# API Is Enough: Conformal Prediction for Large Language Models Without Logit-Access

**Jiayuan Su**[1], **Jing Luo**[2], **Hongwei Wang**[1]*, **Lu Cheng**[3]*
[1]ZJU-UIUC Institute, Zhejiang University
[2]School of Mathematics and Statistics, Shandong University
[3]Department of Computer Science, University of Illinois Chicago
{jiayuan.23, hongweiwang}@intl.zju.edu.cn
{luojing020713}@gmail.com
{lucheng}@uic.edu

## Abstract

This study aims to address the pervasive challenge of quantifying uncertainty in large language models (LLMs) without logit-access. Conformal Prediction (CP), known for its model-agnostic and distribution-free features, is a desired approach for various LLMs and data distributions. However, existing CP methods for LLMs typically assume access to the logits, which are unavailable for some API-only LLMs. In addition, logits are known to be miscalibrated, potentially leading to degraded CP performance. To tackle these challenges, we introduce a novel CP method that (1) is tailored for API-only LLMs without logit-access; (2) minimizes the size of prediction sets; and (3) ensures a statistical guarantee of the user-defined coverage. The core idea of this approach is to formulate nonconformity measures using both coarse-grained (i.e., sample frequency) and fine-grained uncertainty notions (e.g., semantic similarity). Experimental results on both close-ended and open-ended Question Answering tasks show our approach can mostly outperform the logit-based CP baselines.

## 1 Introduction

Large Language Models (LLMs) have made significant advancements (Thoppilan et al., 2022; Wei et al., 2022, 2023; Xie et al., 2023), highlighting the research potential of natural language generation (Peinl and Wirth, 2023). However, they often generate information that is not accurate, factual, or grounded in reality, referred to as "hallucination" (LeCun, 2023). Therefore, it is crucial to quantify LLM uncertainty to ensure responsible responses.

However, uncertainty quantification (UQ) for LLMs is challenging due to the complex data distributions and inner model mechanism, as well as the often limited access to logit information. A potential solution is to use conformal prediction (CP) (Vovk et al., 2005; Angelopoulos and Bates, 2021; Kato et al., 2023; Wang et al., 2023), which is known for being model-agnostic and distribution-free, and with rigorous coverage guarantees. Given a user-defined error rate $\alpha$, CP provides a guaranteed coverage rate for prediction sets/intervals. It measures the uncertainty from a model prediction using nonconformity score functions, e.g., $1 - f(X)_Y$ (Sadinle et al., 2019), where $f(X)_Y$ is the softmax score for the true label $Y$.

Most of the existing CPs for LLMs rely on the access to model logits to measure nonconformity scores. For instance, Kumar et al. (2023) define nonconformity scores as softmax scores for logits of different options in the multi-choice question answering (MCQ) task and Quach et al. (2023) apply the conformal risk control framework (Angelopoulos et al., 2021), an extension of CP, to LLMs by utilizing model-based log probability. However, for some API-only LLMs like Bard (Manyika and Hsiao, 2023), logit-access is almost impossible for end users. Even though the logits are available (e.g., GPT 4V (OpenAI, 2023)), they are known to be miscalibrated and can lead to degraded performance of CP w.r.t. estimating the prediction sets or intervals (Nguyen and O'Connor, 2015; Lin et al., 2022), e.g., a large set size (i.e., low efficiency).

To enable CP without logit-access, a straightforward way is to calculate the frequency of each response via sampling and approximate model-based probabilities. However, we theoretically prove that this approach is extremely computationally expensive (Lemma 3.1). As nonconformity scores typically measure the level of uncertainty, CP depends on the *ranking* of the nonconformity measures rather than their actual *values* (Shafer and Vovk, 2008). Therefore, we propose to sample responses for a certain number of times (e.g., 30) for each input and then utilize the frequency of each response as a *coarse-grained* uncertainty notion. This approach reduces the overall sampling costs and eliminates the dependence on the logits.

*Corresponding authors.

*Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 979–995
November 12-16, 2024 ©2024 Association for Computational Linguistics

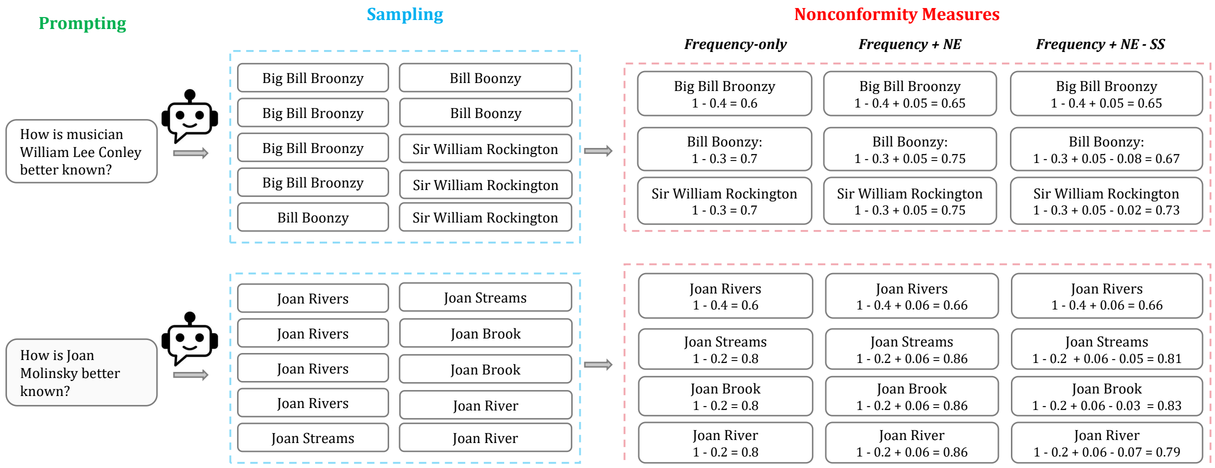

Prompting
Sampling
Nonconformity Measures
Frequency-only
Frequency + NE
Frequency + NE - SS
How is musician William Lee Conley better known?
Big Bill Broonzy
Bill Boonzy
Big Bill Broonzy
Bill Boonzy
Big Bill Broonzy
Sir William Rockington
Big Bill Broonzy
Sir William Rockington
Bill Boonzy
Sir William Rockington
Big Bill Broonzy 1 - 0.4 = 0.6
Big Bill Broonzy 1 - 0.4 + 0.05 = 0.65
Big Bill Broonzy 1 - 0.4 + 0.05 = 0.65
Bill Boonzy: 1 - 0.3 = 0.7
Bill Boonzy: 1 - 0.3 + 0.05 = 0.75
Bill Boonzy: 1 - 0.3 + 0.05 - 0.08 = 0.67
Sir William Rockington 1 - 0.3 = 0.7
Sir William Rockington 1 - 0.3 + 0.05 = 0.75
Sir William Rockington 1 - 0.3 + 0.05 - 0.02 = 0.73
How is Joan Molinsky better known?
Joan Rivers
Joan Streams
Joan Rivers
Joan Brook
Joan Rivers
Joan Brook
Joan Rivers
Joan River
Joan Streams
Joan River
Joan Rivers 1 - 0.4 = 0.6
Joan Rivers 1 - 0.4 + 0.06 = 0.66
Joan Rivers 1 - 0.4 + 0.06 = 0.66
Joan Streams 1 - 0.2 = 0.8
Joan Streams 1 - 0.2 + 0.06 = 0.86
Joan Streams 1 - 0.2 + 0.06 - 0.05 = 0.81
Joan Brook 1 - 0.2 = 0.8
Joan Brook 1 - 0.2 + 0.06 = 0.86
Joan Brook 1 - 0.2 + 0.06 - 0.03 = 0.83
Joan River 1 - 0.2 = 0.8
Joan River 1 - 0.2 + 0.06 = 0.86
Joan River 1 - 0.2 + 0.06 - 0.07 = 0.79

Figure 1: Illustrations of the proposed problem and solution. Three uncertainty notions for measuring nonconformity: (1) Frequency-only, where the nonconformity score is calculated as $1-$ *the frequency of a response out of 10 samplings*. Concentration issues arise at scores of **0.6, 0.7, and 0.8**. For instance, responses from different prompts (e.g., "Big Bill Broonzy" and "Joan Rivers") have the same score of 0.6, as well as responses within the same prompt (e.g., "Bill Boonzy" and "Sir William Rockington") which both have a score of 0.7, and so forth. (2) Frequency combined with NE, where the nonconformity score is calculated as $1-$ *frequency* $+$ *NE*, revealing concentration issues at scores of **0.75 and 0.86**. (3) Frequency, NE, and SS combined, where the nonconformity score is calculated as $1-$ *frequency* $+$ *NE* $-$ *SS*, with **no observed concentration issues**.

However, when using frequency as the only nonconformity measure, we observe that nonconformity scores concentrate on certain values as some responses may share the same frequency even if they have varied levels of uncertainty (see Figure 1), consequently diminishing the efficiency of prediction sets.

