# SConU: Selective Conformal Uncertainty in Large Language Models

Anonymous ACL submission


## Abstract

As large language models are increasingly utilized in real-world applications, guarantees of task-specific performance are essential for their reliable deployment. Recent studies have introduced various conformal uncertainty criteria grounded in split conformal prediction, which offer user-specified correctness coverage. However, existing frameworks often fail to identify uncertainty data outliers that violate the exchangeability assumption, leading to unbounded miscoverage rates and unactionable prediction sets. In this paper, we propose a novel approach termed Selective Conformal Uncertainty (SConU), which, for the first time, implements significance tests, by developing two conformal p-values that are instrumental in determining whether a given sample deviates from the uncertainty distribution of the calibration set at a specific manageable risk level. Our approach not only facilitates rigorous management of miscoverage rates across both single-domain and interdisciplinary contexts, but also enhances the efficiency of predictions. Furthermore, we comprehensively analyze the components of the conformal procedures, aiming to approximate conditional coverage, particularly in high-stakes question-answering tasks.


## 1 Introduction

Large language models (LLMs) have been increasingly deployed in real-world natural language generation (NLG) tasks, including question-answering (QA) (Duan et al., 2024; Wang et al., 2025). However, their generations often reveal deficiencies in trustworthiness and robustness (Yona et al., 2024; Farquhar et al., 2024; Kaur et al., 2024). These issues have sparked significant interest in developing guarantees for task-specific performance metrics, such as correctness miscoverage rate (Wang et al., 2024d; Quach et al., 2024), factuality (Mohri and Hashimoto, 2024; Cherian et al., 2024), and disparities in generation quality across diverse user populations (Deng et al., 2023; Zollo et al., 2024).

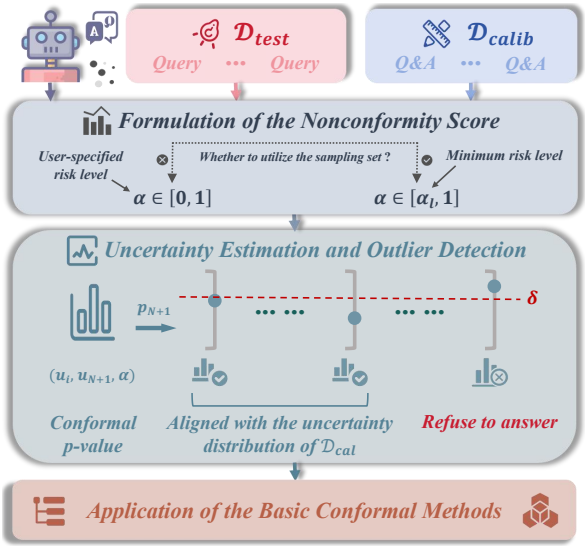



Figure 1: Pipeline of the SConU framework. We achieve rigorous coverage of correct generations on test samples at various user-specified risk levels based on the calibration set after automatic outliers detection.

Split conformal prediction (SCP) (Papadopoulos et al., 2002; Bates et al., 2021; Angelopoulos and Bates, 2021) can offer distribution-free and model-agnostic coverage guarantees to new samples based on a calibration set. Recent studies have introduced various criteria of conformal uncertainty (ConU), which allow user-specified risk levels (e.g., $\alpha$) for the coverage of acceptable responses in practical NLG tasks, by correlating the nonconformity score (NS) with the uncertainty state of ground-truth answers (Quach et al., 2024; Su et al., 2024; Wang et al., 2024d,b; Kaur et al., 2024). However, these frameworks are vulnerable to uncertainty data outliers and sensitive to internal units, such as the uncertainty notion and split ratio, compromising their statistical rigor and operational efficiency (Cresswell et al., 2024b; Plassier et al., 2024).

To conduct comprehensive research, we first revisit a crucial precondition for prior frameworks: the combined sequence of the given test QA sample and all calibration data points should be ex-

changeable (Kumar et al., 2023). In practical QA tasks, however, this condition is hard to characterize and verify specifically, often being violated due to the conditional nature of language generation approaches (Ulmer et al., 2024). More concerning, we observe significant coverage anomalies within single-domain contexts, as illustrated in Figure 2a, which contradict the assumptions made in previous studies (Ye et al., 2024; Quach et al., 2024; Su et al., 2024; Wang et al., 2024d; Kaur et al., 2024; Wang et al., 2024b). Furthermore, miscalibration issues become even more pronounced in interdisciplinary scenarios (Kumar et al., 2023), as demonstrated in Figure 2b. The conceptual and fragile nature of exchangeability renders the prediction sets produced by existing ConU frameworks unreliable and less actionable (Cresswell et al., 2024a).

Within prior ConU frameworks, the NSs are derived from various uncertainty notions linked with reliable generations and then utilized to select responses by a user-specified quantile. As supported by Figure 2, our key insight is that employing different models will affect how well the uncertainty distribution of the calibration set covers test QA samples at a specific risk level (Lin et al., 2024; Ye et al., 2024), thus determining the exchangeability among the NSs. For instance, if the deployed model excels in health but struggles with math, the NSs from the health dataset will significantly differ from (lower than) those from the math dataset, thus leading to miscalibration, while a powerful proprietary model with comprehensive knowledge across both domains can yield an approximate uncertainty distribution. Furthermore, ConU methods manually remove calibration samples that fail to contain acceptable answers within the sampling space (Su et al., 2024; Kaur et al., 2024; Wang et al., 2024d,b), which constrains the quantity of test QA samples that the calibration set can handle, as demonstrated in Section 3.2. At this point, our goal is to derive the minimum risk level manageable by the original calibration set, and then eliminate uncertainty data outliers undermining exchangeability. Subsequently, the remaining test samples are expected to allow for user-specified marginal coverage.

Inspired by prior work on outlier detection (OD) and permutation test (Vovk et al., 2003; Angelopoulos and Bates, 2021; Guan and Tibshirani, 2022; Bates et al., 2023), we propose **s**elective **con**formal **u**ncertainty (SConU), which gathers statistical evidence for nonexchangeable data sequences via hypothesis testing. Specifically, we construct a conformal p-value (Jin and Candès, 2023; Angelopoulos et al., 2024a; Gui et al., 2024) for each test data to identify whether its uncertainty state significantly deviates from the calibration data distribution, using it as a baseline for exchangeability assessment. Furthermore, recognizing that uncertainty data anomalies in the calibration set compromise their reference value and statistical rigor, we provide an optimized version by incorporating the prediction status of each calibration data point at a specific risk level into the counting criterion of the conformal p-value. After filtering out uncertainty data outliers within the test set, we achieve rigorous management of the miscoverage rates in both single-domain and cross-domain QA datasets.

Additionally, practical NLG applications focus on conditional coverage for a particular input. However, this property is infeasible in most NLG cases (Angelopoulos and Bates, 2021; Plassier et al., 2024; Angelopoulos et al., 2024a). In this paper, we investigate the impact of the exchangeability condition, split ratio, and uncertainty measurements on conditional performance, aiming to approximate conditional coverage in high-stakes QA scenarios. Finally, we disclose significant semantic redundancy within prediction sets in human-in-the-loop QA applications (Cresswell et al., 2024b).

Our contributions can be summarized as follows:

- We propose **s**elective **con**formal **u**ncertainty (SConU), 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 calibration set and derive the minimum manageable risk level after deploying the language model.
- We explore internal components of SConU to enhance conditional performance and operational efficiency of the prediction sets.

## 2 Related Work

**Split Conformal Prediction.** SCP guarantees ground-truth coverage on fresh test samples based on a calibration set (Papadopoulos et al., 2002; Angelopoulos and Bates, 2021; Angelopoulos et al., 2024b,a). We briefly outline the conformal procedures of the SCP framework in Appendix A. Despite the statistical rigor, SCP assumes the NSs of all the $N$ calibration data points and the given test sample to be exchangeable (Tibshirani et al.,

2019; Bates et al., 2021; Barber et al., 2023; Farinhas et al., 2024). Formally, the sequence of data points $Z_1, Z_2, \cdots, Z_N, Z_{N+1}$ is considered exchangeable if, for any permutation $\pi$, the sequence $\left(Z_{\pi(1)}, Z_{\pi(2)}, \ldots, Z_{\pi(N)}, Z_{\pi(N+1)}\right)$ has the same joint distribution as $(Z_1, Z_2, \ldots, Z_N, Z_{N+1})$. Intuitively, this condition is hard to represent and verify concretely in NLG tasks (Campos et al., 2024).

**Conformal Uncertainty in QA Tasks.** Recently, researchers have attempted to apply SCP to LLMs for reliable language generation. In white-box settings, several studies (Kumar et al., 2023; Ye et al., 2024; Kostumov et al., 2024; Kaur et al., 2024; Quach et al., 2024; Angelopoulos et al., 2024b) develop ConU frameworks for multiple-choice query-answering (MCQA) and open-ended QA tasks by correlating the NS with a certain uncertainty notion of reliable responses (e.g., normalized logit-based probability of each option). Meanwhile, researchers also establish criteria in black-box scenarios (Wang et al., 2024d; Su et al., 2024; Wang et al., 2024b) based on self-consistency. Our work SConU applies to both settings and retains existing frameworks: We do not process calibration samples manually but instead derive the minimum risk level, which allows for handling more QA samples from diverse subjects. Then, we perform the conformal p-value to eliminate uncertainty data outliers violating the exchangeability precondition, and apply ConU frameworks based on the type of problems.

Additionally, real-world QA applications often focus on conditional coverage over a particular input (Gibbs et al., 2023; Ding et al., 2023; Kim et al., 2024; Cresswell et al., 2024a), while in the most practical NLG case, this property is impossible to achieve (Vovk, 2012; Plassier et al., 2024). This paper examines internal factors of SConU, such as the reliability measurements in the formulation of the NS, seeking to approximate conditional coverage across various set sizes (Angelopoulos et al., 2024a; Su et al., 2024; Wang et al., 2024b).

# 3 Method

## 3.1 Preliminaries

Formally, we have a held-out set of $N$ calibration data points, $\mathcal{D}_{cal} = \{(x_i, y_i^*)\}_{i=1}^N$, where $x_i$ and $y_i^*$ denote the $i$-th question and ground-truth answer, respectively. For each data point, we sample multiple (e.g., $M$) responses from the output space of the language model to construct a candidate set for the corresponding question, denoted as $\left\{y_j^{(i)}\right\}_{j=1}^M$. We can calculate the reliability score of each generation or semantic cluster utilizing various uncertainty measurements within the candidate set (Su et al., 2024; Wang et al., 2024d; Kaur et al., 2024). For instance, we can express the confidence score of each option in MCQA task as $w_l \cdot F_l\left(y_j^{(i)}\right) + w_f \cdot F_f\left(y_j^{(i)}\right)$, where $F_l\left(y_j^{(i)}\right)$ represents the probability derived from model logit, $F_f\left(y_j^{(i)}\right)$ denotes the frequency score of $y_j^{(i)}$ within the candidate set, and $w_l$ and $w_f$ are the respective weights assigned to each score. Then, the NS of each MCQA sample is $1 - w_l \cdot F_l\left(y_i^*\right) - w_f \cdot F_f\left(y_i^*\right)$ $(w_l + w_f = 1)$.

