Improving Decision-Making in Open-World Agents with Conformal Prediction and Monty Hall

Hart Vishwakarma *
Department of Computer Science
University of Wisconsin
Madison, WI 53706, USA
hvishwakarma@cs.wisc.edu

Alan Mishler, Thomas Cook, Niccolò Dalmasso, Natraj Raman, Sumitra Ganesh
JPMorganChase AI Research
New York, NY 10017, USA
{alan.mishler,thomas.cook,niccolo.dalmasso,natraj.raman,sumitra.ganesh}
@jpmchase.com

Abstract

Large language models (LLMs) are empowering decision-making in open-world agents in several applications, including tool or API usage and answering multiple choice questions (MCQs). However, they often make overconfident, incorrect predictions or "hallucinations", which can be risky in high-stakes settings like healthcare, and finance. To improve safety, we leverage conformal prediction (CP), a model-agnostic framework that provides distribution-free uncertainty quantification. CP transforms a score function, which measures how well an output "conforms" to a given input, into prediction sets that contain the true answer with high probability. While CP ensures this coverage guarantee for arbitrary scores, the quality of the scores significantly impacts the size of prediction sets. Prior works have relied on LLM logits or other heuristic scores, lacking guarantees on their quality. To address this issue, we propose an optimization framework (CP-OPT) to learn scores that minimize set sizes while maintaining coverage guarantees. Furthermore, leveraging the coverage guarantees of CP, we propose a conformal revision of questions (CROQs) to revise the problem by narrowing down the available choices to those in the prediction set. Our results on MMLU and ToolAlpaca datasets with Llama3 and Phi-3 models show that optimized CP scores reduce the set sizes by up to 13% and CROQs improves accuracy relatively by up to 4.6% overall and up to 15% in non-trivial parts of the input space.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in various natural language tasks, including tool usage and multi-choice question answering (MCQ) tasks [14, 20, 8]. However, despite their success, LLMs often operate as black boxes, offering limited insight into the uncertainty of their predictions. This lack of transparency becomes especially concerning in tool usage, where the model must select the correct tool or API to complete a task. Missteps in such scenarios, especially when made with high confidence, can lead to significant errors. To address these challenges, we focus on improving uncertainty quantification in MCQ and tool usage tasks.

*Some of this work was performed while at JPMorganChase.

Workshop on Open-World Agents (OWA) at NeurIPS, 2024.

There is an emerging line of work proposing to use conformal prediction for UQ in LLMs [10, 19, 12, 15]. Conformal prediction (CP) [23, 3] is a promising framework for uncertainty quantification in machine learning models, offering *distribution-free* and *model-agnostic* guarantees on the reliability of predictions. By producing prediction sets that contain the true outcome with a user-specified confidence level (e.g. 95%), CP provides a *rigorous* and *interpretable* measure of uncertainty.

A critical component in CP is the *score function*, which measures how well a candidate output "conforms" to a given input. For example, in classification settings, it is common to use the classifier logits corresponding to each class for a given input as the score function [2]. While conformal prediction gives coverage guarantee for *any* scoring function, the size of the output set depends on the *quality* of scores – bad scores will lead to large sets. Previous works that apply CP for MCQs have used readily available scores such as the logits (softmax) output from the LLM [10] or have designed some heuristic scores based for example on repeated querying of the LLM [19]. Logits can be overconfident and may have biases for some options [27] and heuristic scores lack performance guarantees. Thus we need principled solutions to obtain scores that are guaranteed to minimize set sizes while maintaining the coverage guarantee.

The interpretability of conformal prediction (CP) sets offers potential benefits beyond uncertainty quantification in LLMs. While prior work has primarily focused on integrating CP with LLMs to produce prediction sets, we believe these sets can be leveraged in more versatile ways. Specifically, inspired by the Monty Hall problem [17], we hypothesize that by revising a multiple-choice question to include only the options within the CP prediction set, the LLM is more likely to provide the correct answer due to the reduced number of choices, leading to improved accuracy.

To summarize, we focus on the following two questions while applying CP in solving MCQs with LLMs. **R1.** *How can we get optimal scores for conformal prediction?* **R2.** *Can the prediction sets from CP be utilized to improve accuracy?*

To address R1, we design a post-hoc procedure to learn the optimal scores for conformal prediction. To address R2, we conduct experiments in which we re-prompt the LLM with multi-choice questions using the answer sets generated by conformal prediction. In all cases, we assume white-box access to the LLM such that we are able to access the model logits and internal weights. We summarize our main contributions as follows,

1. We design a post-hoc score function optimization framework (CP-OPT) that can be applied to any pre-trained LLM. Moving away from the unreliable LLM logits and heuristic scores, it provides a principled way to learn the scores for conformal prediction. Empirically, we show that our procedure leads to a relative reduction in average set sizes by up to 13%, in contrast to the baselines using LLM logits as scores, at the same level (95%) of marginal coverage.
2. Extending the utility of CP beyond uncertainty quantification, we propose the conformal revision of questions (CROQs), in which we revise the question by narrowing down the choices to those in the prediction sets output by CP. Then the LLM is re-prompted with the revised question. We show that this procedure can lead to relative accuracy improvements up to 4.6% overall and up to 15% in non-trivial parts of the input space.