To distinguish between responses that share the same frequency, we first identify two potential causes: the respective concentration issues across different prompts and within the same prompt, which indicates we need to integrate prompt-wise and response-wise notions to respectively mitigate these two causes. We then propose two additional *fine-grained* uncertainty notions: normalized entropy (NE), measuring prompt-wise self-consistency to alleviate concentration issues across different prompts; and semantic similarity (SS), measuring response-wise similarity to the most frequent response within the same prompt, to mitigate internal concentration issues specific to the prompt. Figure 1 illustrates the different scores defined using *frequency-only*, *frequency combined with NE*, and *frequency combined with NE and SS* as nonconformity measures, respectively. By considering various uncertainty information, the proposed nonconformity score function can better distinguish the uncertainty of different responses.

Our contributions are summarized as follows:

- To our knowledge, this is the first CP work dedicated to LLMs without logit-access that provides a coverage guarantee for the prediction set with a small size.
- We propose a novel CP approach that uses both coarse-grained and fine-grained uncertainty notions as the non-conformity measures. We also theoretically prove (1) it is computationally infeasible to use response frequency to approximate model output probability, and (2) our approach ensures a rigorous statistical coverage guarantee.
- We conduct experiments on both close- and open-ended QA tasks and demonstrate the effectiveness of our method. Notably, we mostly surpass all baselines, including four logit-access methods and one method without logit-access.

## 2 Preliminaries of Conformal Prediction

Conformal prediction (CP) (Vovk et al., 2005) is a model-agnostic method offering distribution-free uncertainty quantification, which produces prediction sets/intervals containing ground-truth labels with a desired error rate $\alpha$. One of the widely used CP methods is split CP. Formally, let $(X, Y)$ be a sample, where $X$ represents features and $Y$ represents the outcome. We denote the calibration set as $(X_i, Y_i)_{i=1,\ldots,n}$ and the test set as $(X_{\text{test}}, Y_{\text{test}})$. CP presents the following nesting property:

$$\alpha_1 > \alpha_2 \Rightarrow C_{1-\alpha_1}(X) \subseteq C_{1-\alpha_2}(X). \qquad (1)$$

*where $C : X \to 2^Y$ is a set-valued function that generates prediction sets over the powerset of $Y$ given an input $X$.*

**Theorem 2.1** (Conformal coverage guarantee)**.** *Suppose $(X_i, Y_i)_{i=1,\ldots,n}$ and $(X_{test}, Y_{test})$ are* *independent and identically distributed (i.i.d.).* $C_{1-\alpha}(X_{test})$ *is a set-valued mapping satisfying the nesting property in Eq. 1. Then the following holds:*

$$P(Y_{test} \in C_{1-\alpha}(X_{test})) \geq 1-\alpha, \tag{2}$$

*where* $\alpha \in (0,1)$ *is the user-defined error rate.*

**Nonconformity Measures.** The nonconformity measure $N$ is a core element in CP. It measures uncertainty in the model's output by assessing the deviation of a specific instance or output from patterns observed in the training data. Typically, we have logit access to models to measure nonconformity, e.g., $1 - f(X)_Y$. For LLMs, $N$ is typically derived from the post-hoc logits.

**Split CP Steps.** Split CP typically involves four steps (Angelopoulos and Bates, 2021):

1. Establish heuristic uncertainty notions.
2. Define the nonconformity measures/score function $N(x, y) \in \mathbb{R}$.
3. Compute $\hat{q}$ as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the nonconformity scores.
4. Use $\hat{q}$ to generate prediction sets for new examples: $C(X_{\text{test}}) = \{Y : N(X_{\text{test}}, Y) \leq \hat{q}\}$.

## 3 Methodology

Our method considers two pivotal challenges arising from the LLMs without logit-access: how to approximate the logit information of LLMs; and how to further improve CP efficiency, i.e., small prediction sets. We propose the **Lo**git-**free** **C**onformal **P**rediction for LLMs (**LofreeCP**), where its nonconformity measures consist of three notions: *frequency*, representing coarse-grained uncertainty; *NE*, representing prompt-wise fine-grained uncertainty; and *SS*, representing response-wise fine-grained uncertainty.

### 3.1 Frequency As the Rankings Proxy

A straightforward way is to approximate real predictive probabilities through a sufficiently large number of samplings. However, as we show in Lemma 3.1, a minimum of 9,604 samples is required to achieve a 95% confidence level with a 1% margin of error. Therefore, the implementation is impractical due to computational constraints.

**Lemma 3.1** (Minimum Sample Size for Confident Probability Estimation)**.** *Let* $freq(Y_i)$ *be the absolute frequency of outcome* $Y_i$ *in the sampling,* $N_{total}$ *be the total number of samplings,* $p_i$ *be the desired estimated probability,* $\epsilon$ *be the estimation error, and* $\delta$ *be the target confidence level. To determine the minimum sample size for confident probability estimation, for any given* $\epsilon > 0$ *and* $0 < \delta < 1$*, the following inequality must hold:*

$$P\left\{\left|\frac{freq(Y_i)}{N_{total}} - p_i\right| \leq \epsilon\right\} \geq \delta. \tag{3}$$

*Then, the minimum sample size* $N_{total}$ *satisfying Inequality 3 is given by:*

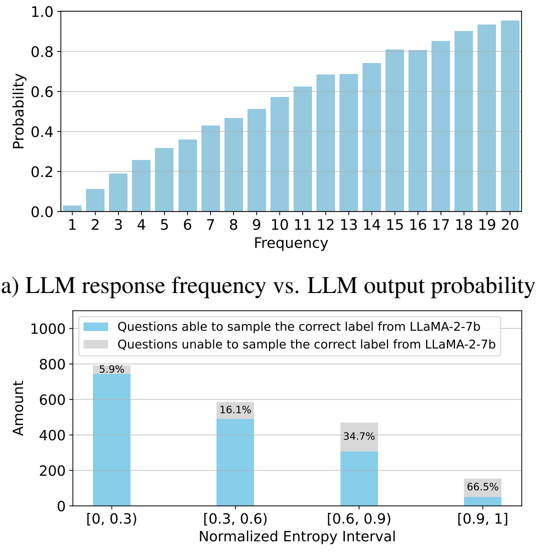
$$N_{total} \geq \left(\frac{u_{1-(1-\delta)/2}}{2\epsilon}\right)^2, \tag{4}$$

*where* $u_{1-(1-\delta)/2}$ *is the quantile of the standard normal distribution corresponding to the confidence level* $1-(1-\delta)/2$*. The proof of Lemma 3.1 is given in Appendix A.1.*

Since nonconformity measures are grounded in assessing the model's predictive uncertainty (Shafer and Vovk, 2008), the primary focus lies in the *rankings* of uncertainty inherent in nonconformity measures rather than the absolute values themselves. Further, self-consistency theory (Wang et al., 2022; Li et al., 2022) states that a repetitively sampled response is viewed as a form of consistency linked to higher confidence in the response. To empirically validate this intuition, we randomly select 2000 questions from the TriviaQA dataset (Joshi et al., 2017). We conduct 20 samplings from the Llama-2-7b model (Touvron et al., 2023), extracted logits, and subsequently computed model output probabilities. The observed results depicted in Figure 2a indicate a direct positive correlation between response frequency and average real probability. As the response frequency climbs, there is a corresponding increase in the average real probability, suggesting a growing level of confidence and certainty in the model's responses. Therefore, we propose to use frequency as the proxy of probability ranking. It is defined as

$$F(\hat{y}_a^{(i)}, m) = \frac{\tilde{p}[\hat{y}_a^{(i)}]}{m}, \tag{5}$$

where $\tilde{p}$ represents the empirical absolute frequency, $\hat{y}_a^{(i)}$ is the $a$-th non-repeated sampled response for $i$-th prompt, $m$ is the sampling quantity from LLMs for each prompt. However, only using response frequency as nonconformity measures results in the concentration of nonconformity scores on certain values. This issue makes it challenging to discern nonconformity differences among responses with the same scores, rendering ineffective calibration in CP.



(a) LLM response frequency vs. LLM output probability.

(b) NE vs. the proportion of prompts unable to sample correct labels from Llama-2-7b.

Figure 2: Empirical findings with TriviaQA dataset.

### 3.2 Fine-grained Uncertainty Notions

To resolve the concentration issue, we propose two fine-grained uncertainty measures. Firstly, inspired by self-consistency theory (Wang et al., 2022; Li et al., 2022), we incorporate NE, a prompt-wise fine-grained uncertainty notion, to mitigate the concentration issue across different prompts. NE is a measure of the uncertainty or diversity in the model's predictions when generating responses to a given prompt. It is defined as

$$H(x^{(i)}|\{\hat{y}_j^{(i)}\}_{j=1}^{m}) = |\frac{\sum_{a=1}^{n} \tilde{F}(\hat{y}_a^{(i)}) \log(\tilde{F}(\hat{y}_a^{(i)}))}{\log m}|, \quad (6)$$

where $x^{(i)}$ is the $i$-th instance of the prompt dataset, $m$ is the number of sampled responses, $n$ is the number of non-repeated responses, $\hat{y}_j^{(i)}$ is the $j$-th sampled response. Following experiments in Section 3.1, we show that as NE increases, the number of unanswered questions also increases (Figure 2b), indicating a rise in uncertainty.

