & \mathbb{E}[l(\mathcal{C}_m(X_{test}), Y_{test}^*)] \\ &= \frac{A_{N+1}(m)}{N+1} \\ &\in \left\{ \frac{A_N(m)}{N+1}, \frac{A_N(m)+1}{N+1} \right\}. \end{aligned} \quad (9)$$

Since we have demanded that at least one acceptable response is sampled into the candidate set for each calibration data (i.e., $A_N(m) = 0$), we obtain $\mathbb{E}[l(\mathcal{C}_m(X_{test}), Y_{test}^*)] \in \left\{0, \frac{1}{N+1}\right\}$ and Assumption (1) holds in this case.

C Implementation Details

C.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., 2024) 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.

C.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.

D Robustness of Conformal Uncertainty Criterion

We verify the correctness coverage guarantees on the other 6 LLMs across 4 datasets. As demonstrated in Figures 5 ~ 10, 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.

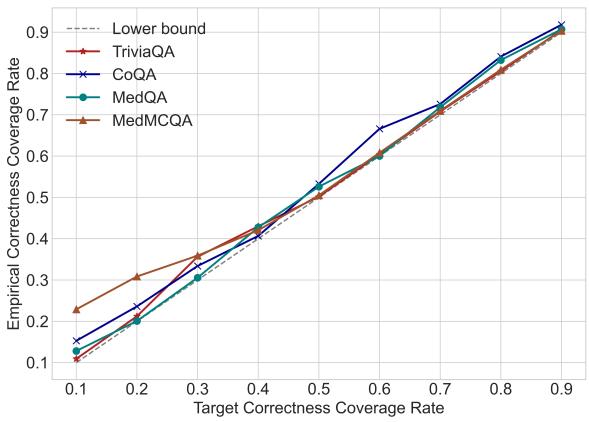


Figure 5: Target vs. empirical correctness coverage rate. We test the 4 datasets utilizing the Mistral-7B-Instruct-v0.3 model as the generator.

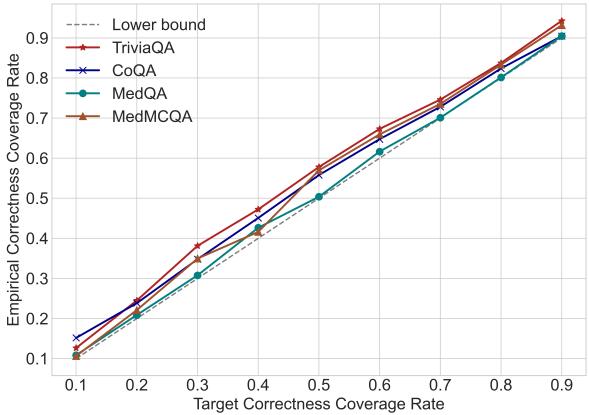


Figure 6: Target vs. empirical correctness coverage rate. We test the 4 datasets utilizing the LLaMA-3-8B-Instruct model as the generator.