2 Preliminaries

We provide background on solving MCQ tasks with LLMs and conformal prediction.

2.1 Solving Multiple Choice Questions (MCQs) Using LLMs

MCQ Setup. MCQs are a general abstraction for expressing problems where we need to select the correct choice(s) from a given set of choices for each problem. These are common in conventional question-answering tasks such as MMLU [8] and recently with the versatility of modern language models several tasks such as tool learning can be expressed as MCQs [20, 14]. An MCQ consists of the question text Q_i, i.e. a sequence of tokens, and a set of choices $O_i = \{(Y_1, V_1^{(i)}), (Y_2, V_2^{(i)}), \ldots, (Y_{m_i}, V_{m_i}^{(i)})\}$. Here each Y_j is a unique character from the English alphabet and we assume the number of choices $m \leq$ size of the alphabet and $V_j^{(i)}$ is the option text for jth option. Denote the whole MCQ instance as $x_i = (Q_i, O_i)$. Let $\mathcal{X}_m$ denote the space of

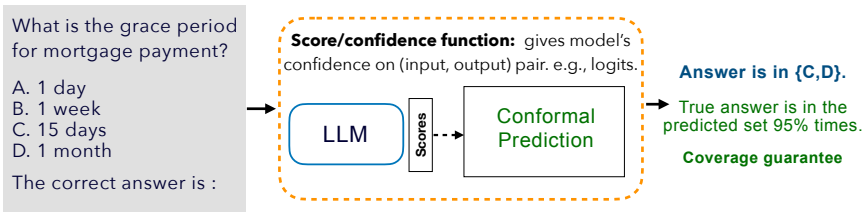

Figure 1: In the high-stake applications using conformal prediction allows the answer of the LLM to both include multiple answers as well as enjoy a coverage guarantee, i.e., a probabilistic statement about the predicted set containing the true answer with high probability.

MCQs with m choices and $\mathbb{P}_{\mathcal{X}_m}$ is a distribution over $\mathcal{X}_m$, from which samples for training, validation, and testing are drawn independently. Here we assume for each question Q_i there is only one true answer key $y_i^\star \in \{Y_1, Y_2, \dots Y_{m_i}\} = \mathcal{Y}_{m_i}$.

Prompt for MCQ. We concatenate the question text Q_i and the choices O_i, all separated by a new line character, and in the end append text "`The correct answer is: `". The expectation is that on this input prompt the next token predicted by the LLM will be one of the options keys. See Appendix B for a prompt example. We consider zero-shot prompts and do not include example questions and answers in the prompt. We also add the prefix and suffix tokens to the prompt as recommended by the language model providers. Since these are fixed modifications to x_i, we will use x_i to denote the final prompt and the MCQ instance analogously.

LLM Inference. We run the forward pass of the auto-regressive LLM [21, 7, 1] on the input prompt x_i. The whole prompt x_i is a sequence of tokens $t_1^{(i)}, t_2^{(i)}, \dots t_{n_i}^{(i)}$.

$$\boldsymbol{l}_1^{(i)}, \boldsymbol{l}_2^{(i)}, \dots, \boldsymbol{l}_{n_i}^{(i)} \leftarrow \texttt{LLM}\big(t_1^{(i)}, t_2^{(i)}, \dots t_{n_i}^{(i)}\big) \tag{1}$$

Here $\boldsymbol{l}_j^{(i)} \in \mathbb{R}^{|V|}$ and V is the universal set of tokens for the given LLM and $|V|$ is its size. The logits $\boldsymbol{l}_j^{(i)}$ express the likelihood of the next token after $t_1^{(i)}, \dots, t_j^{(i)}$. Thus the last token's logits $\boldsymbol{l}_{n_i}^{(i)}$ are expected to have a higher value for the correct answer key. We extract the logit vector $\bar{\boldsymbol{l}}_i \in \mathbb{R}^{m_i}$ corresponding to the option keys as follows,

$$\bar{\boldsymbol{l}}_i := \big[\, \boldsymbol{l}_{n_i}^{(i)}[Y_1], \boldsymbol{l}_{n_i}^{(i)}[Y_2], \dots, \boldsymbol{l}_{n_i}^{(i)}[Y_{m_i}] \,\big]. \tag{2}$$