Due to the randomness of sampling and potential limitations in model capability, we may not always obtain an acceptable response that aligns with the ground-truth answer by sampling $M$ times for each QA sample. Unlike prior work (Wang et al., 2024d; Kaur et al., 2024; Wang et al., 2024b), we do not demand that samples employed as the calibration data must encompass acceptable responses within their candidate sets. On one hand, given that SCP is model-agnostic, we cannot guarantee that all employed language models in practical applications will be capable of addressing the same questions. Furthermore, we aim for the calibration set to cover data distributions across various domains comprehensively. While the lower bound of the error rate that the calibration set can control is constrained at this point, we can accommodate a greater volume of test QA samples by easing the risk level of $\alpha$.

## 3.2 Selective Conformal Uncertainty

Inspired by prior research (Pitman, 1937; Jin and Candès, 2023; Bates et al., 2023; Gui et al., 2024; Angelopoulos et al., 2024a), we collect statistical evidence for nonexchangeable sequences of NSs arising from uncertainty data outliers via hypothesis testing. Specifically, we define the null hypothesis $\mathcal{H}_0$ for the test data point $x_{N+1}$ with the significance level of $\delta$ as follows: $\{(x_i, y_i^*)\}_{i=1}^N$ can serve as the calibration set for $x_{N+1}$ with coverage guarantees. Rejecting $\mathcal{H}_0$ indicates sufficient evidence of the prediction set with an unbounded miscoverage rate when tackling $x_{N+1}$ based on $\{(x_i, y_i^*)\}_{i=1}^N$. To this end, we construct a finite-sample valid conformal p-value associating $\mathcal{H}_0$ as

$$p_{N+1} = \frac{1 + \sum_{i=1}^N \mathbf{1}\left\{u_i \geq u_{N+1}\right\}}{N+1} \quad (1)$$

In the formulation, $u_i$ indicates the uncertainty of the language model addressing the $i$-th question $x_i$, measured by an uncertainty notion $U$, and we utilize the predictive entropy (PE) (Kadavath et al., 2022; Duan et al., 2024; Wang et al., 2025). Note that the uncertainty corresponds to the output distribution of a particular QA sample, while the NS reflects the model's uncertainty regarding a specific generation, representing the disagreement between the current response and the query.

As mentioned, we consider that uncertainty data anomalies may present in the calibration set and compromise statistical rigor. To examine the reference quality of each calibration data point at a specific risk level, we refine the conformal p-value:

$$p'_{N+1} = \frac{1 + \sum_{i=1}^{N} \mathbf{1}\left\{u_i \geq u_{N+1}, y_i^* \in E\left(x_i, \mathcal{D}_{cal}, \alpha\right)\right\}}{N+1}, \tag{2}$$

where $y_i^* \in E\left(x_i, \mathcal{D}_{cal}, \alpha\right)$ determines whether the prediction set established for $x_i$, calibrated by all data points in $\mathcal{D}_{cal}$ except for $x_i$, contains $y_i^*$ at a risk level of $\alpha$. If not, we intuitively consider that the model may encounter hallucination issues when processing $x_i$ (Kuhn et al., 2023; Farquhar et al., 2024), or that the uncertainty of its output distribution is abnormally high, which results in high NSs of its reliable generations and miscoverage. At this point, $u_i \geq u_{N+1}$ lacks statistical validity at the risk level of $\alpha$, and $\mathbf{1}\{\cdot\}$ does not count.

For simplicity, we refer to the conformal procedure employing two conformal p-values as SConU and SConU-Pro in the following text. We demonstrate that the two conformal p-values adhere to the statistical definition of p-values in Appendix C, and present a more rigorous framework to detect when test points do not come from the same distribution.

**Maintenance of the calibration set.** As mentioned, we do not remove calibration data that fail to cover acceptable responses within their candidate sets. In this section, we demonstrate the practical significance by defining the minimum sampling size of each calibration QA sample as

$$m_i = \inf\left\{M_i : \forall M_i' \geq M_i, y_i^* \in \left\{y_j^{(i)}\right\}_{j=1}^{M_i}\right\}, \tag{3}$$

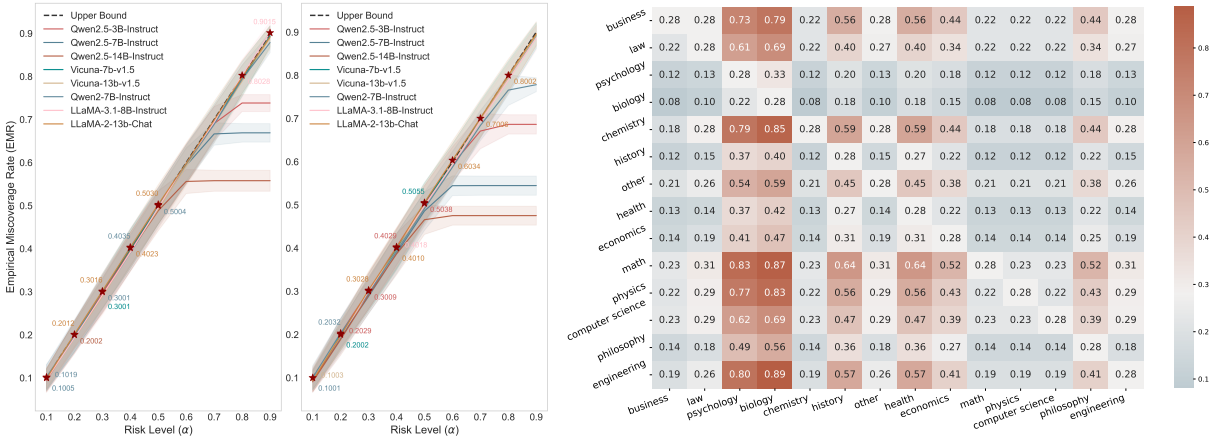
which ensures that there is at least one correct answer in the $i$-th candidate set of size $m_i$. Then, we sort the $N$ minimum sampling sizes and calculate their $\frac{\lceil (1-\beta)(1+N) \rceil}{N}$ quantile: $\hat{m} = m_{\lceil (1-\beta)(1+N) \rceil}$, where $\beta$ represents the error rate (similar to $\alpha$). If the test sample is exchangeable with $N$ calibration data points, we have $\mathbb{P}\left(m_{N+1} \leq m_i\right) = \frac{i}{N+1}$. We then set the sampling size of the test QA sample to $\hat{m}$ and obtain

$$\begin{aligned} &\mathbb{P}\left(y_{N+1}^* \in \left\{y_j^{(N+1)}\right\}_{j=1}^{\hat{m}}\right) = \mathbb{P}\left(m_{N+1} \leq \hat{m}\right) \\ &= \lceil (1-\beta)(1+N) \rceil / (N+1) \geq 1-\beta \end{aligned}. \tag{4}$$

Following the requirement of previous research, where at least one correct answer exists in the candidate set of fixed size $M$ for each calibration data, we have $M \geq \max\{m_i\}_{i=1}^{N}$ and $\beta \to 0$. At this point, $y_{N+1}^* \in \left\{y_j^{(N+1)}\right\}_{j=1}^{M}$ is a certain event, which is infeasible in practical NLG tasks. Additionally, removing calibration samples will narrow the uncertainty distribution of the calibration set, which diminishes its adaptability to new test QA samples. Therefore, we explore the minimum risk level controlled by the original calibration set.

**Minimum risk level.** Building on prior research (Angelopoulos et al., 2024b; Farinhas et al., 2024), we post-process the candidate set of each calibration data point into a set of reliable responses with sufficiently high confidence scores, $\mathcal{C}_\lambda\left(x_i\right) = \left\{y_j^{(i)} : F\left(y_j^{(i)}\right) \geq 1-\lambda\right\}$ ($\lambda \in [0,1]$), where $F(\cdot)$ can be any measurement that reflect the trustworthiness of each sampled response. Then, we calculate the loss of miscoverage, $l\left(\mathcal{C}_\lambda\left(x_i\right), y_i^*\right) = \mathbf{1}\left\{y_i^* \notin \mathcal{C}_\lambda\left(x_i\right)\right\}$, abbreviated as $l_i(\lambda)$, and set $L_N(\lambda) = \frac{1}{N}\sum_{i=1}^{N} l_i(\lambda)$. Suppose $l_{N+1}(\lambda)$ follows $\text{Uniform}\left(\left\{l_1(\lambda), \cdots, l_{N+1}(\lambda)\right\}\right)$ by exchangeability, we have $\mathbb{E}\left[l_{N+1}(\lambda)\right] = \frac{1}{N+1}\sum_{i=1}^{N+1} l_i(\lambda) = \frac{N L_N(\lambda) + l_{N+1}(\lambda)}{N+1}$. Obviously, $L_N(\lambda)$ is non-increasing in $\lambda$. Then, we set $\lambda$ to its upper bound (i.e., 1) and obtain the minimum value, $L_N(1)$.

When $\lambda$ is set to 1, $\mathcal{C}_\lambda\left(x_i\right) = \left\{y_j^{(i)}\right\}_{j=1}^{M}$, and at this point, the problem simplifies to calculating the proportion of candidate sets in the calibration set that do not contain an acceptable response: $L_N(1) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left\{y_i^* \notin \left\{y_j^{(i)}\right\}_{j=1}^{M}\right\}$. Since $\mathbb{E}\left[l_{N+1}(\lambda)\right]$ should be controlled by a user-




(a) Single-domain Miscalibration.

(b) Cross-domain Miscalibration.