Secondly, to address concentration issues within a prompt, we introduce SS as a response-wise fine-grained uncertainty measure. This metric semantically assesses the similarity between each non-top-1 response and the top-1 response within a prompt. Intuitively, when two non-top-1 responses share the same frequency, the one more semantically similar to the top-1 response is more likely to express high confidence and low uncertainty. We use the cosine similarity to express SS. It is defined as

$$SS(\hat{y}_a^{(i)}, P_{\text{highest}}^{(i)}) = \frac{\mathbf{v}(\hat{y}_a^{(i)}) \cdot \mathbf{v}(P_{\text{highest}}^{(i)})}{\|\mathbf{v}(\hat{y}_a^{(i)})\| \cdot \|\mathbf{v}(P_{\text{highest}}^{(i)})\|}, \quad (7)$$

where $\mathbf{v}(x)$ is the vector representation of $x$, $P_{\text{highest}}^{(i)}$ is the response having the highest frequency for $i$-th prompt. However, if the response to be measured is the one with the highest frequency, we do not consider SS with itself.

### 3.3 CP for LLMs Without Logit-Access

Considering both the coarse-grained and fine-grained uncertainty notions, the final nonconformity score function of LofreeCP is defined as

$$\begin{aligned} N^{(i)} = &-F(\hat{y}_a^{(i)}, m) + \lambda_1 \cdot H(x^{(i)}|\{\hat{y}_j^{(i)}\}_{j=1}^{m}) \\ &- \lambda_2 \cdot SS(\hat{y}_a^{(i)}, P_{\text{highest}}^{(i)}), \end{aligned} \quad (8)$$

where $\lambda = (\lambda_1, \lambda_2)$ representing a hyperparameter configuration controls the balance between the coarse-grained and fine-grained uncertainty notions. LofreeCP has the coverage guarantee:

**Proposition 3.2** (Coverage guarantee of LofreeCP)**.** *Suppose $(X_i, Y_i)_{i=1,\ldots,n}$ and $(X_{test}, Y_{test})$ are i.i.d. Let $C_{1-\alpha}(X_{test})$ be defined as in Step 3. Then we have the coverage guarantee:*

$$P\left\{Y_{test} \in C_{1-\alpha}\left(X_{test}\right)\right\} \geq 1 - \alpha,$$

*where $\alpha \in (0, 1)$ denotes the desired error rate. The proof of the coverage guarantee of LofreeCP is provided in Appendix A.2.*

LofreeCP consists of three stages: calibration, validation, and testing. The calibration stage aims to find the quantile based on the desired error rate. We sample $m$ responses from the LLM for each prompt and store them in a response pool. Then, we obtain the nonconformity scores of the true labels with the following rules: if the true label exists in the pool, we use the nonconformity measures from Equation 8 to calculate its nonconformity score; otherwise, we set the nonconformity score as $\infty$ to signify that it is nearly impossible to for the LLM to generate the true response. After obtaining all nonconformity scores of the calibration set, we find the quantile based on the desired error rate. We use this quantile as a threshold value for both the validation and test stages.

We then use the validation set to choose the optima hyperparameter configuration $\lambda = (\lambda_1, \lambda_2)$. Subsequently, we conduct evaluations on the test set using the chosen configuration. Both stages follow identical sampling steps to the calibration, traversing all responses and calculating the nonconformity scores. We preserve the responses whose nonconformity score is less than the threshold in our final prediction set. The pseudocode of the LofreeCP method is provided in Appendix B.10.

## 4 Experiments

### 4.1 Experimental Setup

**Backbone LLMs and Evaluation Tasks.** Since we need to compare LofreeCP with logit-based methods, from where logits can be retrieved directly, we consider different open-source LLMs, including Llama-2-7B, Llama-2-13B, WizardLM-v1.2(13b) (Xu et al., 2023) and Vicuna-v1.5(7b) (Chiang et al., 2023) models as our backbone models. Note that our method uses these LLMs as if they were API-only LLMs, i.e., it assumes no access to any internal information of LLMs. We use both open-ended Question-Answering (QA) and close-ended Multi-Choice Question-Answering (MCQ) tasks for evaluation.

**Datasets.** We use standard benchmarking datasets TriviaQA and MMLU (Hendrycks et al., 2020), following (Kumar et al., 2023) and (Quach et al., 2023). We also include the WebQuestions benchmark (Berant et al., 2013). For QA, we use the TriviaQA dataset, which consists of trivia questions spanning a wide range of topics such as history and science, and the WebQuestions dataset, which is focused on questions asked by users on a search engine. MMLU dataset, covering 57 subjects (e.g., mathematics, history), is used for MCQ. We focus on a subset of 16 subjects out of the total 57, as in Kumar et al. (2023).

**Baselines.** Baselines include methods without logit-access and those based on logit:

- **Top-K$_{\text{white}}$.** A logit-based non-CP method without coverage guarantee, which includes responses with the first $K$ highest probabilities for each prompt in the prediction set.
- **Standard Split Conformal Prediction (SCP)** (Vovk et al., 2005). A logit-based CP method, which follows the steps shown in Section 2.
- **Sorted Adaptive Prediction Sets (SAPS)** (Huang et al., 2023). A logit-based CP method, which uses the highest probability and replaces other probabilities with some weighted values to mitigate the miscalibration issue.
- **Top-K$_{\text{black}}$.** A non-CP method without logit-access and coverage guarantee, which includes responses with the first K highest frequency for each prompt in the prediction set.
- **Conformal Language Modeling (CLM)** (Quach et al., 2023). The state-of-the-art logit-based CP method, which uses the general risk control framework. This baseline is only used in QA as it is not applied to MCQ.

**Metrics.** We use following metrics for evaluation (Angelopoulos and Bates, 2021):

- Empirical Coverage Rate (ECR) assesses whether the conformal procedure has the correct coverage with the theoretical guarantee.
- Size-Stratified Coverage (SSC) (Angelopoulos et al., 2020) assesses the worst coverage rate of each bin among different set sizes.
- Average Prediction Set Size (APSS) assesses the efficiency of CP. We expect the APSS of an efficient CP method to be small.

### 4.2 Results for QA

We perform QA using TriviaQA and WebQuestions datasets. The results for Llama-2-13b are reported in Tables 1-2, those for Llama-2-7b are shown in the sensitivity analysis of Section 4.5 and those for WizardLM-v1.2(13b) and Vicuna-7b-v1.5 can be found in Appendix D. In Table 1, the LofreeCP method excels on TriviaQA across all error rate settings, outperforming the second-best method, CLM, by 7.7% in terms of APSS at an error rate of 0.25. Regarding SSC, our LofreeCP method surpasses the second-best method, First-K$_{\text{white}}$, by 1.6%. In Table 2, our method demonstrates superior performance on WebQuestions in most settings. For instance, at an error rate of 0.45, our LofreeCP method outperforms the second-best method, CLM, by 11.6% in terms of APSS. Regarding SSC, we outperform the second-best method, SCP, by 4.3%. WizardLM-v1.2(13b) and Vicuna-7b-v1.5 exhibit similar trends to Llama-2-13b.

The smallest APSS indicates that our method can produce the most efficient prediction sets. The highest SSC indicates that our method is attentive to the conditional coverage rate, achieving well-calibrated uncertainty estimates within diverse size categories. The rationale behind the observed superior performance is that our nonconformity measure can capture the coarse-grained uncertainty of responses and effectively optimize nonconformity through fine-grained considerations, thereby miti-

Table 1: Results for TriviaQA using Llama-2-13b: Among all baselines, only *First-K$_{white}$* and *First-K$_{black}$* are non-CP-based, while the rest are CP-based methods. In the results, **bold** indicates that the method produces the best performance among all methods; ✗ denotes that the method fails to produce the set with the desired error rate.

| Methods | Logit-Access | Error Rate | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.2 | | | 0.25 | | | 0.3 | | | 0.35 | | | 0.4 | | | 0.45 | | |
| | | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ |
| First-K$_{\text{white}}$ | ✓ | 82.1 | 76.6 | 3.39 | 76.1 | 72.9 | 1.90 | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | 62.4 | 62.5 | 1.00 | ✗ | ✗ | ✗ |
| CLM | ✓ | 80.2 | 73.4 | 2.29 | 75.2 | 69.1 | 1.55 | 70.1 | 68.3 | 1.28 | 65.0 | 69.3 | 0.96 | 60.1 | 72.7 | 0.81 | 55.2 | 83.3 | 0.70 |
| SCP | ✓ | 80.3 | 75.7 | 2.25 | 75.1 | 70.0 | 1.59 | 70.3 | 74.5 | 1.21 | 65.1 | 76.4 | 1.02 | 60.3 | 75.7 | 0.85 | 55.3 | 82.5 | 0.74 |
| SAPS | ✓ | 80.0 | 77.9 | 2.74 | 75.1 | 64.2 | 1.80 | 70.0 | 49.4 | 1.55 | 65.1 | 57.4 | 1.28 | 60.1 | 70.7 | 0.85 | 55.1 | 76.5 | 0.72 |
| First-K$_{\text{black}}$ | ✗ | 80.1 | 76.8 | 2.70 | 76.4 | 72.2 | 1.90 | ✗ | ✗ | ✗ | 66.5 | 66.5 | 1.00 | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| **LofreeCP (Ours)** | ✗ | 80.1 | **79.0** | **2.19** | 75.3 | **74.5** | **1.43** | 70.3 | **76.7** | **1.08** | 65.1 | **78.5** | **0.90** | 60.0 | **81.0** | **0.75** | 55.2 | **84.1** | **0.66** |