Here $\boldsymbol{l}_{n_i}^{(i)}[Y_j]$ means the logit value corresponding to the token Y_j in the last token's logits $\boldsymbol{l}_{n_i}^{(i)}$. The logits $\bar{\boldsymbol{l}}_i$ are converted to softmax scores $s(x_i)$. The softmax score of point x_i and option key y is given by $s(x_i, y)$ and the predicted answer key $\hat{y}_i$ corresponds to the maximum softmax value.

$$s(x_i) := \texttt{softmax}(\bar{\boldsymbol{l}}_i), \qquad s(x_i, y) := s(x_i)[y], \qquad \hat{y}_i := \underset{y \in \{Y_1, \dots Y_{m_i}\}}{\arg\max}\; s(x_i, y) \tag{3}$$

Note that prior works on solving MCQs using LLMs [10, 19] use similar steps but lack precise details, we outlined these here for clarity and to make the paper self-contained.

2.2 Conformal Prediction

Conformal prediction (CP) [23, 3] is a framework for quantifying uncertainty in machine learning models. It provides a flexible and user-friendly approach to output statistically valid *prediction sets* (or intervals) on model predictions. The key strength of conformal prediction lies in its *distribution-free* guarantees – it ensures that the constructed prediction sets are valid regardless of the underlying data distribution and model. This property is particularly desirable in the context of language models, as it is hard to characterize the language data distributions or put specific assumptions/restrictions on the LLMs.

Score Function. Let $g : \mathcal{X}_m \times \mathcal{Y}_m \mapsto \mathbb{R}$ be a *conformity* score function – larger scores indicate better agreement between x and y. A common choice of scoring function is the softmax scores from the given model as in equation (3). There can be several other heuristic choices for score function e.g. self-consistency based scores [19]. As such, CP can work with *any* scoring function but the quality of the scores reflects in the size of the prediction sets.

Prediction Sets. Given a threshold τ on the scores, the prediction set for point any $x \in \mathcal{X}_m$ is defined as follows,

$$C(x \mid g, \tau) := \{y \in \mathcal{Y}_m : g(x, y) \geq \tau\}. \tag{4}$$

The size of the prediction set quantifies the uncertainty, i.e., the smaller the prediction set the smaller the prediction uncertainty and vice versa.

Split Conformal Prediction. Similar to prior works [10, 19], we use *Split Conformal Prediction* [13, 11] due to its popularity, ease of use, and computational efficiency. For a given scoring function $g : \mathcal{X}_m \times \mathcal{Y}_m \mapsto \mathbb{R}$, Split Conformal Prediction first fits an arbitrary model on train dataset, $D_{\text{train}} = \{x_i, y_i^\star\}_{i=1}^{n_{\text{train}}}$. Then a separate calibration dataset $D_{\text{cal}} = \{x_i, y_i^\star\}_{i=1}^{n_{\text{cal}}}$ is used to compute a threshold $\hat{\tau}$ such that

$$\hat{\tau} = \min \left\{ q : \frac{1}{n_{\text{cal}}} \sum_{i=1}^{n_{\text{cal}}} \mathbb{1}\left(g(x_i, y_i^\star) \leq q\right) \geq \alpha \right\}. \tag{5}$$

When $\hat{\tau}$ is used in place of the unknown parameter, τ, in Equation 4, we enjoy a marginal coverage guarantee for prediction sets constructed on unseen test data points, which we formalize below in Proposition 2.1.

Proposition 2.1. *(Marginal Coverage Guarantee) Let g being a (given) conformity score function and $\hat{\tau}$ be an α threshold computed via Split Conformal Prediction on $D_{\text{cal}} = \{x_i, y_i^\star\}_{i=1}^{n_{\text{cal}}} \sim \mathbb{P}_{\mathcal{X}_m \times \mathcal{Y}_m}$. Then, for a new sample $(\tilde{x}, \tilde{y}^\star) \sim \mathbb{P}_{\mathcal{X}_m \times \mathcal{Y}_m}$, we have that*

$$\mathbb{P}_x(\tilde{y}^\star \in C(\tilde{x} \mid g, \hat{\tau})) \geq 1 - \alpha. \tag{6}$$

The above coverage guarantee makes CP an attractive tool for safely deploying LLMs. Figure 1 illustrates the usage of conformal prediction by the prior works in answering MCQs with LLMs. While using CP is promising, the issue of score quality (critical to CP) can stymie its effectiveness. Next, we discuss our solutions to improve CP and its utility in solving MCQs with LLMs.

