SConU: Selective Conformal Uncertainty in Large Language Models

Zhiyuan Wang[†1], Qingni Wang[†1], Yue Zhang[2], Tianlong Chen[3], Xiaofeng Zhu[1], Xiaoshuang Shi[*1], Kaidi Xu[*2]

[1] University of Electronic Science and Technology of China
[2] Drexel University
[3] University of North Carolina at Chapel Hill

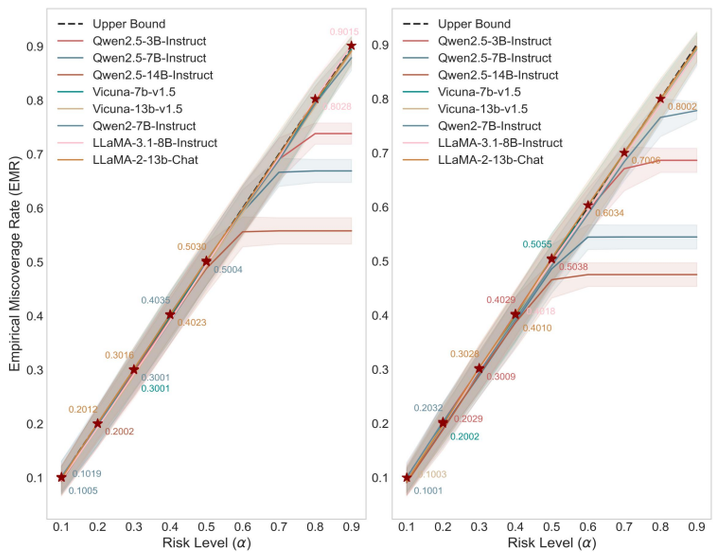
Abstract

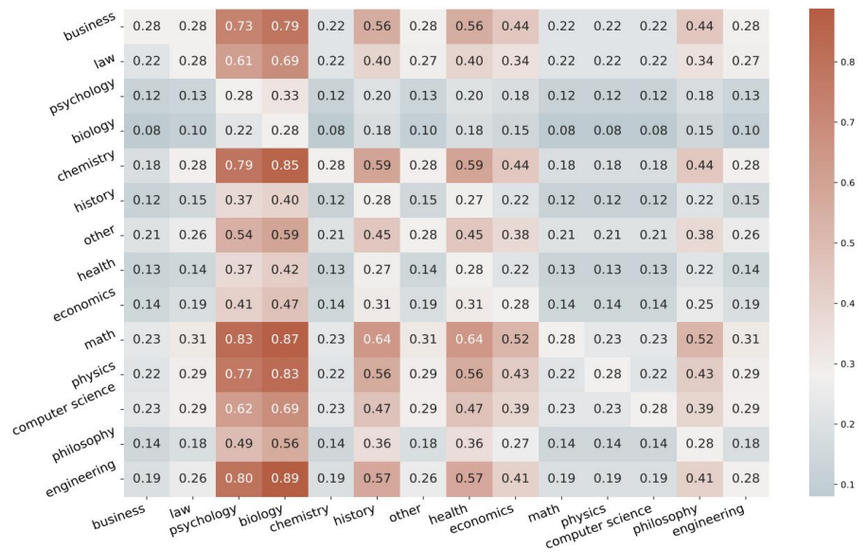
To ensure the reliable deployment of large language models, we introduce Selective Conformal Uncertainty (**SConU**)—a novel framework that combines conformal prediction with significance testing. By computing two **conformal p-values**, SConU identifies **out-of-distribution uncertainty patterns** and controls miscoverage at a user-specified risk level across both single-domain and interdisciplinary scenarios. This approach improves the reliability and efficiency of predictions, particularly in high-stakes QA tasks, and approximates **conditional coverage** across diverse domains.

Motivation and Contribution

A key assumption underlying conformal uncertainty frameworks is the **exchangeability** between the test sample and calibration data. However, in practical QA tasks, this assumption is often violated due to the **conditional nature of language generation**, making it difficult to verify or enforce. Empirically, we observe **significant miscoverage** even in **single-domain** scenarios, and more severe miscalibration in **interdisciplinary** contexts.

- We propose **selective conformal uncertainty**, which for the first time implements **significance tests** to filter out uncertainty data outliers that violate the **exchangeability** precondition at a specific risk level.
- We **maintain the integrity of the calibration set** and derive the **minimum risk level**.





Method

We have to **maintain the calibration set** to cover data distributions across various domains comprehensively.