Table 2: Results for WebQuestions using Llama-2-13b.

| Methods | Logit-Access | Error rate | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.35 | | | 0.4 | | | 0.45 | | | 0.5 | | |
| | | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ |
| First-K$_{\text{white}}$ | ✓ | 66.4 | 57.5 | 6.18 | 61.6 | 58.1 | 3.81 | 57.5 | 55.0 | 2.91 | 50.6 | 49.0 | 1.97 |
| CLM | ✓ | 65.3 | 50.5 | **4.54** | 60.5 | 52.9 | 2.86 | 55.0 | 51.6 | 1.81 | 50.1 | 56.8 | 1.27 |
| SCP | ✓ | 65.1 | 46.7 | 4.61 | 61.6 | 49.3 | 3.01 | 55.2 | 55.8 | 2.02 | 50.2 | 57.8 | 1.39 |
| SAPS | ✓ | 65.2 | 46.2 | 5.19 | 60.6 | 56.2 | 3.39 | 55.5 | 37.7 | 2.40 | 50.8 | 21.7 | 1.86 |
| First-K$_{\text{black}}$ | ✗ | 65.1 | 54.9 | 6.20 | 60.0 | 55.3 | 3.78 | 56.9 | 54.4 | 2.91 | 53.7 | 52.4 | 1.97 |
| **LofreeCP (Ours)** | ✗ | 65.1 | **61.1** | 5.33 | 60.0 | **60.0** | **2.68** | 55.1 | **60.1** | **1.60** | 50.3 | **59.9** | **1.06** |

gating the inherent miscalibration issue in LLMs.

### 4.3 Ablation Study

To demonstrate the impact of our fine-grained uncertainty notions (NE and SS) on mitigating the concentration issues, we conduct a series of ablation studies using the TriviaQA dataset with a sampling quantity of 20. We compare LofreeCP with its different variants: we remove one fine-grained notion at a time (Freq&SS, removing the NE notion; and Freq&NE, removing the SS notion), and finally remove both fine-grained notions (Freq-Only). We report APSS and ECR, the direct indicators of the concentration issue, in Figure 3.

**Impact of Concentration Issue.** As introduced in Section 3, the concentration issue occurs when the nonconformity score is concentrated on certain values. When we use the frequency-only variant (Freq-Only), this issue can be observed in all error rate settings, as shown in Figure 3: Freq-Only has the largest APSS and the most conservative ECR. Due to its coarse-grained uncertainty notion, Freq-Only tends to generate similar nonconformity scores clustered into several groups, making it hard to differentiate granular uncertainties to produce efficient prediction sets.

**Full Method Mitigates Concentration Issue.** We further observe that the concentration issue is mitigated in all error rate settings by incorporating fine-grained notions (NE & SS). For example, at an error rate of 0.2, Freq-Only exhibits an APSS of nearly 6.5, while the full method LofreeCP has an APSS of 4.27, resulting in a drop of more than 23%. The method including only SS or NE also mitigates the concentration issue to some extent, while the full method performs the best in terms of APSS and ECR. The results suggest that NE and SS both have a significant impact on improving the efficiency of prediction sets by mitigating concentration issues of nonconformity scores.

### 4.4 Results for MCQ

In addition to open-ended tasks, e.g. QA, LofreeCP is also effective at close-ended tasks that can be converted into a generation pipeline, e.g. MCQ. We conduct MCQ experiments on the MMLU dataset using Llama-2-13b with a sampling quantity of 20. We present the results in Figure 4.[1] LofreeCP exhibits superior performance. When compared with SCP and SAPS across all 16 subjects, LofreeCP

[1] We omit the results from top-K methods as they exhibit much larger APSS than other methods for MCQ.

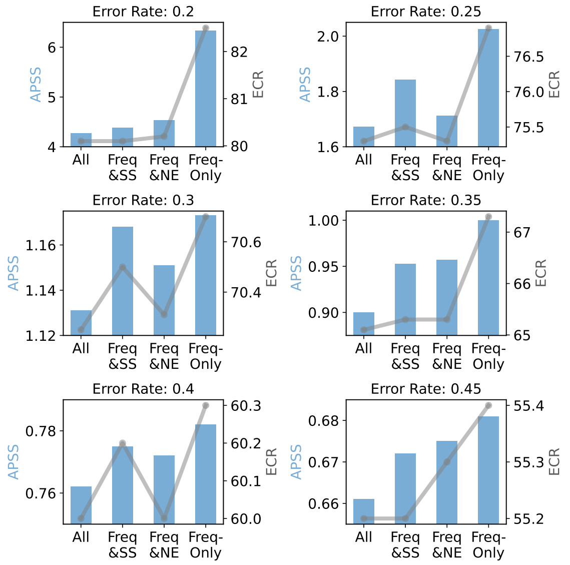


Figure 3: Ablation study. The blue bar chart represents APSS, while the gray line represents ECR.

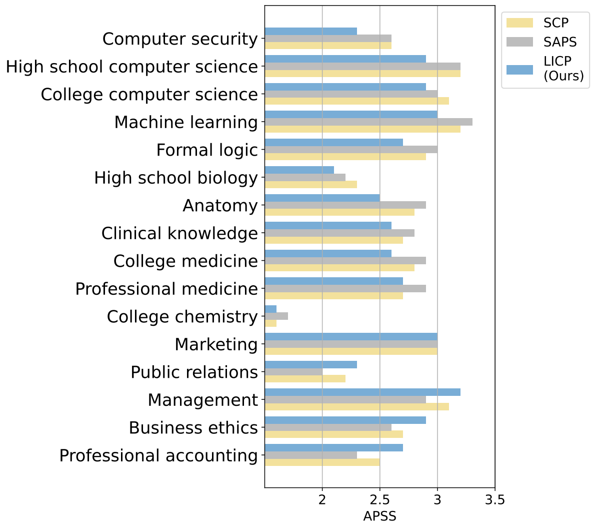


Figure 4: Results on MMLU for MCQ task, with the error rate of 0.2. Our method and baselines are applied individually to each of the 16 subjects.

achieves the best performance in 9 subjects and ties for the best in subjects of professional medicine, college chemistry, and marketing, resulting in the overall best performance in 12 out of 16 subjects. In contrast, SCP only ties for the best in 3 subjects. SAPS achieves the solo best performance in 3 subjects and ties for the best in 1 subject.

An intriguing observation is related to subjects in the business and management (B&M) category (e.g., marketing and public relations). When using LofreeCP method, these subjects show slightly larger APSS than the two logit-based methods, SCP and SAPS. This suggests that the logits for responses to B&M questions predicted by the Llama-2-13b model are better calibrated than the remaining subjects from the Science, Technology, Engineering, and Mathematics (STEM) category. Our LofreeCP method mitigates the model miscalibration issue by refraining from directly using logits.

## 4.5 Sensitivity Analyses

**BackBone Models.** To investigate the influence of different backbone models on the performance of LofreeCP, we conduct experiments using Llama-2-7b and Llama-2-13b with a sampling quantity of 20. Results of SSC and APSS are shown in Figure 5. We observe that better performance of APSS and SSC in the 13b setting than in the 7b setting. We believe this is because Llama-2-13b is more powerful than Llama-2-7b, and produces more confident and calibrated responses, thereby providing more efficient prediction sets. Results for Vicuna-v1.5(7b) are provided in Appendix D, indicating that Vicuna-v1.5(7b) can only produce prediction sets with higher error rates compared to Llama-2 backbones. This is because Vicuna-v1.5(7b) is less powerful for these two datasets. This demonstrates that CP performance for LLMs is largely dependent on the performance of the backbone models.

**Sampling Quantity** The sampling quantity regulates the number and types of sampled responses acquired from LLMs, thereby influencing frequency, NE and SS. We vary the sampling quantity from 10 to 40 on the TriviaQA dataset using Llama-2-13b, incrementing by 5 each time. Results shown in Figure 6 suggest that a larger sampling quantity tends to present better performance w.r.t. efficiency. This is because, with a higher sampling quantity, the frequency notion more accurately represents response rankings. Of particular interest is that, at an error rate of 0.2, the sampling quantity of 15 exhibits inferior performance compared to the quantity of 10. We hypothesize it is because a sampling quantity

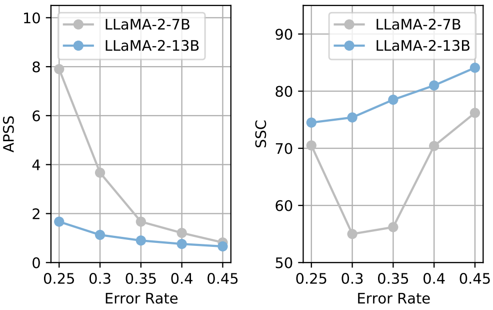


Figure 5: Results of the sensitivity analysis for different backbone models: Llama-2-7b and Llama-2-13b.

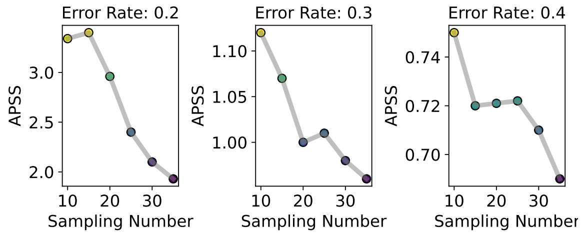


Figure 6: Results of sensitivity analysis of sampling quantity.