Figure 2: **(a)** Empirical miscoverage rate (EMR) at various risk levels on the MMLU-Pro dataset utilizing 8 LLMs. Results on the left are from the Health discipline, while results on the right are from the Economics discipline. Solid lines give the mean over 100 trials and shaded regions show $+/-$ the standard deviation (std). We set the split ratio between the calibration and test set to 0.5 for all trials. The ★ indicates that even the mean miscoverage rate at the corresponding risk level is higher than the upper bound, and the shaded regions exceeding the upper bound reflect significant data point anomalies. **(b)** Significant violation in the management of EMR when we use data points from different disciplines for the calibration set and the test set within the MMLU-Pro dataset employing the LLaMA-3.1-8B-Instruct model at the risk level of 0.28. Note that we calculate the minimum reliable risk level on each subject based on Eq. (5) and set $\alpha$ to the maximum. All data on the diagonal is manually set to equal $\alpha$.

specified risk level of $\alpha$ (i.e., $\mathbb{E}\left[l_{N+1}(\lambda)\right] \leq \alpha$), and $l_{N+1}(\lambda) \in \{0, 1\}$, we obtain $\mathbb{E}\left[l_{N+1}(\lambda)\right] \geq \frac{NL_N(1)}{N+1}$, and at this point,

$$\alpha_l = NL_N(1) / (N+1) \tag{5}$$

Finally, for any risk level of $\alpha \geq \alpha_l$, we can rigorously manage the correctness miscoverage rate leveraging the given calibration set.

**Workflow of SConU.** As illustrated in Figure 1, after deploying the LLM, we first calculate the minimum risk level if we utilize the candidate set when formulating NS. Then we conduct significance tests to identify whether the given test QA sample aligns with the uncertainty distribution of the calibration set at a user-specified risk level. A low conformal p-value suggests a violation of the exchangeability precondition, and we decline to respond.

## 4 Experiments

### 4.1 Experimental Settings

**Datasets.** We utilize 3 closed-ended QA datasets: MMLU (Hendrycks et al., 2021) for multitask language understanding, more challenging MMLU-Pro (Wang et al., 2024c), and MedMCQA (Pal et al., 2022) for real-world medical entrance exam, and 2 open-domain datasets: TriviaQA (Joshi et al., 2017) for closed-book QA and CoQA (Reddy et al., 2019) for open-book conversational QA. More details are presented in Appendix B.2.

**Metrics.** We utilize the Empirical Miscoverage Rate (EMR) to assess whether conformal methods produce prediction sets that meet statistical guarantees (Wang et al., 2024b; Quach et al., 2024) after outlier elimination. For conditional coverage, we apply the Size-stratified Miscoverage Rate (SMR) that evaluates error rates across various set sizes (Angelopoulos and Bates, 2021; Kumar et al., 2023; Su et al., 2024). We also explore the operational efficiency through the Average Prediction Set Size (APSS) on the test set (Wang et al., 2024d,b; Su et al., 2024; Angelopoulos et al., 2024a).

Our utilized LLMs and additional experimental settings are presented in Appendix B.

### 4.2 Empirical Results

**Marginal Coverage.** As illustrated in Figure 2a, we apply ConU to single-domain datasets and observe that the mean EMR results exceed the user-specified risk levels for some LLMs (e.g., Qwen-2-7B-Instruct). Moreover, the shaded area significantly surpassing the dashed line indicates substantial issues unbounded EMR in 100 trials. Note that we employ the typical conformal framework (Kumar et al., 2023; Ye et al., 2024; Kostumov et al., 2024; Campos et al., 2024) for MCQA tasks, de-

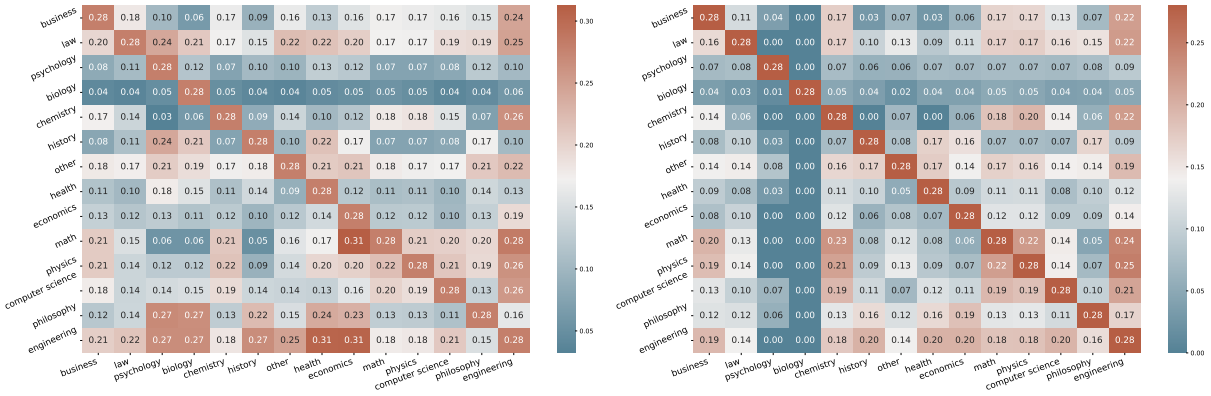
Table 1: The mean, std, and median of EMR results obtained from 100 trials on the MMLU-Pro dataset. **Note that** the mean and median metrics only need to be below the corresponding risk level, and they are not required to be as low as possible. ✗ indicates the employment of the basic ConU framework for MCQA, and ✓ represents utilizing our SConU criterion, eliminating uncertainty data outliers within the test set. Red indicates out of risk level.

| Disciplinary | Metric | OD | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Qwen-2-7B-Instruct Model. | | | | | | | | | | | |
| Health | Mean | ✗ | 0.1019 | 0.1977 | 0.3001 | 0.4035 | 0.5004 | 0.5964 | 0.6888 | 0.7938 | 0.8788 |
| | | ✓ | 0.0938 | 0.1943 | 0.2972 | 0.3957 | 0.4937 | 0.5915 | 0.6819 | 0.7876 | 0.8754 |
| | Std ↓ | ✗ | 0.0285 | 0.0372 | 0.0420 | 0.0434 | 0.0424 | 0.0441 | 0.0420 | 0.0323 | 0.0232 |
| | | ✓ | **0.0283** | **0.0362** | 0.0423 | **0.0425** | **0.0358** | **0.0429** | **0.0384** | **0.0323** | **0.0227** |
| | Median | ✗ | 0.1080 | 0.1960 | 0.2960 | 0.4120 | 0.5080 | 0.5960 | 0.6760 | 0.7880 | 0.8760 |
| | | ✓ | 0.0960 | 0.1920 | 0.2920 | 0.3960 | 0.4920 | 0.5920 | 0.6800 | 0.7960 | 0.8800 |
| Economics | Mean | ✗ | 0.1001 | 0.2032 | 0.2951 | 0.3928 | 0.4916 | 0.5871 | 0.6838 | 0.7658 | 0.8783 |
| | | ✓ | 0.0965 | 0.1951 | 0.2950 | 0.3928 | 0.4877 | 0.5853 | 0.6820 | 0.7630 | 0.8767 |
| | Std ↓ | ✗ | 0.0279 | 0.0338 | 0.0367 | 0.0408 | 0.0384 | 0.0366 | 0.0347 | 0.0352 | 0.0161 |
| | | ✓ | **0.0210** | **0.0281** | **0.0275** | **0.0395** | **0.0294** | **0.0253** | **0.0294** | **0.0272** | 0.0226 |
| | Median | ✗ | 0.1040 | 0.2080 | 0.2880 | 0.3920 | 0.4960 | 0.5920 | 0.6880 | 0.7640 | 0.8760 |
| | | ✓ | 0.0960 | 0.1960 | 0.2920 | 0.3960 | 0.4880 | 0.5880 | 0.6840 | 0.7680 | 0.8720 |
| LLaMA-3.1-8B-Instruct Model. | | | | | | | | | | | |
| Health | Mean | ✗ | 0.0961 | 0.1933 | 0.2922 | 0.3912 | 0.4957 | 0.5988 | 0.6936 | 0.8028 | 0.9015 |
| | | ✓ | 0.0975 | 0.1925 | 0.2926 | 0.3935 | 0.4978 | 0.5966 | 0.6883 | 0.7913 | 0.8941 |
| | Std ↓ | ✗ | 0.0273 | 0.0364 | 0.0459 | 0.0447 | 0.0481 | 0.0459 | 0.0404 | 0.0362 | 0.0257 |
| | | ✓ | **0.0214** | **0.0300** | **0.0412** | **0.0431** | **0.0457** | **0.0420** | 0.0426 | **0.0357** | **0.0241** |
| | Median | ✗ | 0.0960 | 0.1920 | 0.2960 | 0.3920 | 0.4960 | 0.5880 | 0.6920 | 0.7960 | 0.9040 |
| | | ✓ | 0.0960 | 0.1960 | 0.3000 | 0.3880 | 0.4840 | 0.5920 | 0.6960 | 0.7960 | 0.8920 |
| Economics | Mean | ✗ | 0.0947 | 0.1952 | 0.2997 | 0.4018 | 0.4985 | 0.5932 | 0.6936 | 0.7889 | 0.8867 |
| | | ✓ | 0.0916 | 0.1902 | 0.2913 | 0.3875 | 0.4855 | 0.5879 | 0.6855 | 0.7897 | 0.8863 |
| | Std ↓ | ✗ | 0.0363 | 0.0373 | 0.0424 | 0.0443 | 0.0458 | 0.0447 | 0.0385 | 0.0326 | 0.0279 |
| | | ✓ | **0.0242** | **0.0368** | **0.0415** | **0.0427** | **0.0455** | **0.0388** | **0.0294** | **0.0285** | **0.0250** |
| | Median | ✗ | 0.1000 | 0.1880 | 0.2920 | 0.4080 | 0.4880 | 0.5840 | 0.6800 | 0.7880 | 0.8840 |
| | | ✓ | 0.0920 | 0.1920 | 0.2920 | 0.3960 | 0.4880 | 0.5960 | 0.6920 | 0.7920 | 0.8880 |

tailed in Appendix D. We implement our SConU framework under the same settings using the Qwen-2-7B-Instruct and LLaMa-3.1-8B-Instruct models as examples. We also consider the median metric as mentioned in several studies (Deng et al., 2023; Snell et al., 2023; Zollo et al., 2024). As shown in Table 1, both the mean and median of the EMR results obtained from SConU are rigorously confined within the risk level, and the variance metric is significantly lower than that of the basic ConU framework on the Health and Economics subsets, highlighting the effectiveness of our approach.

In real-world QA tasks, LLMs often face queries from diverse disciplines (Kumar et al., 2023). However, as shown in Figure 2b, considerable issues of unbounded EMR emerge when the uncertainty distribution of test samples deviates from that of the provided calibration set, compromising the reliability of their prediction sets. For instance, when utilizing calibration data from the Psychology domain to address test samples from 13 other subsets, EMR values typically exceed the risk level of 0.28, peaking at 0.83 in the Math subject. Moreover, we may have no access to model logit. At this point, we incorporate the frequency score into the NS formulation and set $w_l = 0, w_f = 1$ following the study (Wang et al., 2024b). Then, we employ our SConU framework, which filters out uncertainty data outliers within each test subset. As illustrated in Figure 3a, the EMR metric for the Math discipline decreases to 0.15, while the results for other subjects remain confined by the minimum risk level of 0.28. When subsets from other disciplines are utilized as the calibration set, EMR results generally meet the guarantee of marginal coverage.

