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

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

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

Background

Large language models (LLMs) are increasingly being employed for human-in-the-loop decision making and human-AI teams. However, LLMs are proven to generate information that is not grounded in reality and deviates from user instructions. Uncertainty quantification (UQ) can provide valuable insights into the reliability of model responses, facilitating risk control and assessment. Split conformal prediction (CP) is a distribution-free and model-agnostic approach to UQ, which transforms any heuristic notion of uncertainty into a statistically rigorous one by calibrating prediction sets.

Motivation

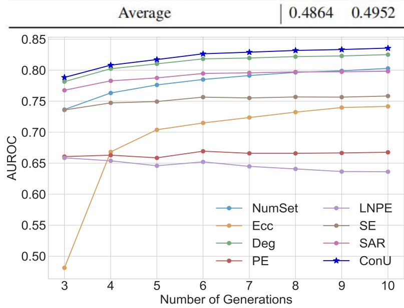
Prior studies adapt split CP to multiple-choice question-answering (MCQA) tasks, where the acceptable response is selected from a fixed set of options, limiting its deployment in real-world open-ended applications. Additionally, in open-ended context, existing work either relies on the model logits or fails to achieve strict risk control at various user-specified error rates.

Our Work

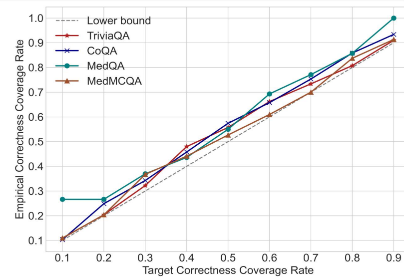
- We propose a novel black-box uncertainty measure, termed as *ConU*, based on self-consistency theory.
- We devise a conformal uncertainty criterion, by connecting the nonconformity score with the uncertainty condition, estimated by *ConU*, aligned with correctness.

ConU generally outperforms 8 baseline methods in distinguishing between correct and incorrect responses.

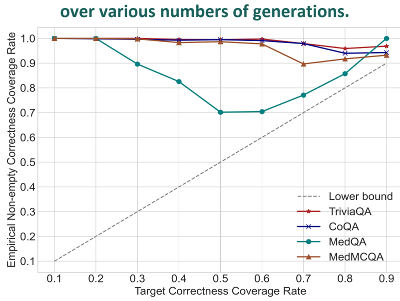
Dataset	LLMs	White-box				Black-box				
		PE	*LNPE*	*SE*	*SAR*	*LS*	*NumSet*	*Ecc*	*Deg*	*ConU*
TriviaQA	LLaMA-2-7B-Chat	0.6587	0.6459	0.7495	0.7876	0.5571	0.7763	0.7839	0.8103	**0.8198**
	Mistral-7B-Instruct-v0.3	0.6620	0.5968	0.7845	0.8306	0.5969	0.8491	0.8596	0.8596	**0.8671**
	LLaMA-3-8B-Instruct	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**
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**



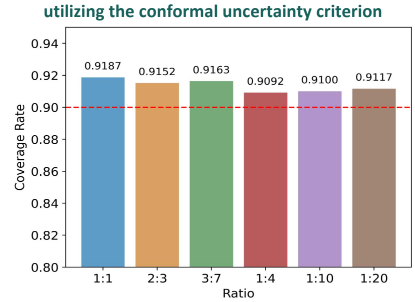
ConU consistently outperforms 7 baselines over various numbers of generations.



We bound the correctness coverage rate utilizing the conformal uncertainty criterion



Empirical correctness coverage rate on non-empty prediction sets



Empirical correctness coverage rate at various splitting ratios

Step one: ConU

We sample M generations to each query, denoted as $\{\hat{y}_m\}_{m=1}^{M}$, and employ the most frequent response as the evaluation object, denoted as $\hat{y}_{mst}$, to estimate the model uncertainty based on self-consistency theory.

Specifically, we conduct semantic clustering and obtain K non-repeated clusters, where semantically equivalent generations share the same frequency score, equaling to the cluster size.

Then, we propose a heuristic uncertainty measure, termed *ConU*, which combines the frequency score of $\hat{y}_{mst}$ with the semantic diversity between it and response samples sharing other semantics.

$$\mathcal{U}(\{\hat{y}_m\}_{m=1}^{M}|x)$$
$$= 1 - \lambda \cdot \mathcal{F}(\hat{y}_{mst}) - (1-\lambda) \cdot \frac{1}{K}\sum_{k=1}^{K} \mathcal{S}(\hat{y}_{mst}, \hat{y}_k)\mathcal{F}(\hat{y}_k)$$

Step two: Conformal Uncertainty Criterion

We devise the nonconformity score (NS) in CP by linking it with the uncertainty condition strictly aligned with the acceptable responses, which leads to robust correctness coverage guarantees in i.i.d. test data points.

$$r(x_i, y_i^*) = \mathcal{U}\left(\mathrm{argmax}_{\hat{y}_m}\, \mathcal{S}(\hat{y}_m, y_i^*)\mathbf{1}\{\hat{y}_m \Leftrightarrow y_i^*\}\right)$$

Finally, we construct prediction sets based on the $1-\alpha$ quantile of NSs, denoted as $\hat{q}$.

$$\mathcal{P}(x_{test}) = \{\hat{y}_m : r(x_{test}, \hat{y}_m) \le \hat{q}\},$$

and $\mathbb{P}\left(y_{test}^* \in \mathcal{P}(x_{test})\right) \ge 1-\alpha$.

EMNLP 2024 Findings