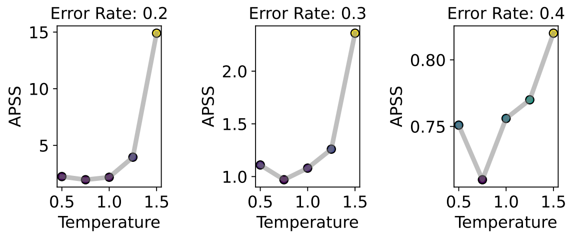


Figure 7: Results of sensitivity analysis of temperature.

of 15 remains insufficient to adequately represent rankings meanwhile introducing more non-robust randomness in responses. In addition, we observe a larger impact of the sampling quantity on APSS when a small error rate guarantee is required.

**Temperature Scaling.** The temperature (Hinton et al., 2015) in LLMs adjusts the randomness in generated outputs by scaling logits during the softmax operation. Higher temperatures boost the diversity of the output, which may further affect the performance of LofreeCP. In this experiment, we vary temperatures[2] (0.5, 0.75, 1.0, 1.25, and 1.5) in the Llama-2-13b model. Results for the TriviaQA dataset are presented in Figure 7. The smallest (best) APSS is observed at a temperature of 0.75. We observe an overall growing trend as the temperature increases from 0.75 to 1.50. This indicates that excessive diversity can result in uncertain and suboptimal predictions. The decline from 0.50 to 0.75 implies that too much determinism may hurt CP efficiency due to a lack of randomness and diversity. We also note a significant temperature influence on APSS when aiming for low error rates.

## 5 Related Work

**Conformal Prediction for NLP.** CP has already found diverse applications in NLP, e.g., text infilling and part-of-speech prediction Dey et al. (2021), sentiment analysis Maltoudoglou et al. (2020), and Automatic Speech Recognition Ernez et al. (2023). In the application of CP to LLMs, existing methods are predominantly logit-based. For instance, Kumar et al. (2023) apply standard CP (Vovk et al., 2005) to Llama-2-13b (Touvron et al., 2023) for the MCQ task by computing softmax scores of token logits for options to measure nonconformity. Similarly, Quach et al. (2023) extend CP to LLMs using the general risk control framework (Angelopoulos et al., 2021). However, recent studies have pointed out that relying solely on logits may be flawed (Xi et al., 2024) due to hallucinations in LLMs (LeCun, 2023). Consequently, there is ongoing research aiming to reduce reliance on logits. Huang et al. (2023) propose to use the highest probability and replace other probabilities with weighted values. All these methods involve the utilization of logits.

**Uncertainty Estimation in LLMs.** Recent developments in LLMs have highlighted the importance of estimating their uncertainty. While there has been significant research on uncertainty in NLP (Van Landeghem et al., 2022; Ulmer et al., 2022; Wang et al., 2024; Da et al., 2024), several methods exist to estimate the confidence of LLMs, including Deep Ensemble methods (Lakshminarayanan et al., 2017), Monte Carlo dropout (Gal and Ghahramani, 2016), Density-based estimation (Yoo et al., 2022), Confidence learning (DeVries and Taylor, 2018), as well as approaches based on logits. However, recent studies highlight concerns that LLMs may generate unfaithful and nonfactual content (Maynez et al., 2020; Yang et al., 2024). Additionally, logits of LLMs often exhibit overconfidence when producing incorrect answers, indicating that logits alone may not be entirely reliable for studying uncertainty (Desai and Durrett, 2020; Miao et al., 2021; Vasconcelos et al., 2023).

## 6 Conclusion

We study the critical problem of CP for API-only LLMs without logit-access. We propose a novel solution to define the nonconformity score function by leveraging uncertainty information from diverse sources. In particular, under a limited sampling budget, we first use the response frequency as the coarse-grained proxy of uncertainty levels. We then propose two fine-grained uncertainty notions (NE and SS) to further distinguish uncertainty at a nuanced level. Our proposed approach does not rely on model logits and can alleviate the known miscalibration issue when using logits. Experiments demonstrate the superior performance of our approach compared to logit-based and logit-free

[2] Temperature ranges between 0 and 2.

baselines. Our work opens up a new avenue to uncertainty estimation in LLMs without logit-access.

## Limitations

Our approach encounters a common limitation of open-ended Natural Language Generation (NLG) tasks: the unbounded output space. In our work, we address this challenge by sampling a fixed number of times for every prompt from LLMs to achieve a comprehensive output space, but we recognize the potential for more effective and convincing approaches to handle this issue within the framework of CP. Secondly, another future direction is to expand our CP method to non-exchangeability scenarios, particularly in NLG domains, where calibration and test sets may not adhere strictly to the assumption of being independent and identically distributed (i.i.d.). Finally, due to financial constraints, we do not evaluate our approach on several proprietary LLMs (e.g., GPT 4) that allow users to obtain token log probabilities. Thus future work can validate our method on these models.

## Acknowledgements

Lu Cheng is supported by the National Science Foundation (NSF) Grant #2312862, NIH #R01AG091762, and a Cisco gift grant.

Hongwei Wang and Jiayuan Su are supported in part by Zhejiang Provincial Natural Science Foundation of China (LDT23F02023F02).

## References

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. 2020. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Anastasios N Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. 2021. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

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

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers.

Terrance DeVries and Graham W. Taylor. 2018. Learning confidence for out-of-distribution detection in neural networks.

Neil Dey, Jing Ding, Jack Ferrell, Carolina Kapper, Maxwell Lovig, Emiliano Planchon, and Jonathan P Williams. 2021. Conformal prediction for text infilling and part-of-speech prediction.

Fares Ernez, Alexandre Arnold, Audrey Galametz, Catherine Kobus, and Nawal Ould-Amer. 2023. Applying the conformal prediction paradigm for the uncertainty quantification of an end-to-end automatic speech recognition model (wav2vec 2.0). In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pages 16–35. PMLR.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. 2023. Conformal prediction for deep classifier via label ranking. *arXiv preprint arXiv:2310.06430*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Yuko Kato, David MJ Tax, and Marco Loog. 2023. A review of nonconformity measures for conformal prediction in regression. *Conformal and Probabilistic Prediction with Applications*, pages 369–383.

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*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Y LeCun. 2023. Do large language models need sensory grounding for meaning and understanding? In *Workshop on Philosophy of Deep Learning, NYU Center for Mind, Brain, and Consciousness and the Columbia Center for Science and Society*.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2022. On the advance of making language models better reasoners. *arXiv preprint arXiv:2206.02336*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Lysimachos Maltoudoglou, Andreas Paisios, and Harris Papadopoulos. 2020. Bert-based conformal predictor for sentiment analysis. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pages 269–284. PMLR.

James Manyika and Sissie Hsiao. 2023. An overview of bard: an early experiment with generative ai. *AI. Google Static Documents*, 2.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. 2021. Prevent the language model from being overconfident in neural machine translation.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. *arXiv preprint arXiv:1508.05154*.

OpenAI. 2023. GPT-4v(ision): technical work.

René Peinl and Johannes Wirth. 2023. Evaluation of medium-large language models at zero-shot closed book generative question answering. *arXiv preprint arXiv:2305.11991*.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. 2023. Conformal language modeling. *arXiv preprint arXiv:2306.10193*.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. 2019. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234.

Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022. Exploring predictive uncertainty and calibration in NLP: A study on the impact of method & data scarcity. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2707–2735, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jordy Van Landeghem, Matthew Blaschko, Bertrand Anckaert, and Marie-Francine Moens. 2022. Benchmarking scalable predictive uncertainty in text classification. *IEEE Access*, 10:43703–43737.

Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q. Vera Liao, and Jennifer Wortman Vaughan. 2023. Generation probabilities are not enough: Exploring the effectiveness of uncertainty highlighting in ai-powered code completions.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Fangxin Wang, Lu Cheng, Ruocheng Guo, Kay Liu, and Philip S Yu. 2023. Equal opportunity of coverage in fair regression. *Advances in Neural Information Processing Systems*, 36.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2024. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. *arXiv preprint arXiv:2407.00499*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Huajun Xi, Jianguo Huang, Lei Feng, and Hongxin Wei. 2024. Does confidence calibration help conformal prediction? *arXiv preprint arXiv:2402.04344*.

Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Self-improving for zero-shot named entity recognition with large language models. *arXiv preprint arXiv:2311.08921*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Shu Yang, Jiayuan Su, Han Jiang, Mengdi Li, Keyuan Cheng, Muhammad Asif Ali, Lijie Hu, and Di Wang. 2024. Dialectical alignment: Resolving the tension of 3h and security threats of llms. *arXiv preprint arXiv:2404.00486*.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. Detection of word adversarial examples in text classification: Benchmark and baseline via robust density estimation.