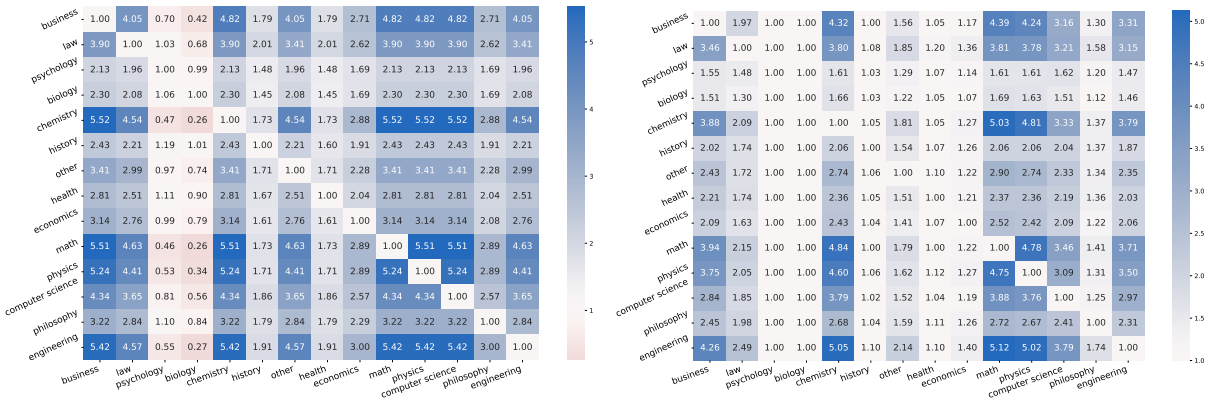
Despite the theoretical guarantee of SCP being rigorous, there can be minor fluctuations in practice due to finite-sample variability (Angelopoulos and Bates, 2021; Ye et al., 2024; Angelopoulos et al., 2024a). We notice EMR deviations in the results of SConU. To address this, we apply SConU-Pro by incorporating the prediction status of each calibration data point into the counting criteria of the conformal p-value, which evaluates the reference




(a) EMR results of SConU.

(b) EMR results of SConU-Pro.

Figure 3: Results of EMR after applying our two frameworks utilizing the LLaMA-3.1-8B-Instruct model on the MMLU-Pro dataset. Note that all data on the diagonal is manually set to equal $\alpha$ ($\alpha_l = 0.2723$).




(a) Original APSS results.

(b) APSS results of SConU-Pro.

Figure 4: Results of APSS before and after performing SConU-Pro, utilizing the LLaMA-3.1-8B-Instruct model on the MMLU-Pro dataset. Note that all data on the diagonal is manually set to 1.

values of the calibration samples across various risk levels. As demonstrated in Figure 3b, we achieve rigorous management of the EMR metric (i.e., $\leq \alpha$) in cross-domain scenarios. Furthermore, we compare the APSS metric before and after implementing SConU-Pro. As illustrated in Figure 4, when employing the Psychology or Biology subset as the calibration set, we observe APSS being less than 1 in the test sets of other disciplines, indicating that many test QA samples have empty prediction sets. Following the application of SConU-Pro, we attain an APSS metric of 1 for all selected test samples with the majority of EMR metrics equal to 0, suggesting that we accurately identify the correct answer for each test QA sample. In other calibration settings, the APSS results also exhibit a significant decline, thereby enhancing prediction efficiency.

More details of our performed conformal procedures can be found in Appendix D, and additional experimental results are presented in Appendix E.

**Conditional Coverage.** Given the critical importance of correctness coverage for individual samples in high-stakes QA tasks, we explore four key factors: exchangeability, the reliability of the NS in representing disagreements between query-answer pairs, split ratio, and model performance, and examine EMR across various set sizes. Our analysis focuses on the MedMCQA task and the Clinical Knowledge subset of the MMLU dataset. As presented in Table 2, when employing logit-based NSs, EMR values exceed the risk threshold at set sizes of 1 and 3. By incorporating the frequency score into the NS formulation and appropriately increasing the sample size, we observe a reduction in the SSM metric. Moreover, while more calibration samples enhance conditional performance, the SSM metric remains above the acceptable risk level. To address this, we utilize the conformal p-value to eliminate outliers, achieving approximate conditional coverage, with the SSM metric falling below the risk

Table 2: Results of the SSM metric obtained from 100 trials, under different settings on the MedMCQA dataset, utilizing the Qwen-2.5-14B-Instruct model (Mean $\pm$ Std). Red indicates out of risk level.

| $w_l$ **(Logit)** | $w_f$ **(Frequency)** | $M$ **(Sampling)** | **OD** | **Size = 1** | **Size = 2** | **Size = 3** | **SSM** $\downarrow$ |
|---|---|---|---|---|---|---|---|
| Split ratio is fix at 0.5 and $\alpha$ is set to 0.34 ($\alpha_l = 0.3342$). | | | | | | | |
| 1 | 0 | 10 | ✗ | $0.3428 \pm 0.0151$ | $0.2800 \pm 0.0277$ | $0.1056 \pm 0.1860$ | 0.3579 |
| 1 | 0 | 10 | ✓ | $0.3060 \pm 0.0054$ | $0.0348 \pm 0.0047$ | 0 | 0.3114 |
| 0.5 | 0.5 | 10 | ✗ | $0.3428 \pm 0.0144$ | $0.2971 \pm 0.0240$ | $0.1487 \pm 0.1594$ | 0.3572 |
| 0 | 1 | 10 | ✗ | $0.3391 \pm 0.0149$ | $0.2874 \pm 0.0251$ | $0.2177 \pm 0.1027$ | 0.3540 |
| 0 | 1 | 10 | ✓ | $0.3025 \pm 0.0067$ | $0.2795 \pm 0.0766$ | 0 | **0.3092** |
| Split ratio is fix at 0.7 and $\alpha$ is set to 0.34 ($\alpha_l$=0.3294). | | | | | | | |
| 1 | 0 | 10 | ✗ | $0.3404 \pm 0.0168$ | $0.2764 \pm 0.0395$ | $0.1212 \pm 0.2499$ | 0.3711 |
| 0.5 | 0.5 | 10 | ✗ | $0.3407 \pm 0.0157$ | $0.2955 \pm 0.0378$ | $0.1350 \pm 0.2069$ | 0.3564 |
| 0.5 | 0.5 | 20 | ✗ | $0.3402 \pm 0.0102$ | $0.2916 \pm 0.0337$ | $0.1160 \pm 0.1713$ | 0.3504 |
| 0.5 | 0.5 | 20 | ✓ | $0.3023 \pm 0.0112$ | $0.2665 \pm 0.0293$ | 0 | 0.3135 |
| 0 | 1 | 20 | ✗ | $0.3382 \pm 0.0154$ | $0.2927 \pm 0.0353$ | $0.1287 \pm 0.1156$ | 0.3536 |
| 0 | 1 | 20 | ✓ | $0.3006 \pm 0.0121$ | $0.2539 \pm 0.0109$ | 0 | **0.3127** |

Table 3: Mean of SSM results obtained from 100 trials at the risk level of 0.3 on the Clinical Knowledge subject of MMLU dataset. Note that we fix the split ratio to 0.5 and set $w_l = w_f = 0.5$ in the formulation of NS.

| **LLMs** | **OD** | **Size = 1** | **Size = 2** | **Size = 3** |
|---|---|---|---|---|
| Vicuna-7B-v1.5 | ✗ | 0.3233 | 0.3811 | 0.2113 |
| ($a_l = 0.2857$) | ✓ | 0.3229 | 0.2971 | 0.0733 |
| Vicuna-13B-v1.5 | ✗ | 0.3045 | 0.2769 | 0.2811 |
| ($a_l = 0.2556$) | ✓ | **0.2973** | **0.1813** | **0** |

Table 4: Results of mean APSS before and after semantic deduplication (SD) within prediction sets.

| **Dataset** | **LLMs** | **SD** | **Mean APSS** |
|---|---|---|---|
| TriviaQA | Qwen2.5-3B-Instruct | ✗ | 8.07 |
| | | ✓ | 1.08 |
| | Qwen2.5-7B-Instruct | ✗ | 8.77 |
| | | ✓ | 1.05 |
| | Qwen2.5-14B-Instruct | ✗ | 9.18 |
| | | ✓ | 1.03 |
| CoQA | Qwen2.5-3B-Instruct | ✗ | 8.65 |
| | | ✓ | 1.12 |
| | Qwen2.5-7B-Instruct | ✗ | 8.80 |
| | | ✓ | 1.04 |
| | Qwen2.5-14B-Instruct | ✗ | 8.84 |
| | | ✓ | 1.03 |

threshold at both split ratios. For instance, with a split ratio of 0.5, we attain an SSM value of 0.3092 using the frequency score derived from the candidate set of size 10. As shown in Table 3, model performance also plays a significant role in influencing conditional coverage, and our SConU-Pro framework consistently enhances the SSM metric.

We conclude that we can design NS using more reliable uncertainty measures based on the internal model information and the true sampling distribution. Additionally, we can appropriately increase the scale of the calibration data, although this will increase computational costs. Most importantly, it is essential to ensure exchangeability among QA samples. Finally, deploying task-specific models can further improve conditional performance.

**Prediction Efficiency.** In open-domain QA tasks, we observe significant semantic redundancy in the prediction sets generated by previous ConU frameworks (Wang et al., 2024d; Su et al., 2024). As shown in Table 4, the mean APSS from 100 trials decreases markedly before and after semantic deduplication, suggesting that there is considerable potential for improving the action efficiency of these prediction sets while maintaining the guarantee.

## 5 Conclusion

In this paper, we introduce SConU, a modular and principled framework aimed at eliminating uncertainty data outliers that violate the exchangeability precondition inherent in existing conformal approaches. We develop two conformal p-values to identify whether the given test QA sample significantly deviates from the uncertainty distribution of the calibration set as a user-specified risk level. Experimental results demonstrate the rigorous guarantees of marginal coverage and efficient prediction of SConU. Additionally, we derive the minimum risk level manageable by the calibration set without manually handling calibration data points post-deployment of the language model. Furthermore, we approximate conditional coverage across various sizes of the prediction set by analyzing several internal components of the conformal procedures.

## Limitations

Our SConU framework excludes test QA samples that significantly deviate from the uncertainty distribution of the calibration set. In future work, we will investigate strategies to address nonexchangeable data sequences by analyzing the degree of uncertainty distribution shift between the given test sample and the calibration set. Moreover, we achieve approximate conditional coverage at various prediction set sizes in high-stakes QA tasks, prompting us to conduct more comprehensive studies on the mechanisms influencing conditional performance on particular data points in subsequent research.

