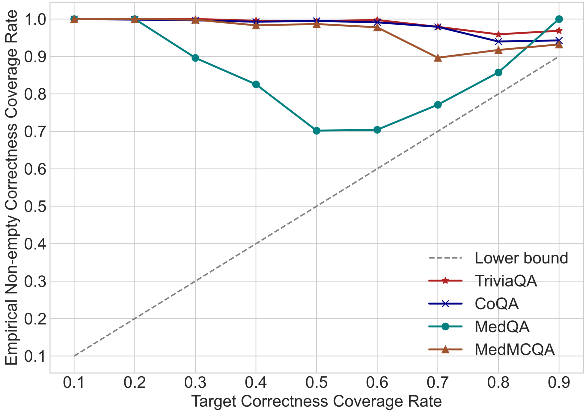

Figure 2: Target correctness coverage rate vs. empirical correctness coverage rate on non-empty prediction sets. We test the 4 datasets utilizing the LLaMA-2-7B-Chat model. We can almost obtain absolute coverage of correct answers in non-empty calibrated prediction sets even at a strict user-accepted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

our methods on real-world high-stakes NLG applications. We will 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. (5) indeed achieve rigorous correctness coverage guarantees under various user-specified error rates as described in Eq. (6). 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 D.

Following the study (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 various 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.

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

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

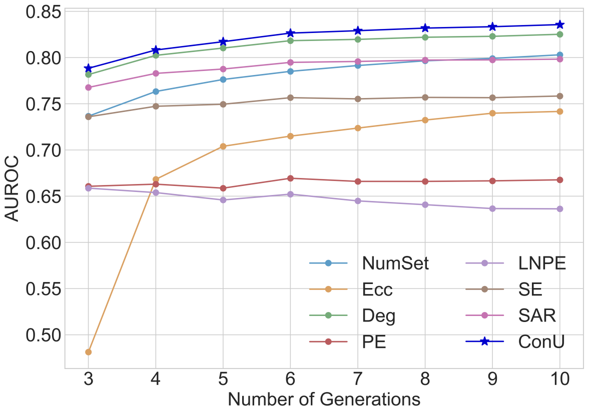

Figure 3: The performance of UQ over various numbers of generations. Results are obtained from the LLaMA-3-8B-Instruct model on the TriviaQA dataset. Our method consistently surpasses 7 baseline methods.

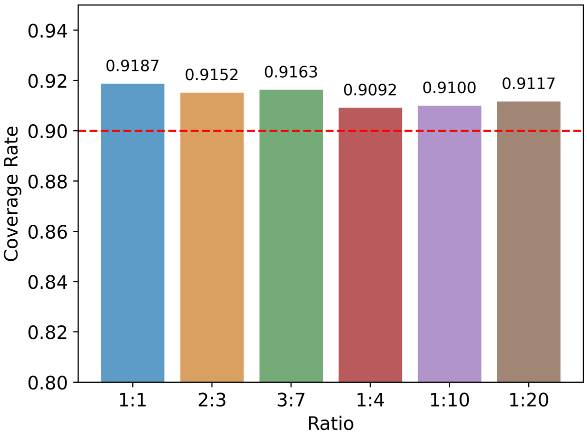

Figure 4: The average coverage rate across 4 datasets at different ratios between the calibration and test set utilizing the LLaMA-3-8B-Instruct model. The red dashed line indicates the lower bound at 0.9 (i.e., $\alpha = 0.1$).

rate increases. In the MedQA dataset, while the 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 3, 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 4. Despite various ratios of set size, we can always obtain a strict lower bound of the coverage rate 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 7 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.

Acknowledgments

Zhiyuan Wang, Xiaoshuang Shi, and Xiaofeng Zhu were supported by the National Key Research & Development Program of China under Grant (No. 2022YFA1004100).

Limitations

Our approach has some limitations. We need to develop an uncertainty criterion to verify whether the correct answer has been sampled from the 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*.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2024. Conformal risk control. In *The Twelfth International Conference on Learning Representations*.

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.

Longchao Da, Tiejin Chen, Lu Cheng, and Hua Wei. 2024. Llm uncertainty quantification through directional entailment graph and claim level response augmentation. *arXiv preprint arXiv:2407.00994*.

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

Jinhao Duan, Shiqi Wang, James Diffenderfer, Lichao Sun, Tianlong Chen, Bhavya Kailkhura, and Kaidi Xu. 2024b. Reta: Recursively thinking ahead to improve the strategic reasoning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2232–2246.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. 2024c. Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. *arXiv preprint arXiv:2402.12348*.

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*.