# A Theoretical Proofs

## A.1 Proof of Lemma 3.1

*Proof.* When $N_{total}$ is sufficiently large, the Lindeberg–Lévy central limit theorem yields the following equation:

$$\frac{\frac{freq(Y_i)}{N_{total}} - p_i}{\sqrt{p_i(1-p_i)/N_{total}}} \sim N(0,1),$$

From this, we conclude that

$$P\left\{\left|\frac{\frac{freq(Y_i)}{N_{total}} - p_i}{\sqrt{p_i(1-p_i)/N_{total}}}\right| \leq u_{1-(1-\delta)/2}\right\} \geq \delta.$$

Approximately replacing $p_i$ in the denominator with $\frac{\text{freq}(Y_i)}{N_{total}}$, we obtain

$$P\left\{\left|\frac{\frac{freq(Y_i)}{N_{total}} - p_i}{\sqrt{\frac{\text{freq}(Y_i)}{N_{total}}(1-\frac{\text{freq}(Y_i)}{N_{total}})/N_{total}}}\right| \leq u_{1-(1-\delta)/2}\right\}$$
$$\geq \delta.$$

This equation is equivalent to

$$P\{-u_{1-(1-\delta)/2} \leq \frac{\frac{freq(Y_i)}{N_{total}} - p_i}{\sqrt{\frac{\text{freq}(Y_i)}{N_{total}}(1-\frac{\text{freq}(Y_i)}{N_{total}})/N_{total}}}$$
$$\leq u_{1-(1-\delta)/2}\} \geq \delta.$$

We can then reformulate the above equation as:

$$P\{\frac{freq(Y_i)}{N_{total}}$$
$$- u_{1-(1-\delta)/2} \cdot \sqrt{\frac{\frac{freq(Y_i)}{N_{total}}(1-\frac{freq(Y_i)}{N_{total}})}{N_{total}}}$$
$$\leq p_i \leq \frac{freq(Y_i)}{N_{total}}$$
$$+ u_{1-(1-\delta)/2} \cdot \sqrt{\frac{\frac{freq(Y_i)}{N_{total}}(1-\frac{freq(Y_i)}{N_{total}})}{N_{total}}}\}$$
$$\geq \delta.$$

In the left part of this equation, $\frac{\text{freq}(Y_i)}{N_{total}}$ represents the absolute frequency, $p_i$ represents the desired estimated probability, and $u_{1-(1-\delta)/2} \cdot \sqrt{\frac{\frac{freq(Y_i)}{N_{total}}(1-\frac{freq(Y_i)}{N_{total}})}{N_{total}}}$ is the error term between them. Recall that we aim to ensure:

$$P\left\{\left|\frac{freq(Y_i)}{N_{\text{total}}} - p_i\right| \leq \epsilon\right\} \geq \delta.$$

Therefore, we need to guarantee:

$$u_{1-(1-\delta)/2} \cdot \sqrt{\frac{\frac{freq(Y_i)}{N_{total}}(1-\frac{freq(Y_i)}{N_{total}})}{N_{total}}} \cdot 2 \leq 2\epsilon.$$

This implies that we must control the error term to not exceed our predetermined estimation error. Note that the left part of this equation reaches its maximum value when $freq(Y_i) = \frac{1}{2}$. Hence, to achieve this, we only require:

$$\sqrt{\frac{1/4}{N_{total}}} \cdot u_{1-(1-\delta)/2} \cdot 2 \leq 2\epsilon.$$

This simplifies to

$$N_{total} \geq \left(\frac{u_{1-(1-\delta)/2}}{2\epsilon}\right)^2$$

□

### A.2 Proof of Proposition 3.2

*Proof.* Let $N$ denote the nonconformity measures of the calibration set $(X_i, Y_i)_{i=1,\ldots,n}$, and let $\alpha_1$ and $\alpha_2$ be the desired error rates, where $\alpha_1 > \alpha_2$. As indicated in Step 2, we have $\hat{q_1} \leq \hat{q_2}$. Given $C(X_{\text{test}}) = \{Y : N(X_{\text{test}}, Y) \leq \hat{q}\}$, it follows that $C_{1-\alpha_1}(X) \subseteq C_{1-\alpha_2}(X)$. Consequently, the nesting property, as defined in Equation 1, is satisfied. Therefore, Proposition 3.2 holds. □

## B Implementation Details

### B.1 Dataset

The TriviaQA benchmark (available at https://nlp.cs.washington.edu/triviaqa/ or can be accessed from Hugging Face at https://huggingface.co/datasets/trivia_qa) and the WebQuestions benchmark (available at worksheets.codalab.org or can be accessed from Hugging Face at https://huggingface.co/datasets/web_questions) are employed for QA. Both datasets operate within a closed-book setting, where LLMs refrain from using supporting text when answering questions.

The MMLU benchmark (can be accessed from Hugging Face at https://huggingface.co/datasets/lukaemon/mmlu) is designed for MCQ, which covers 57 subjects across STEM, the humanities, the social sciences, and more. For our MCQ experiments, we leverage the dataset containing 16 subjects from the MMLU: 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, professional accounting.

For the TriviaQA dataset, we randomly select 10,000 question-answer pairs. Similarly, for the WebQuestions dataset, we randomly select 5,000 question-answer pairs. Regarding the MMLU dataset, we use all available data for each of the 16 subjects. Across all three datasets, we apply the same splitting strategy: 50% of the data serves as the calibration set, 25% as the validation set, and 25% as the test set for each trial.

### B.2 Backbone LLMs

We utilize the Hugging Face API to access open-source LLMs in our experiments, including Llama-2-7B (accessible at huggingface.co/meta-llama/Llama-2-7b-hf), Llama-2-13B (accessible at huggingface.co/meta-llama/Llama-2-13b-hf), WizardLM-v1.2(13b) (accessible at huggingface.co/WizardLM/WizardLM-13B-V1.2), and Vicuna-v1.5(7b) (accessible at huggingface.co/lmsys/vicuna-7b-v1.5). Access to Llama-2-7b and Llama-2-13b requires requesting approval via the Meta website (https://llama.meta.com/). Upon approval, access to these resources will be granted.

### B.3 Length-Normalization

We use length normalization (Wu et al., 2016) on logits to obtain response probability/likelihood:

$$p(x, y_k) = \text{exp}(\frac{\text{log}p_\theta(y_k|x)}{lp(y_k)})$$

where

$$lp(y) = \frac{(5 + |y|)^{0.6}}{(5 + 1)^{0.6}}$$

### B.4 Evaluation

We extract an answer by analyzing the text until we encounter the first line break, comma, or period. This implies that in the dataset, we will disregard data whose answers contain line breaks, commas, or periods. Following this, we standardize the answers by converting them to lowercase, removing articles, punctuation, and duplicate whitespace. The generated answers are then evaluated using the exact match metric, where an answer is considered correct only if it exactly matches the provided answer. These guidelines align with those described in Quach et al. (2023).

For SSC, We focus exclusively on bins with a set size greater than 0 and a sample number exceeding 10% of the total test samples. This is because bins with a size of 0 and fewer samples lack reliability for coverage measurement.

### B.5 LLMs Parameters

We employ the default Transformer generative LMs parameters for our experiments, using default standard sampling with do_sample set to True, top_k set to 0, top_p set to 1, and Temperature set to 1, except when conducting model hyperparameter-tuning experiments. In such hyperparameter-tuning cases, we explicitly mention the parameters in main body of the paper.

## B.6 Semantic Similarity

The measure of semantic similarity was established leveraging the FastText model available within the gensim package. The configuration parameters were carefully selected, defining a vector size of 200 and imposing a minimum count threshold of 1 to ensure robustness and inclusivity in the model's representations.

## B.7 Experiment trails

We conduct 50 trials for all experiments, then average the results to eliminate randomness during the calibration.

## B.8 Error Rate Settings

We do not apply the same error rate settings across different models or datasets. This is because each model varies in its coverage ability for the same dataset. Likewise, the same model doesn't possess identical coverage abilities for different datasets. Therefore, we adjust error rate settings for different combinations of model and dataset accordingly.

## B.9 GPUs

We utilize six NVIDIA RTX 3090 graphics cards to support experiments.

## B.10 Pseudocode

We show the pseudocode in Method 1, where we do not explicitly display the repetitive process of using various hyperparameter configurations to determine the best one. In our actual implementations, we explore the range [0:0.05:2] for both $\lambda_1$ and $\lambda_2$. This range spans from 0 to 2, with each step incrementing by 0.05, thus covering values such as 0, 0.05, 0.1, 0.15, and so forth up to 2. Subsequently, we form different combinations to execute the calibration and validation stages. Ultimately, we utilize the best hyperparameter configurations for testing purposes.