## References


AI@Meta. 2024. Llama 3 model card.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. 2024a. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*.

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. 2024b. Conformal risk control. In *The Twelfth International Conference on Learning Representations*.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. 2021. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. 2023. Conformal prediction beyond exchangeability. *The Annals of Statistics*.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. 2021. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. 2023. Testing for outliers with conformal p-values. *The Annals of Statistics*.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*.

Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário AT Figueiredo, and André FT Martins. 2024. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*.

John Cherian, Isaac Gibbs, and Emmanuel Candes. 2024. Large language model validity via enhanced conformal prediction methods. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jesse C Cresswell, Bhargava Kumar, Yi Sui, and Mouloud Belbahri. 2024a. Conformal prediction sets can cause disparate impact. *arXiv preprint arXiv:2410.01888*.

Jesse C. Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. 2024b. Conformal prediction sets improve human decision making. In *Forty-first International Conference on Machine Learning*.

Zhun Deng, Thomas P Zollo, Jake Snell, Toniann Pitassi, and Richard Zemel. 2023. Distribution-free statistical dispersion control for societal applications. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tiffany Ding, Anastasios Nikolas Angelopoulos, Stephen Bates, Michael I. Jordan, and Ryan J. Tibshirani. 2023. Class-conditional conformal prediction with many classes. *arXiv preprint arXiv:2306.09335*.

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

António Farinhas, Chrysoula Zerva, Dennis Thomas Ulmer, and Andre Martins. 2024. Non-exchangeable conformal risk control. In *The Twelfth International Conference on Learning Representations*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. 2023. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*.

Leying Guan and Robert Tibshirani. 2022. Prediction and outlier detection in classification problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

Yu Gui, Ying Jin, and Zhimei Ren. 2024. Conformal alignment: Knowing when to trust foundation models with guarantees. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Linhui Huang, S. Lala, and Niraj Kumar Jha. 2024. Confine: Conformal prediction for interpretable neural networks. *arXiv preprint arXiv:2406.00539*.

Ying Jin and Emmanuel J Candès. 2023. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*.

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)*.

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*.

Ramneet Kaur, Colin Samplawski, Adam D Cobb, Anirban Roy, Brian Matejek, Manoj Acharya, Daniel Elenius, Alexander Michael Berenbeim, John A Pavlik, Nathaniel D Bastian, et al. 2024. Addressing uncertainty in llms to enhance reliability in generative ai. In *Neurips Safe Generative AI Workshop 2024*.

Jungeum Kim, Sean O'Hagan, and Veronika Rockova. 2024. Adaptive uncertainty quantification for generative ai. *arXiv preprint arXiv:2408.08990*.

Vasily Kostumov, Bulat Nutfullin, Oleg Pilipenko, and Eugene Ilyushin. 2024. Uncertainty-aware evaluation for vision-language models. *arXiv preprint arXiv:2402.14418*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

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*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*.

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*.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. 2002. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning*.

E. J. G. Pitman. 1937. Significance tests which may be applied to samples from any populations. ii. the correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society*.

Vincent Plassier, Alexander Fishkov, Mohsen Guizani, Maxim Panov, and Eric Moulines. 2024. Probabilistic conformal prediction with approximate conditional validity. *arXiv preprint arXiv:2407.01794*.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal language modeling. In *The Twelfth 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*.

Jake Snell, Thomas P Zollo, Zhun Deng, Toniann Pitassi, and Richard Zemel. 2023. Quantile risk control: A flexible framework for bounding the probability of high-loss predictions. In *The Eleventh International Conference on Learning Representations*.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. 2024. API is enough: Conformal prediction for large language models without logit-access. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. 2019. Conformal prediction under covariate shift. *Advances in neural information processing systems*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael

Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melissa Hall Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Dennis Ulmer, Chrysoula Zerva, and Andre Martins. 2024. Non-exchangeable conformal language generation with nearest neighbors. In *Findings of the Association for Computational Linguistics: EACL 2024*.

Vladimir Vovk. 2012. Conditional validity of inductive conformal predictors. *Machine Learning*.

Vladimir Vovk, Ilia Nouretdinov, and Alexander Gammerman. 2003. Testing exchangeability on-line. In *Proceedings of the 20th international conference on machine learning (ICML-03)*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024a. Openchat: Advancing open-source language models with mixed-quality data. In *The Twelfth International Conference on Learning Representations*.

Qingni Wang, Tiantian Geng, Zhiyuan Wang, Teng Wang, Bo Fu, and Feng Zheng. 2024b. Sample then identify: A general framework for risk control and assessment in multimodal large language models. *arXiv preprint arXiv:2410.08174*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024c. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. 2024d. ConU: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Zhiyuan Wang, Jinhao Duan, Chenxi Yuan, Qingyu Chen, Tianlong Chen, Yue Zhang, Ren Wang, Xiaoshuang Shi, and Kaidi Xu. 2025. Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond. *Engineering Applications of Artificial Intelligence*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Ke-Yang Chen, Kexin Yang, Mei Li, Min Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yunyang Wan, Yunfei Chu, Zeyu Cui, Zhenru Zhang, and Zhi-Wei Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking LLMs via uncertainty quantification. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? *arXiv preprint arXiv:2405.16908*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Thomas P Zollo, Todd Morrill, Zhun Deng, Jake Snell, Toniann Pitassi, and Richard Zemel. 2024. Prompt risk control: A rigorous framework for responsible deployment of large language models. In *The Twelfth International Conference on Learning Representations*.

# A An illustration of the SCP framework

SCP can transform any heuristic notion of uncertainty from any model into a rigorous one (Angelopoulos and Bates, 2021). Let's illustrate the basic SCP framework by classification problems (Angelopoulos et al., 2021): Given the calibration set of size $N$, we define the NS of each sample as one minus the softmax output for the true class. Then we calculate the $\frac{\lceil (N+1)(1-\alpha) \rceil}{N}$ quantile of the $N$ sorted (ascending) NSs and employ it as the threshold to select possible classes for a new test sample. If the softmax output of a certain class falls below the threshold, by the exchangeability condition, we consider it to have an approximate probability of $1-\alpha$ to be the true label and add it to the prediction set. Finally, we achieve marginal correctness coverage on the finite-sample test set. The complete framework is presented as follows:

1. Given the calibration data set $\{(X_i, Y_i^*)\}_{i=1}^n$ (i.i.d.) and pretrained model $\hat{f}(\cdot)$ ($\hat{f}(X_i) \in [0,1]^{(K)}$). The probability of each true class (label) is denoted as $\hat{f}(X_i)_{Y_i^*}$.

2. Define and sort the nonconformity scores (uncertainty state associated with the true class of each calibration sample): $s_i = s(X_i, Y_i^*) = 1 - \hat{f}(X_i)_{Y_i^*}$ ($\{s_1 \leq \cdots \leq s_n\}$).

3. Obtain the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of $\{s_i\}_{i=1}^n$: $\hat{q} = \inf\left\{q : \frac{|\{i : s_i \leq q\}|}{n} \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right\} = s_{\lceil (n+1)(1-\alpha) \rceil}$.

4. Create the prediction set for $X_{test}$ following: $\mathcal{C}(X_{test}) = \{y \in [K] : s(X_{test}, y) \leq \hat{q}\}$

5. The event $Y_{test}^* \in \mathcal{C}(X_{test})$ is equivalent to $s(X_{test}, Y_{test}^*) \leq \hat{q}$. As long as $s(X_{test}, Y_{test}^*) \leq \hat{q}$ is satisfied, $Y_{test}^*$ is encompassed by $\mathcal{C}(X_{test})$, and then we obtain the prediction set that contains the true label.

6. By the exchangeability of $N+1$ data points, we have $\mathbb{P}(s_{test} \leq s_i) = \frac{i}{n+1}$.

7. Then we conclude: $\mathbb{P}(Y_{test}^* \in \mathcal{C}(X_{test})) = \mathbb{P}(s_{test} \leq \hat{q}) = \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \geq 1-\alpha$.

# B Additional Experimental Settings

## B.1 Base LLMs

We conduct experiments utilizing 4 popular series of "off-the-shelf" LLMs: OpenChat (Wang et al., 2024a), LLaMA (Touvron et al., 2023; AI@Meta, 2024), Vicuna (Zheng et al., 2023), and Qwen (Yang et al., 2024), divided by model size into: ① 3B: LLaMA-3.2-3B-Instruct and Qwen-2.5-3B-Instruct. ② 7B: Qwen-2-7B-Instruct, Qwen-2.5-7B-Instruct, and OpenChat-3.5. ③ 8B: LLaMA-3-8B-Instruct and LLaMA-3.1-8B-Instruct. ④ 13B: LLaMA-2-13B-Chat and Vicuna-13B-v1.5. ⑤ 14B: Qwen-2.5-14B-Instruct. ⑥ 32B: Qwen-2.5-32B-Instruct.

## B.2 Details of Datasets

**MMLU**[1] is a massive multi-task test consisting of multiple-choice questions from 57 subjects such as anatomy, astronomy, and business ethics. Following prior studies (Kumar et al., 2023; Su et al., 2024), we consider a subset of 16 subjects: computer security, high school computer science, college computer science, machine learning, formal logic, high school biology, anatomy, clinical knowledge, college medicine, professional medicine, college chemistry, marketing, public relations, management, business ethics, and professional accounting. Table 5 presents the number of samples employed for each subject from the MMLU dataset. Note that there is a slight deviation in the actual number of samples utilized for each model due to a few individual samples that do not comply with user instructions in all sampled responses (i.e., each response is not among A, B, C, or D).

Table 5: The number of samples employed for each subject from the MMLU dataset.

| Subjects | Number of Samples |
|---|---|
| computer security | 100 |
| high school computer science | 100 |
| college computer science | 100 |
| machine learning | 112 |
| formal logic | 126 |
| high school biology | 310 |
| anatomy | 135 |
| clinical knowledge | 265 |
| college medicine | 173 |
| professional medicine | 272 |
| college chemistry | 100 |
| marketing | 234 |
| public relations | 110 |
| management | 103 |
| business ethics | 100 |
| professional accounting | 282 |

[1] https://huggingface.co/datasets/cais/mmlu

**MMLU-Pro**[2] is a more robust and challenging multi-task understanding dataset. It expands samples from MMLU by increasing the 4 options for each question to 10, and the subjects are enhanced with questions from STEM Website, TheoremQA, and SciBench. This dataset totally contains 12,000 complex questions across various disciplines. qIn order for a balanced distribution of sample quantities across different subjects, we employ a maximum of 500 samples for each subject. The detailed sample quantities are shown in Table 6. Note that the number of samples applied for each model may have slight deviations (i.e., each response is not among A, B, C, D, E, F, G, H, I, or J).