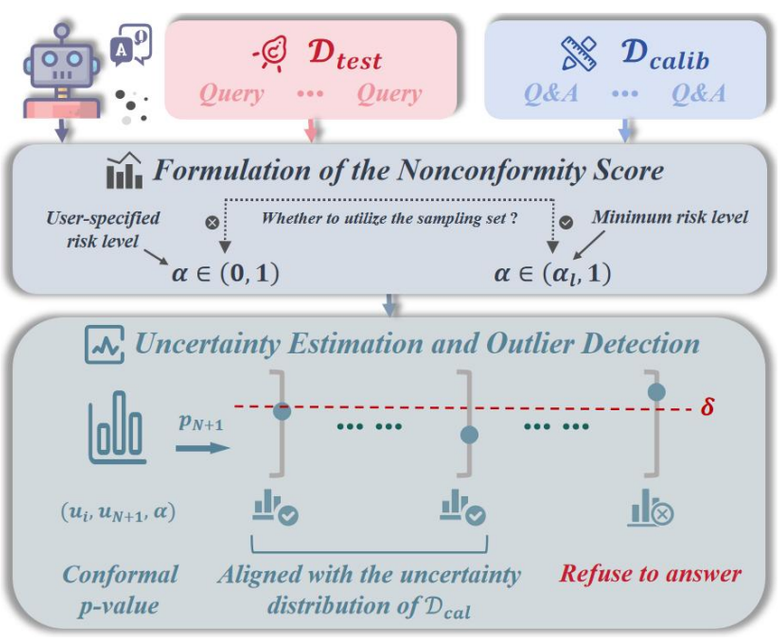
$$m_i = \inf\left\{ M_i : \forall M_i' \geq M_i, y_i^* \in \left\{ y_j^{(i)} \right\}_{j=1}^{M_i'} \right\}$$

We first compute the **minimum risk level** if we utilize the sampling set when formulating the nonconformity score.

$$L_N(1) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left\{ y_i^* \notin \left\{ y_j^{(i)} \right\}_{j=1}^{M} \right\} \qquad \alpha_l = N L_N(1) / (N+1)$$

Given that existing ConU methods rely on **heuristic uncertainty scores** as **nonconformity** measures—while the **exchangeability** assumption fundamentally pertains to the distribution of these scores across samples—we adopt ideas from **conformal outlier detection** and **permutation testing** to gather statistical evidence for non-exchangeable data sequences via **hypothesis testing**. Specifically, we design two types of **conformal p-values** to identify and exclude out-of-distribution samples, thereby ensuring that the remaining subset adheres to the desired coverage guarantees.

$$p_{N+1} = \frac{1 + \sum_{i=1}^{N} \mathbf{1}\{u_i \geq u_{N+1}\}}{N+1} \qquad p'_{N+1} = \frac{1 + \sum_{i=1}^{N} \mathbf{1}\{u_i \geq u_{N+1}, y_i^* \in E(x_i, \mathcal{D}_{cal}, \alpha)\}}{N+1}$$