**Method 1** LofreeCP method

```
Require: Prompt $x^{(i)}$, LLM $f_\theta$, response $\hat{y}_j^{(i)}$, current sampling number $j$, required sampling number $m$, response pool $P^{(i)}$, response with the highest frequency $P_{highest}^{(i)}$, semantic similarity between response $a$ and $b$: $S$(a,b)
 1: for $x^{(i)}$, $i = 1$ to $n$ do
        $P^{(i)} = \{\}$                                   ▷ Calibration stage starts
 2:     for $j = 1$ to $m$ do
            $\hat{y}_j^{(i)} \leftarrow f_\theta(x^{(i)})$      ▷ Sample response from LLM given the prompt
 3:         if $\hat{y}_j^{(i)}$ in $P^{(i)}$ then
                $\tilde{p}[\hat{y}_j^{(i)}]$ ++                ▷ Increment frequency for existing response
 4:         else
                $\tilde{p}[\hat{y}_j^{(i)}] = 1$               ▷ Initialize frequency for new response
 5:         end if
 6:     end for
 7:     Sort($P^{(i)}$)
 8:     Get $P_{highest}^{(i)}$                          ▷ Get the response with the highest frequency
 9:     if $y^{(i)}$ in $P^{(i)}$ then
            $N^{(i)} = \frac{\tilde{p}[\hat{y}_a^{(i)}]}{m} + \lambda_1 \cdot H(x^{(i)}|\{\hat{y}_j^{(i)}\}_{j=1}^{m})$ - $\lambda_2 \cdot S(\hat{y}_a^{(i)}, P_{highest}^{(i)})$
10:     else
            $N^{(i)} = \infty$                           ▷ Nonconformity measures
11:     end if
12: end for
13: $\hat{q}_\alpha$ = Quantile($\{N^{(1)}, N^{(2)}, ..., N^{(n)}\}, \frac{\lceil (n+1)(1-\alpha) \rceil}{n}$) ▷ Find quantile $\hat{q}_\alpha$   ▷ Calibration stage ends
14: for sampling same as 1 ~ 7 do                    ▷ Validation / Test stage starts
15:     for each $\hat{y}_\alpha^{(i)}$ in $P^{(i)}$ do
            $N_\alpha^{(i)} = \frac{P^{(i)}[\hat{y}_\alpha^{(i)}]}{m} + \lambda_1 \cdot H(x^{(i)}|\{\hat{y}_j^{(i)}\}_{j=1}^{m})$ - $\lambda_2 \cdot S(\hat{y}_\alpha^{(i)}, P_{highest}^{(i)})$
16:     end for
17:     $C(x_{\text{test}}^{(i)}) = \{\hat{y}_a^{(i)} : N_\alpha^{(i)} \leq \hat{q}\}$     ▷ Nonconformity measures
18: end for                                          ▷ Validation / Test stage ends
```

# C Prompts

## C.1 Few-shot Prompts of TriviaQA

We use the 32-shot question-answer pair prompts from the TriviaQA dev set, the same as those in Quach et al. (2023).

```
Answer these questions.
Q: Which American-born Sinclair won the
Nobel Prize for Literature in 1930?
A: Sinclair Lewis
Q: Where in England was Dame Judi Dench
born?
A: York
Q: In which decade did Billboard
magazine first publish an American hit
chart?
A: 30s
Q: From which country did Angola achieve
independence in 1975?
A: Portugal
Q: Which city does David Soul come from?
A: Chicago
Q: Who won Super Bowl XX?
A: Chicago Bears
Q: Which was the first European country
to abolish capital punishment?
A: Norway
Q: In which country did the widespread
use of ISDN begin in 1988?
```

```
A: Japan
Q: What is Bruce Willis' real first name?
A: Walter
Q: Which William wrote the novel Lord Of The Flies?
A: Golding
Q: Which innovation for the car was developed by Prince Henry of Prussia in 1911?
A: Windshield wipers
Q: How is musician William Lee Conley better known?
A: Big Bill Broonzy
Q: How is Joan Molinsky better known?
A: Joan Rivers
Q: In which branch of the arts is Patricia Neary famous?
A: Ballet
Q: Which country is Europe's largest silk producer?
A: Italy
Q: The VS-300 was a type of what?
A: Helicopter
Q: At which university did Joseph Goebbels become a doctor of philosophy?
A: Heidelberg
Q: Which prince is Queen Elizabeth II's youngest son?
A: Edward
Q: When did the founder of Jehovah's Witnesses say the world would end?
A: 1914
Q: Who found the remains of the Titanic?
A: Robert Ballard
Q: Who was the only Spice Girl not to have a middle name?
A: Posh Spice
Q: What are the international registration letters of a vehicle from Algeria?
A: DZ
Q: How did Jock die in Dallas?
A: Helicopter accident
Q: What star sign is Michael Caine?
A: Pisces
Q: Who wrote the novel Evening Class?
A: Maeve Binchy
Q: Which country does the airline Air Pacific come from?
A: Fiji
Q: In which branch of the arts does Allegra Kent work?
A: Ballet
Q: Banting and Best pioneered the use of what?
A: Insulin
Q: Who directed the movie La Dolce Vita?
A: Federico Fellini
Q: Which country does the airline LACSA come from?
A: Costa Rica
Q: Who directed 2001: A Space Odyssey?
A: Stanley Kubrick
Q: Which is the largest of the Japanese Volcano Islands?
A: Iwo Jima
Q: (Question)
A:
```

### C.2 Prompts of Webquestions

We also use 32-shot question-answer pair prompts from the Webquestions train set.

```
Answer these questions.
Q: What country is the Grand Bahama Island in?
A: Bahamas
Q: What two countries invaded Poland in the beginning of WW2?
A: Germany
Q: Which countries border the US?
A: Canada
Q: Where is Rome, Italy located on a map?
A: Rome
Q: What is Nina Dobrev's nationality?
A: Bulgaria
Q: What country does Iceland belong to?
A: Iceland
Q: What does Thai mean?
A: Language
Q: Who was Ishmael's mom?
A: Hagar
Q: What are the major cities in France?
A: Paris
Q: What city did Esther live in?
A: Susa
Q: What sport do the Toronto Maple Leafs play?
A: Ice Hockey
Q: What is Martin Cooper doing now?
A: Inventor
```

Q: What county is the city of Hampton, VA in?
A: Hampton
Q: What county is Heathrow Airport in?
A: London
Q: What type of car does Michael Weston drive?
A: Wishcraft
Q: What was Tupac’s name in Juice?
A: Bishop
Q: Who does Maggie Grace play in Taken?
A: Kim
Q: What style of music did Louis Armstrong play?
A: Jazz
Q: Where does Jackie French live?
A: Australia
Q: Where is Jack Daniels factory?
A: Tennessee
Q: What is Charles Darwin famous for?
A: Evolution
Q: Where to visit in N. Ireland?
A: Antrim
Q: What are dollars called in Spain?
A: Peseta
Q: Who plays Meg in Family Guy?
A: Mila Kunis
Q: Where did Martin Luther King get shot?
A: Memphis
Q: What was Nelson Mandela’s religion?
A: Methodism
Q: Who will win the 2011 NHL Stanley Cup?
A: Canada
Q: What is Henry Clay known for?
A: Lawyer
Q: What is the money of Spain called?
A: Euro
Q: Where are Sunbeam microwaves made?
A: Florida
Q: Where was Kennedy when he got shot?
A: Dallas
Q: Where did the Casey Anthony case take place?
A: Orlando
Q: (Question)
A:

### C.3 Prompts of MMLU

Each subject in MMLU uses similar prompts. We take the high school biology as examples.

Please engage in the multiple-choice question-answering task. You should generate the option (A, B, C, or D) you think is right. Examples are provided. (Select 8-shot randomly from other subjects)

This is a question from high school biology.
A piece of potato is dropped into a beaker of pure water. Which of the following describes the activity after the potato is immersed into the water?
(A) Water moves from the potato into the surrounding water.
(B) Water moves from the surrounding water into the potato.
(C) Potato cells plasmolyze.
(D) Solutes in the water move into the potato.
The correct answer is option: B.

You are the world’s best expert in high school biology. Reason step-by-step and answer the following question.
From the solubility rules, which of the following is true?
(A) All chlorides, bromides, and iodides are soluble
(B) All sulfates are soluble
(C) All hydroxides are soluble
(D) All ammonium-containing compounds are soluble
The correct answer is option:

# D Additional Results

## D.1 Ablation Study

Table 3: SCC Results of Ablation Study

| Error Rate | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 |
|---|---|---|---|---|---|---|
| Freq-Only | 77.1 | 72.9 | 75.3 | 77.2 | 79.4 | 81.7 |
| Freq + NE | 78.8 | 74.0 | 76.8 | 77.9 | 80.2 | 83.3 |
| Freq + SS | 78.2 | 74.7 | 76.6 | 78.7 | 80.0 | 82.9 |
| All (Ours) | 79.2 | 74.3 | 76.5 | 78.6 | 81.5 | 84.0 |

Table 4: Portion of Concentration

| Method | Portion of Concentration (%) |
|---|---|
| Freq-Only | 66.6 |
| Freq + NE | 45.5 |
| Freq + SS | 59.8 |
| All (Ours) | 35.1 |

## D.2 Sensitivity Experiments

More results regarding sampling quantity and temperature sensitivity are included in Figures 8-9 due to the page limit in the main body.

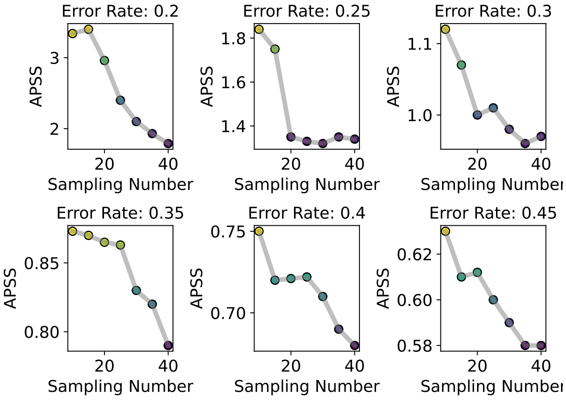


Figure 8: All results of the sensitivity analysis to variations in sampling quantity.