Table 6: The number of samples employed for each subject from the MMLU-Pro dataset.

| Subjects | Number of Samples |
|---|---|
| computer science | 410 |
| math | 500 |
| chemistry | 500 |
| engineering | 500 |
| law | 500 |
| biology | 500 |
| health | 500 |
| physics | 500 |
| business | 500 |
| philosophy | 499 |
| economics | 500 |
| other | 500 |
| psychology | 500 |
| history | 381 |

For both MMLU and MMLU-Pro datasets, we utilize the test set of each subject, sourced from the `test-00000-of-00001.parquet` file.

**MedMCQA**[3] is designed to address real-world medical entrance exam questions. We consider the full validation set, 4,180 MCQA samples, sourced from the `validation-00000-of-00001.parquet` file. Note that several MCQA samples cannot be correctly encoded by the `tokenizer`, specifically non-ASCII characters. We exclude these samples, remaining 3,967 samples.

**TriviaQA**[4] is a reading comprehension dataset containing over 650,000 high-quality query-answer pairs. We utilize the validation set sourced from the `validation-00000-of-00001.parquet` file and randomly select 4,000 QA samples.

**CoQA**[5] is a large-scale conversational QA task, including 127,000 query-answer samples with their corresponding evidence highlighted in the provided context. We also utilize the validation set sourced from the `validation-00000-of-00001.parquet` file and randomly select 4,000 QA samples.

### B.3 Prompt Engineering

For both the MMLU and MMLU-Pro tasks, we randomly select 3 QA examples from the validation set of each subject, to construct a 3-shot prompt, which guides the language model in answering the current question using the specified response format (i.e., providing options like A, B, or C). Notably, all questions within the same subject share the same examples in the 3-shot prompt. For the MedMCQA task, we randomly selected 3 samples from the validation set as few-shot examples and exclude these three samples from subsequent experiments. We apply similar system prompts across the 3 MCQA datasets. Note that each question in the MMLU-Pro dataset generally includes 10 multiple-choice options, though some QA samples have fewer options following a manual review process to eliminate unreasonable choices. In the TriviaQA and CoQA tasks, we develop few-shot prompts following prior work (Duan et al., 2024; Wang et al., 2025). We provide complete prompt examples for 5 datasets, as presented in Figures 10−14.

### B.4 Hyperparameters

Following prior studies (Duan et al., 2024; Wang et al., 2024d,b), We employ multinominal sampling to generate $M$ candidate responses for each data point. For both the MMLU and MedMCQA datasets with 4 options for each question, we set the number of candidate responses, $M$, to 20, maintaining consistency with previous research (Kuhn et al., 2023; Lin et al., 2024; Quach et al., 2024). Since each sample in the MMLU-Pro dataset includes 10 multiple-choice options, we increase $M$ to 50 to better approximate the distribution of model outputs. For the TriviaQA and CoQA tasks, we generate 10 responses for each question (Wang et al., 2024d). Considering that we develop prompts to guide the language model in responding with the most probable option letters (e.g., A, B, or C), as detailed in Appendix B.3, we set the maximum generation length to 1 to accelerate sampling in 3 MCQA tasks. For open-domain QA, we examine

[2] https://huggingface.co/datasets/TIGER-Lab/MMLU-Pro
[3] https://huggingface.co/datasets/openlifescienceai/medmcqa
[4] https://huggingface.co/datasets/mandarjoshi/triviaqa
[5] https://huggingface.co/datasets/stanfordnlp/coqa

| Hyperparameter | Value |
|---|---|
| do_sample | True |
| num_beams | 1 |
| top_p | 0.9 |
| temperature | 1.0 |
| max_length | input_length + 1/36 |

Table 7: Hyperparameters for the generate function. input_length is the embedding length of the input prompt after being encoded by the tokenizer of the current language model.

the maximum length of answers for all randomly selected samples, and set the maximum generation length for 2 tasks to 36. In the generate function, we configure the hyperparameters as presented in Table 7. Moreover, since the conformal p-value detects when test points do not come from the same distribution of the calibration set, we guarantee that it does not return too many false positives and set $\delta$ equal to the user-specified risk level, following prior work (Angelopoulos and Bates, 2021; Jin and Candès, 2023; Gui et al., 2024; Huang et al., 2024).

## C Conformal p-value

In this section, we first demonstrate that the conformal p-value formulated in Eq. (1) adheres to the statistical definition of p-values. As mentioned, $u_i$ represents the uncertainty of the LLM addressing the $i$-th question. At this point, we can denotes $p_{N+1}$ as

$$\begin{aligned} p_{N+1} &= \frac{1+\sum_{i=1}^{N}\mathbf{1}\left\{u_i \geq u_{N+1}\right\}}{N+1}, \\ &= \frac{k}{N+1} \end{aligned} \tag{6}$$

where $k$ is the position of $u_{N+1}$ in the sorted (i.e., ascending) sequence of $N+1$ uncertainty scores, and we have

$$\begin{aligned} \mathbb{P}\left(p_{N+1} \leq \delta\right) &= \mathbb{P}\left(\frac{k}{N+1} \leq \delta\right) \\ &= \mathbb{P}\left(k \leq \lfloor (N+1)\,\delta \rfloor\right) \end{aligned} . \tag{7}$$

Since we apply the consistent uncertainty measure for each QA sample, the $N+1$ uncertainty scores are exchangeable. Then, we obtain

$$\begin{aligned} \mathbb{P}\left(p_{N+1} \leq \delta\right) &= \frac{\lfloor (N+1)\,\delta \rfloor}{N+1} \\ &\leq \frac{(N+1)\,\delta}{N+1} \\ &\leq \delta \end{aligned} . \tag{8}$$

As mentioned in section 4.2, we observe minor fluctuations in the results of SConU under cross-domain scenarios. This arises from the hallucination issues of LLMs. For example, consider two questions with similar sampling distribution. However, in one question's candidate set, nearly all the answers are incorrect, while in the other question's sampling set, most answers are correct. In this case, the scores obtained from the uncertainty measure for the two samples may be the same, but in fact, the answering situations of the two QA samples are opposite, which can affect the exchangeability of the uncertainty scores, leading to slight variations in the performance of outlier detection.

To check whether the uncertainty score of each calibration data point is referenceable at different risk levels, we incorporate their prediction status into the counting criterion. At this point, we denote the count of calibration samples that satisfy both $u_i \geq u_{N+1}$ and $y_i^* \in E\left(x_i, \mathcal{D}_{cal}, \alpha\right)$. Thus the conformal p-value can be expressed as $p_{N+1}^{'} = \frac{1+k}{N+1}$. Here, $k$ can take values from 0 to $N$, so the range of $p_{N+1}^{'}$ is $\left[\frac{1}{N+1}, 1\right]$. Similar to Eq. (7), we have

$$\mathbb{P}\left(\frac{1+k}{N+1} \leq \delta\right) = \mathbb{P}\left(k \leq (N+1)\,\delta - 1\right). \tag{9}$$

Let $m = \lfloor (N+1)\,\delta - 1 \rfloor$. Since $k$ can be at most $N$, if $M < 0$, then $p_{N+1}$ will always be greater than any negative value, so $\mathbb{P}\left(p_{N+1}^{'} \leq \delta\right) = 0 \leq \delta$. If $0 \leq m \leq N$, we have

$$\mathbb{P}\left(k \leq m\right) \leq \frac{m+1}{N+1}. \tag{10}$$

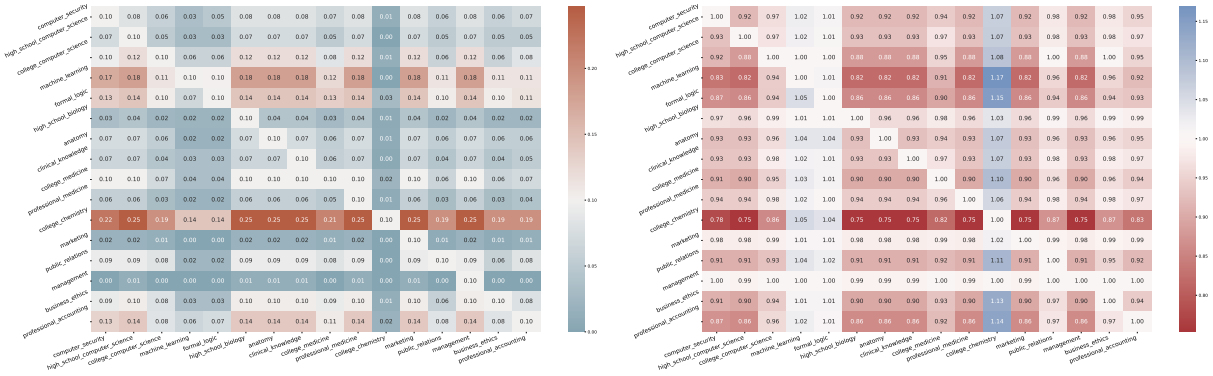
Therefore,

$$\begin{aligned} \mathbb{P}\left(p_{N+1}^{'} \leq \delta\right) &\leq \frac{m+1}{N+1}. \\ &\leq \delta \end{aligned} \tag{11}$$

In summary, we have demonstrated that our developed two conformal p-values satisfy the statistical definition of p-values.

Table 8: The minimum risk level manageable by each subject of the calibration set in the MMLU dataset utilizing the Qwen2.5-32B-Instruct model.

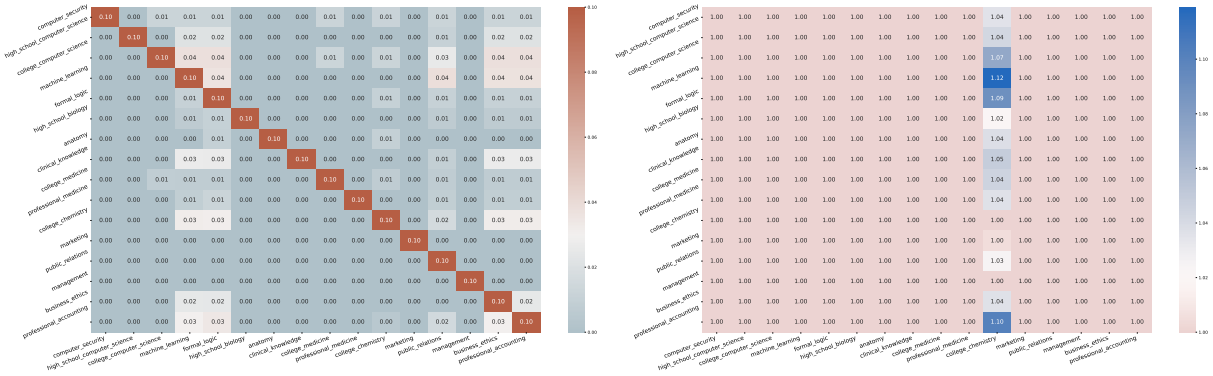
| **Subjects (Computer Science)** | $a_l$ | **Subjects (Medicine)** | $a_l$ | **Subjects (Business)** | $a_l$ |
|---|---|---|---|---|---|
| computer security | 0 | high school biology | 0.007 | marketing | 0 |
| high school computer science | 0 | anatomy | 0.009 | public relations | 0 |
| college computer science | 0 | clinical knowledge | 0.004 | management | 0 |
| machine learning | 0 | college medicine | 0.014 | business ethics | 0.011 |
| formal logic | 0.019 | professional medicine | 0.008 | professional accounting | 0.018 |
| | | college chemistry | **0.051** | | |



(a) EMR results of the basic ConU framework.