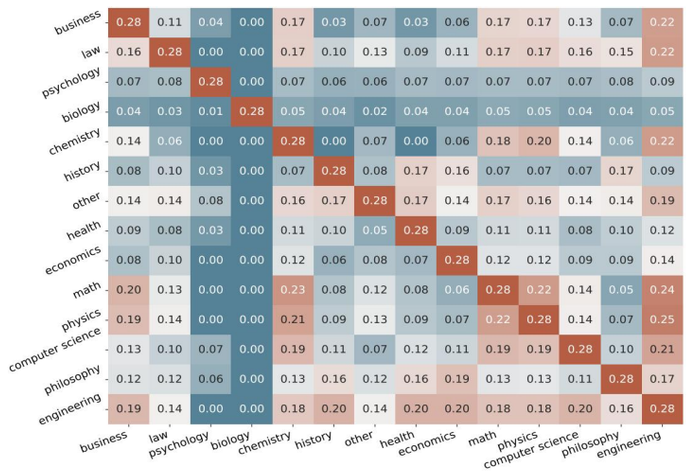


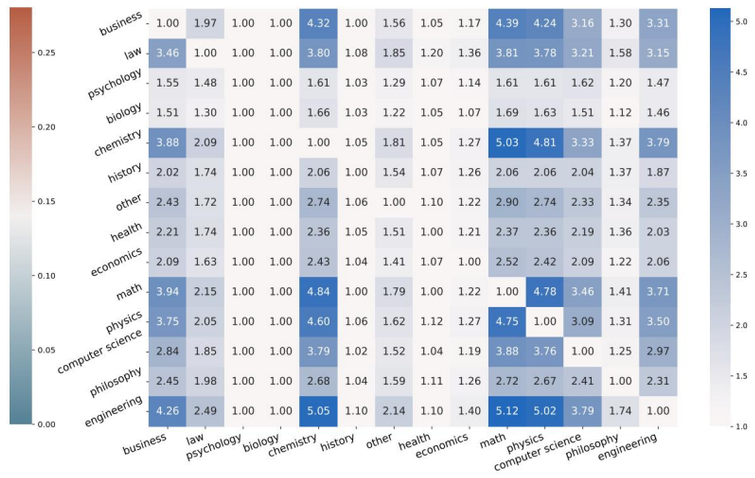
Experiments

➢ Maintenance of the calibration set.

Dataset	TriviaQA (open-ended)			MedMCQA (closed-ended)		
LLMs / β	0.1	0.2	0.3	0.1	0.2	0.3
LLaMA-3.2-3B-Instruct	0.0884 ± 0.0149	0.1767 ± 0.0109	0.2725 ± 0.0194	0.0896 ± 0.0078	0.1823 ± 0.0084	0.2423 ± 0.0072
OpenChat-3.5	0.0848 ± 0.0179	0.1551 ± 0.0391	0.1997 ± 0.0090	0.0911 ± 0.0119	0.1785 ± 0.0265	0.2676 ± 0.0074
LLaMA-3.1B-Instruct	0.0869 ± 0.0060	0.1770 ± 0.0378	0.1965 ± 0.0086	0.0861 ± 0.0067	0.1697 ± 0.0331	0.2771 ± 0.0078
Qwen2.5-14B-Instruct	0.0835 ± 0.0201	0.1731 ± 0.0075	0.1731 ± 0.0075	0.0815 ± 0.0047	0.0815 ± 0.0047	0.0815 ± 0.0047

➢ Cross-domain calibration.





➢ Single-domain calibration.

Disciplinary	Metric	OD	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
			Qwen-2-7B-Instruct Model.							
Health	Mean	✗	0.1019	0.1977	0.3001	0.4035	0.5004	0.5964	0.6888	0.7938
		✓	0.0938	0.1943	0.2972	0.3957	0.4937	0.5915	0.6819	0.7876
	Std ↓	✗	0.0285	0.0372	0.0420	0.0434	0.0424	0.0441	0.0420	0.0323
		✓	**0.0283**	**0.0362**	0.0423	**0.0425**	**0.0358**	**0.0429**	**0.0384**	**0.0323**
	Median	✗	0.1080	0.1960	0.2960	0.4120	0.5080	0.5960	0.6760	0.7880
		✓	0.0960	0.1920	0.2920	0.3960	0.4920	0.5920	0.6800	0.7960
Economics	Mean	✗	0.1001	0.2032	0.2951	0.3928	0.4916	0.5871	0.6838	0.7658
		✓	0.0965	0.1951	0.2950	0.3928	0.4877	0.5853	0.6820	0.7630
	Std ↓	✗	0.0279	0.0338	0.0367	0.0408	0.0384	0.0366	0.0347	0.0352
		✓	**0.0210**	**0.0281**	**0.0275**	**0.0395**	**0.0294**	**0.0253**	**0.0294**	**0.0272**
	Median	✗	0.1040	0.2080	0.2880	0.3920	0.4960	0.5920	0.6880	0.7640
		✓	0.0960	0.1960	0.2920	0.3960	0.4880	0.5880	0.6840	0.7680

➢ Conditional coverage.

w_l (Logit)	w_f (Frequency)	M (Sampling)	OD	Size = 1	Size = 2	Size = 3	SSM ↓
			Split ratio is fix at 0.5 and α is set to 0.34 ($\alpha_l = 0.3342$).				
1	0	10	✗	0.3428 ± 0.0151	0.2800 ± 0.0277	0.1056 ± 0.1860	0.3579
1	0	10	✓	0.3060 ± 0.0054	0.0348 ± 0.0047	0	0.3114
0.5	0.5	10	✗	0.3428 ± 0.0144	0.2971 ± 0.0240	0.1487 ± 0.1594	0.3572
0	1	10	✗	0.3391 ± 0.0149	0.2874 ± 0.0251	0.2177 ± 0.1027	0.3540
0	1	10	✓	0.3025 ± 0.0067	0.2795 ± 0.0766	0	**0.3092**