## D.3 Results for WizardLM-v1.2 (13B) and Vicuna-v1.5 (7B)

To save on computation costs, we use float16 precision (half-precision) for experiments in this section. We use standard sampling with sampling quantity of 30. Results for TriviaQA are shown in Table 5, for WebQuestions are shown in Table 7. Results for TriviaQA are shown in Table 6, for WebQuestions are shown in Table 8. Results for WizardLM-v1.2 (13B) and Vicuna-v1.5 (7B) consistently align with the main body results, demonstrating that the LofreeCP method mostly outperforms baselines.

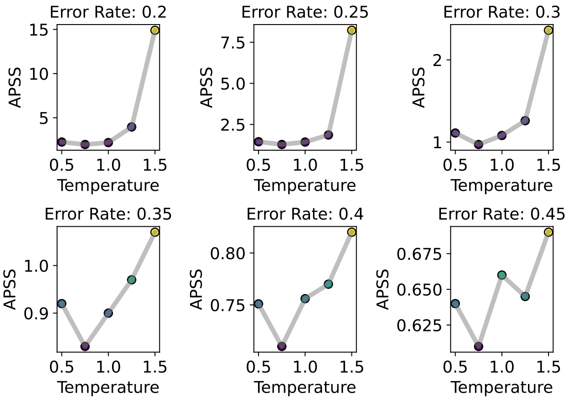


Figure 9: All results of the sensitivity analysis to variations in temperature.

Table 5: Results for TriviaQA using WizardLM-v1.2.

| | | Error Rate | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Methods | Logit-Access | 0.25 | | | 0.3 | | | 0.35 | | |
| | | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ |
| First-K$_{\text{white}}$ | ✓ | 75.1 | 68.7 | 3.19 | 71.0 | 65.8 | 2.56 | 66.4 | 63.3 | 1.84 |
| CLM | ✓ | 75.1 | 63.3 | **3.01** | 70.1 | 64.9 | 2.20 | 65.0 | 63.3 | 1.43 |
| SCP | ✓ | 75.4 | 57.9 | 3.29 | 70.1 | 62.2 | 2.15 | 65.2 | 56.4 | 1.68 |
| SAPS | ✓ | 75.1 | **70.6** | 3.83 | 70.1 | 53.2 | 2.30 | 65.1 | 54.9 | 1.37 |
| First-K$_{\text{black}}$ | ✗ | 75.7 | 58.0 | 4.94 | 71.5 | 66.6 | 2.59 | 68.4 | 65.6 | 1.84 |
| **LofreeCP (Ours)** | ✗ | 75.1 | 68.0 | 4.07 | 70.0 | **67.7** | **1.92** | 65.1 | **70.1** | **1.27** |
| | | Error Rate | | | | | | | | |
| Methods | Logit-Access | 0.4 | | | 0.45 | | | 0.5 | | |
| | | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ |
| First-K$_{\text{white}}$ | ✓ | ✗ | ✗ | ✗ | 55.2 | 56.0 | 0.99 | ✗ | ✗ | ✗ |
| CLM | ✓ | 60.1 | 65.3 | 1.25 | 55.1 | 69.1 | 0.92 | 50.1 | 71.3 | 0.81 |
| SCP | ✓ | 60.0 | 65.9 | 1.30 | 55.1 | 67.8 | 1.01 | 50.1 | 70.1 | 0.82 |
| SAPS | ✓ | 60.0 | 47.3 | 1.37 | 55.2 | 53.7 | 1.05 | 50.1 | 60.6 | 0.83 |
| First-K$_{\text{black}}$ | ✗ | ✗ | ✗ | ✗ | 56.9 | 57.4 | 0.99 | ✗ | ✗ | ✗ |
| **LofreeCP (Ours)** | ✗ | 60.2 | **69.8** | **0.98** | 55.3 | **70.4** | **0.81** | 50.2 | **72.5** | **0.69** |

Table 6: Results for TriviaQA using Vicuna-v1.5.

| | | Error Rate | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Methods | Logit-Access | 0.475 | | | 0.5 | | | 0.525 | | |
| | | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ |
| First-K$_{\text{white}}$ | ✓ | 53.0 | 42.1 | **2.23** | 50.4 | 42.4 | 1.63 | ✗ | ✗ | ✗ |
| CLM | ✓ | 52.5 | **45.1** | 2.60 | 50.1 | 45.5 | 1.39 | 47.5 | 47.7 | 1.21 |
| SCP | ✓ | 52.6 | 39.0 | 2.66 | 50.0 | 40.5 | 1.43 | 47.9 | 49.3 | 1.14 |
| SAPS | ✓ | 52.7 | 40.1 | 2.30 | 50.3 | **48.8** | 1.59 | 47.5 | 45.6 | 1.24 |
| First-K$_{\text{black}}$ | ✗ | 53.4 | 44.1 | 2.75 | 50.9 | 42.3 | 1.62 | ✗ | ✗ | ✗ |
| **LofreeCP (Ours)** | ✗ | 52.5 | 39.3 | 2.27 | 50.0 | 39.1 | **1.33** | 47.6 | **50.1** | **1.12** |
| | | Error Rate | | | | | | | | |
| Methods | Logit-Access | 0.4 | | | 0.45 | | | 0.5 | | |
| | | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ |
| First-K$_{\text{white}}$ | ✓ | 45.0 | 46.7 | 0.99 | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| CLM | ✓ | 45.2 | 50.7 | 1.01 | 42.5 | 50.6 | 0.85 | 40.1 | 56.2 | 0.83 |
| SCP | ✓ | 45.4 | 52.4 | **0.96** | 42.6 | 48.6 | 0.85 | 40.5 | 52.0 | 0.76 |
| SAPS | ✓ | 45.0 | 46.2 | 1.04 | 42.6 | 50.8 | 0.84 | 40.1 | 57.9 | 0.75 |
| First-K$_{\text{black}}$ | ✗ | ✗ | ✗ | ✗ | 44.6 | 46.2 | 0.97 | ✗ | ✗ | ✗ |
| **LofreeCP (Ours)** | ✗ | 45.1 | **55.3** | **0.96** | 42.7 | **58.0** | **0.82** | 40.2 | **58.5** | **0.73** |

Table 7: Results for WebQuestions using WizardLM-v1.2.

| | | Error rate | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | Logit-Access | 0.45 | | | 0.5 | | | 0.55 | | | 0.6 | | |
| | | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ |
| First-K$_{\text{white}}$ | ✓ | 55.5 | 42.5 | 3.40 | 53.0 | 40.6 | 2.70 | 49.1 | 39.0 | 1.91 | ✗ | ✗ | ✗ |
| CLM | ✓ | 55.1 | **52.3** | 3.02 | 50.2 | 40.1 | 2.01 | 45.2 | 28.6 | 1.58 | 40.4 | 31.2 | 1.19 |
| SCP | ✓ | 55.2 | 45.9 | 3.63 | 50.1 | 40.8 | 2.04 | 45.0 | 37.1 | 1.55 | 40.2 | 47.8 | 1.04 |
| SAPS | ✓ | 55.0 | 45.7 | 3.38 | 50.1 | 41.1 | 2.15 | 45.2 | 28.6 | 1.58 | 40.4 | 31.2 | 1.19 |
| First-K$_{\text{black}}$ | ✗ | 56.7 | 43.6 | 3.40 | 50.9 | 45.0 | 1.91 | ✗ | ✗ | ✗ | 41.4 | 41.1 | 1.00 |
| **LofreeCP (Ours)** | ✗ | 55.0 | 45.3 | **2.87** | 50.0 | **46.5** | **1.88** | 45.1 | **49.9** | **1.18** | 40.1 | **51.7** | **0.82** |

Table 8: Results for WebQuestions using Vicuna-v1.5.

| | | Error rate | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | Logit-Access | 0.575 | | | 0.6 | | | 0.625 | | | 0.65 | | |
| | | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ | ECR | SSC↑ | APSS↓ |
| First-K$_{\text{white}}$ | ✓ | 43.2 | 23.8 | 1.99 | 41.7 | 26.9 | 1.57 | ✗ | ✗ | ✗ | 36.6 | 36.6 | 1.00 |
| CLM | ✓ | 42.5 | 32.3 | 1.88 | 40.1 | 36.2 | 1.32 | 37.6 | 38.2 | 1.08 | 35.0 | 41.8 | 0.83 |
| SCP | ✓ | 42.6 | 31.1 | 1.91 | 40.1 | 34.4 | 1.28 | 38.2 | 37.3 | 1.06 | 35.2 | **43.7** | 0.87 |
| SAPS | ✓ | 42.5 | 32.3 | 1.88 | 40.1 | 36.2 | 1.32 | 37.6 | 38.2 | 1.08 | 35.0 | 41.8 | 0.83 |
| First-K$_{\text{black}}$ | ✗ | 43.7 | 25.9 | 2.01 | 40.9 | 25.5 | 1.57 | ✗ | ✗ | ✗ | 36.8 | 36.8 | 1.00 |
| **LofreeCP (Ours)** | ✗ | 42.5 | **32.4** | **1.73** | 40.1 | **36.7** | **1.22** | 37.5 | **39.6** | **0.97** | 35.0 | 39.3 | **0.81** |