(b) APSS results of the basic ConU framework.

Figure 5: Results of the EMR and APSS metrics obtained from the basic ConU framework on the MMLU dataset utilizing the Qwen2.5-32B-Instruct model at the risk level of 0.1.



(a) EMR results of our SConU framework.

(b) APSS results of our SConU framework.

Figure 6: Results of the EMR and APSS metrics obtained from our SConU framework on the MMLU dataset utilizing the Qwen2.5-32B-Instruct model at the risk level of 0.1.

Considering that a single uncertainty notion cannot fully represent the exchangeability among QA samples, we can perform multiple hypothesis testing to identify uncertainty data outliers in practical high-stakes QA applications. As mentioned in Section 3.2, we utilize PE as the uncertainty measure, formulated as $u_i = PE(x_i) = \sum_{o=1}^{O_i} -p_o \log p_o$, where $p_o$ denotes the logit-based confidence score of the $o$-th option and $O_i$ denotes the number of options for the $i$-th question (e.g., 4 or 10). Here, for each QA sample, we define $B$ notions of uncertainty: $\left\{u_b^{(i)}\right\}_{b=1}^{B}$, such as the number of semantics within the candidate set (Lin et al., 2024) and the frequency-based PE. At this point, we check whether its $B$ types of uncertainty significantly deviate from the calibration set for each test data point. If any one of them does not meet the criterion, we consider that the exchangeability condition is violated and decline to provide a prediction set.

We determine the significance level for the p-

value associated with each uncertainty notion by the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). More details can be referred to the study (Jin and Candès, 2023). Finally, for each test QA sample, if a certain conformal p-value associated with one uncertainty notion is lower than the significance level calculated by the BH procedure, we reject the null hypothesis and decline to provide an answer. Conversely, when multiple hypothesis testing indicates that the $N+1$ QA samples are exchangeable, we select task-specific ConU methods. Next, we present several typical frameworks.

## D Details of Conformal Procedures

Similar to Prompt Risk Control (PRC) (Zollo et al., 2024), our approach is orthogonal to some existing conformal frameworks. For MCQA tasks within the same discipline or dataset, we apply the basic procedures in prior studies (Kumar et al., 2023; Ye et al., 2024; Kostumov et al., 2024) and evaluate the EMR metric before and after implementing our developed conformal p-value. In formulation, the NS of each option can be expressed as $1 - w_l \cdot F_l\left(y_i^*\right) - w_f \cdot F_f\left(y_i^*\right)$ as defined in Section 3.1. Here, we only utilize the confidence score obtained from the model logit and set $w_l = 1$ and $w_f = 0$. At this point, we calculate the uncertainty score based on the logit-based PE method.

In more practical cross-domain settings, we investigate employing the black-box frequency score to formulate the NS following the research (Wang et al., 2024b), assuming no access to model internal information, and set $w_l = 0$ and $w_f = 1$. We use the frequency score of each option obtained from the candidate set of size 20 (or 50) to characterize the probability of $p_o$ and calculate frequency-based PE to implement uncertainty data outlier detection. Note that the performance of uncertainty quantification methods to differentiate between correct and incorrect answers affects the effectiveness of conformal p-values in identifying outliers. This is because a single notion of uncertainty cannot fully characterize the exchangeability among data points. Therefore, by applying more efficient uncertainty measures (Lin et al., 2024; Duan et al., 2024; Wang et al., 2025), we can enhance the capability of the NS to represent the disagreement between the current question and response while also improving the statistical rigor of significance tests.

In open-domain QA tasks, we employ the similar ConU framework applicable to black-box settings introduced in the study (Wang et al., 2024d). The NS of each calibration data is formulated as $1 - 0.5 \cdot F\left(y_{ref}^{(i)}\right) - 0.5 \cdot \frac{1}{M}\sum_{j=1}^{M} S\left(y_{ref}^{(i)}, y_j^{(i)}\right) F\left(y_j^{(i)}\right)$, where $y_{ref}^{(i)}$ represent the response in the candidate set that have equivalent semantics to the ground-truth $y_i^*$, $F\left(y_{ref}^{(i)}\right)$ measures the number of generations that is semantically equivalent to $y_{ref}^{(i)}$ (i.e., the frequency score of correct semantic), and $S\left(y_{ref}^{(i)}, y_j^{(i)}\right)$ measures the semantic similarity score between $y_{ref}^{(i)}$ and $y_j^{(i)}$ in the candidate set. Refer to the studies (Wang et al., 2024d,b; Su et al., 2024) for more details. We also link the NS with the uncertainty state of acceptable semantics.

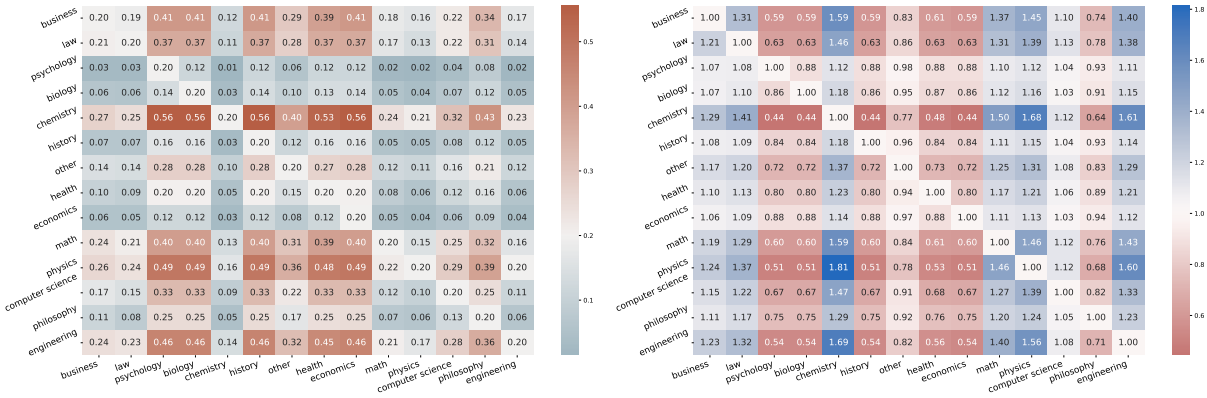
Table 9: The minimum risk level manageable by each subject of the calibration set in the MMLU-Pro dataset utilizing the Qwen2.5-32B-Instruct model.

| **Subjects** | $a_l$ |
|---|---|
| computer science | 0.075 |
| math | 0.123 |
| chemistry | 0.164 |
| engineering | 0.126 |
| law | 0.109 |
| biology | 0.032 |
| health | 0.047 |
| physics | **0.161** |
| business | 0.112 |
| philosophy | 0.045 |
| economics | 0.030 |
| other | 0.092 |
| psychology | 0.015 |
| history | 0.031 |

## E Additional Experimental Results

Note that in all experimental results within the interdisciplinary scenarios, the discipline from the horizontal axis represents the calibration set, while the discipline from the vertical axis represents the test set. When the calibration set and test set belong to the same discipline along the diagonal, all EMR results are directly set equal to the corresponding risk level of $\alpha$, and the APSS results are set to 1.

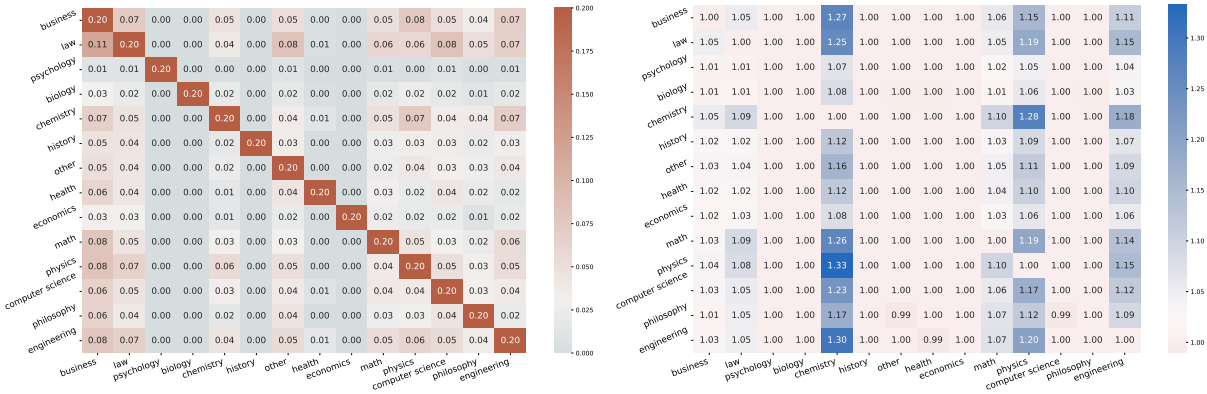
In this section, we also evaluate our SConU framework in cross-domain settings utilizing the MMLU dataset with the Qwen2.5-32B-Instruct model employed as the generator. Firstly, we calculate the minimum risk level manageable by each subject of the calibration set based on Eq. (5), as




(a) EMR results of the basic ConU framework.

(b) APSS results of the basic ConU framework.

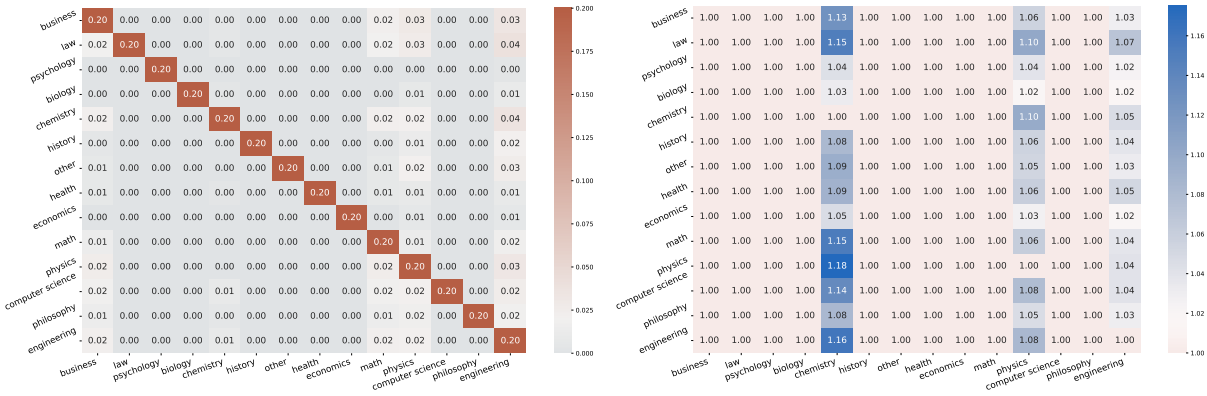
Figure 7: Results of the EMR and APSS metrics obtained from the basic ConU framework on the MMLU-Pro dataset utilizing the Qwen2.5-32B-Instruct model at the risk level of 0.2.




(a) EMR results of our SConU framework.

(b) APSS results of our SConU framework.

Figure 8: Results of the EMR and APSS metrics obtained from our SConU framework on the MMLU-Pro dataset utilizing the Qwen2.5-32B-Instruct model at the risk level of 0.2.




(a) EMR results of our SConU-Pro framework.

(b) APSS results of our SConU-Pro framework.

Figure 9: Results of the EMR and APSS metrics obtained from our SConU-Pro framework on the MMLU-Pro dataset utilizing the Qwen2.5-32B-Instruct model at the risk level of 0.2.

presented in Table 8. Then, we specify the risk level to 0.1 ($\alpha = \delta = 0.1$) and evaluate the results of the EMR metric on each subject of the test set before and after performing our SConU framework. As shown in Figures 5 and 6, before conducting outliers detection and elimination within the test set, significant issues arise: the EMR exceeds the risk level (i.e., $\geq 0.1$) in testing datasets such as college

chemistry, and many datasets report an APSS metric below 1, indicating that a substantial number of QA samples resulted in empty predictions. After filtering out test samples that significantly deviate from the calibration set, our foundational SConU framework achieves strict EMR control, with the APSS metrics of the test set nearly all at 1, highlighting that our method accurately identifies the correct answers.

On the more robust and challenging MMLU-Pro task with 10 options for each query, the minimum manageable risk level of each calibration set significantly increases, as presented in Table 9. We set the risk level of $\alpha$ to 0.2 and present the results of the EMR metric utilizing the basic ConU framework in Figure 7. The phenomenon of EMR surpassing the risk level is both frequent and severe. For example, when utilizing the Biology subset as the calibration to address queries from the Chemistry subject, the EMR metric is 0.56, significantly exceeding 0.2. Similarly, there are numerous QA samples where the prediction sets are empty, resulting in several subjects of the test sets having APSS scores below 1. Our SConU framework consistently maintains strict control over the EMR metric across all calibration-test set pairs, while also achieving higher prediction efficiency, as shown in Figure 8. Furthermore, as demonstrated in Figure 9, SConU-Pro achieves lower EMR and APSS metrics across all calibration and test sets by assessing the reliability of the uncertainty scores of each calibration sample at a specific risk level.

MMLU

### System:
Answer the following multiple-choice question by giving the most appropriate response. Answer should be one among [A, B, C, D].

### User:
What is penetration testing?
A: A procedure for testing libraries or other program components for vulnerabilities; B: Whole-system testing for security flaws and bugs; C: A security-minded form of unit testing that applies early in the development process; D: All of the above
### Assistant:
B

### User:
Suppose a user has an iPhone (running iOS) and downloads an app called Innocent from the Apple App Store and installs it. The user unlocks the phone and runs Innocent. Innocent exploits a bug in the iOS kernel which allows Innocent to redirect execution inside the kernel to code that Innocent controls. Now Innocent can execute any instructions it likes inside the iOS kernel. Innocent is not able to exploit any bugs in the phone's secure enclave. Can Innocent read the user's private information stored in the phone's flash (e.g. Contacts and messages), or will the security measures described in the paper keep the data private? If Innocent is only able to see encrypted data, then the phone has successfully kept the data private. Circle the security features of the phone (if any) that will prevent Innocent from reading information from the flash on the phone.
A: Secure boot chain; B: System software authorization; C: The secure enclave's ephemeral key; D: None of the above
### Assistant:
D

### User:
Why is it that anti-virus scanners would not have found an exploitation of Heartbleed?
A: It's a vacuous question: Heartbleed only reads outside a buffer, so there is no possible exploit; B: Anti-virus scanners tend to look for viruses and other malicious; C: Heartbleed attacks the anti-virus scanner itself; D: Anti-virus scanners tend to look for viruses and other malicious code, but Heartbleed exploits steal secrets without injecting any code
### Assistant:
D

### User:
Which of the following styles of fuzzer is more likely to explore paths covering every line of code in the following program?
A: Generational; B: Blackbox; C: Whitebox; D: Mutation-based
### Assistant:

Figure 10: An example of the prompt in the MMLU task.

MMLU-Pro

### System:
Answer the following multiple-choice question by giving the most appropriate response. Answer should be one among [A, B, C, D, E, F, G, H, I, J].

### User:
In contrast to ____, ____ aim to reward favourable behaviour by companies. The success of such campaigns have been heightened through the use of ____, which allow campaigns to facilitate the company in achieving ____.
A: Boycotts, Buyalls, Blockchain technology, Increased Sales; B: Buycotts, Boycotts, Digital technology, Decreased Sales; C: Boycotts, Buycotts, Digital technology, Decreased Sales; D: Buycotts, Boycotts, Blockchain technology, Charitable donations; E: Boycotts, Buyalls, Blockchain technology, Charitable donations; F: Boycotts, Buycotts, Digital technology, Increased Sales; G: Buycotts, Boycotts, Digital technology, Increased Sales; H: Boycotts, Buycotts, Physical technology, Increased Sales; I: Buycotts, Buyalls, Blockchain technology, Charitable donations; J: Boycotts, Buycotts, Blockchain technology, Decreased Sales
### Assistant:
F

### User:
____ is the direct attempt to formally or informally manage ethical issues or problems, through specific policies, practices and programmes.
A: Operational management; B: Corporate governance; C: Environmental management; D: Business ethics management; E: Sustainability; F: Stakeholder management; G: Social marketing; H: Human resource management; I: N/A; J: N/A
### Assistant:
D

### User:
How can organisational structures that are characterised by democratic and inclusive styles of management be described?
A: Flat; B: Bureaucratic; C: Autocratic; D: Hierarchical; E: Functional; F: Decentralized; G: Matrix; H: Network; I: Divisional; J: Centralized
### Assistant:
A

### User:
Typical advertising regulatory bodies suggest, for example that adverts must not: encourage ____, cause unnecessary ____ or ____, and must not cause ____ offence.
A: Safe practices, Fear, Jealousy, Trivial; B: Unsafe practices, Distress, Joy, Trivial; C: Safe practices, Wants, Jealousy, Trivial; D: Safe practices, Distress, Fear, Trivial; E: Unsafe practices, Wants, Jealousy, Serious; F: Safe practices, Distress, Jealousy, Serious; G: Safe practices, Wants, Fear, Serious; H: Unsafe practices, Wants, Fear, Trivial; I: Unsafe practices, Distress, Fear, Serious
### Assistant:

Figure 11: An example of the prompt in the MMLU-Pro task. Note that the current question has 9 options.

MedMCQA

### System:
Answer the following multiple-choice question by giving the most appropriate response. Answer should be one among [A, B, C, D].

### User:
Kamlesh, a 2 year old girl, has Down's syndrome. Her karyotype is 21/21 translocation. What is the risk ofrecurrence in subsequent pregnancies if the father is a balanced translocation carrier :
A: 100%; B: 50%; C: 25%; D: 0%
### Assistant:
A

### User:
Not a part of ethmoid bone is
A: Inferior turbinate; B: Agar nasi cells; C: Uncinate process; D: Crista galli
### Assistant:
A

### User:
Haddon matrix is related to:
A: Injury prevention; B: Communicable diseases; C: Maternal and child moality; D: Hypeensive disorders
### Assistant:
B

### User:
Which of the following is not true for myelinated nerve fibers:
A: Impulse through myelinated fibers is slower than non-myelinated fibers; B: Membrane currents are generated at nodes of Ranvier; C: Saltatory conduction of impulses is seen; D: Local anesthesia is effective only when the nerve is not covered by myelin sheath
### Assistant:

Figure 12: An example of the prompt in the MedMCQA task.

TriviaQA

### System:
This is a bot that correctly answers questions.

### User:
In 1968, who did radical feminist Valerie Solanas shoot and wound as he entered his New York studio?
### Assistant:
Andy Warhol

### User:
What lake can be found on the border of Vermont and New York?
### Assistant:
Lake Champlain

### User:
Which competition was won by Nadiya Hussain in 2015?
### Assistant:
The Great British Bake-Off

### User:
Who was the man behind The Chipmunks?
### Assistant:

Figure 13: An example of the prompt in the TriviaQA task.

CoQA

### System: This is a bot that correctly answers questions.
Once upon a time, in a barn near a farm house, there lived a little white kitten named Cotton. Cotton lived high up in a nice warm place above the barn where all of the farmer's horses slept. But Cotton wasn't alone in her little home above the barn, oh no. She shared her hay bed with her mommy and 5 other sisters. All of her sisters were cute and fluffy, like Cotton. But she was the only white one in the bunch. The rest of her sisters were all orange with beautiful white tiger stripes like Cotton's mommy. Being different made Cotton quite sad. She often wished she looked like the rest of her family. So one day, when Cotton found a can of the old farmer's orange paint, she used it to paint herself like them. When her mommy and sisters found her they started laughing. "What are you doing, Cotton?!" "I only wanted to be more like you". Cotton's mommy rubbed her face on Cotton's and said "Oh Cotton, but your fur is so pretty and special, like you. We would never want you to be any other way". And with that, Cotton's mommy picked her up and dropped her into a big bucket of water. When Cotton came out she was herself again. Her sisters licked her face until Cotton's fur was all all dry. "Don't ever do that again, Cotton!" they all cried. "Next time you might mess up that pretty white fur of yours and we wouldn't want that!" Then Cotton thought, "I change my mind. I like being special".

### User:
What color was Cotton?
### Assistant:
white

### User:
Where did she live?
### Assistant:
in a barn

### User:
Did she live alone?
### Assistant:
no

### User:
Who did she live with?
### Assistant:

Figure 14: An example of the prompt in the CoQA task.