3 Methodology

In this section, we discuss details of our pipeline for question revision using conformal prediction and our method for learning optimal scores for conformal prediction.

3.1 Conformal Revision of Questions (CROQs)

Conformal prediction in language models can go beyond quantifying uncertainty in the output. While the sizes of prediction sets quantify the uncertainty, the sets are *interpretable* and are backed with the *coverage* guarantee. Inspired by the Monty Hall problem [17], these properties, in the context of LLMs, open up the possibility to revise the question after first round of CP and give the revised question back to the LLM. It is expected that with reduced choices in the question, the chances of the LLM arriving at the right answer will increase. We describe this two step procedure below and it is illustrated in Figure 2,

Step 1: Usual Conformal Procedure. In the first step we run the aforementioned split conformal procedure with coverage level $1-\alpha$, to estimate the threshold $\hat{\tau}$ using the questions in the calibration data. Then we pass each test instance x_i through the LLM and CP procedure to obtain a first stage prediction set, $C(x_i|g, \hat{\tau})$.

Step 2: Revise and Ask LLM Again. After getting the first stage prediction set, $C(x_i|g, \hat{\tau})$, if the set size is greater than 1 and less than m_i, it modifies x_i to $x_i' = (Q_i, O_i')$, where $O_i' = \{(K_j, V_j^{(i)}) : K_j \in C(x_i|g, \hat{\tau})\}$. The keys in O_i' are changed, they start with the first alphabet and go to the alphabet corresponding to the number of choices available. Next, we transform x_i' into a prompt format and input to the LLM. We run the same inference procedure and extract the predicted answer key $\hat{y}_i'$ as in equation (3).

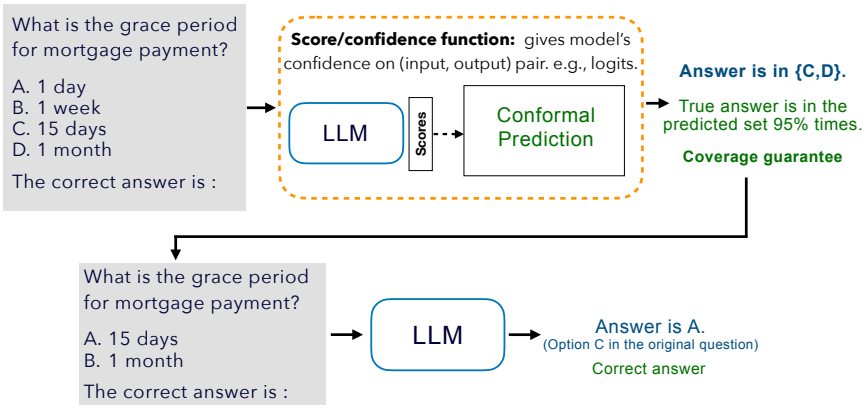

Figure 2: (CROQs) Illustration of conformal revision of questions and prompting the LLM with the revised question. In this example, the initial predicted set by LLM + conformal prediction (CP) is $\{\text{C}, \text{D}\}$. The question is revised to contain only the answer choices in the prediction set. The labels (keys) are revised as well. Then the LLM is prompted with the revised question. Since CP provides rigorous coverage guarantees, we expect that this way of re-prompting LLM with reduced answer choices will improve the chances of obtaining the correct answer. See Section 3.1 for more details.

3.2 Scores Optimization for Conformal Prediction (CP-OPT)

We describe our method for post-hoc learning of the optimal scores for conformal prediction for solving MCQ with LLMs. Similar ideas have been incorporated in the training objective of classifiers [18] so that the classifiers' softmax output is better suited for CP. However, the LLMs are not trained with this objective and we want to apply CP to any given LLM, therefore, we design a post-hoc method to optimize the scores. We first describe the theoretical version to characterize the optimal scores and then instantiate its practical version.

3.2.1 Theoretical Characterization of the optimal scores

For any scoring function $g : \mathcal{X}_m \times \mathcal{Y}_m \mapsto \mathbb{R}$ and threshold τ, the membership of any y in the prediction set $C(x \mid g, \tau)$ is given by $\mathbb{1}(y \in C(x \mid g, \tau)) \iff \mathbb{1}\{g(x, y) \geq \tau\}$. Define the expected set size $S(g, \tau)$ and the marginal coverage $\mathcal{P}(g, \tau)$ as follows,

$$S(g, \tau) := \mathbb{E}_{x \sim \mathcal{X}_m}\Big[\sum_{y \in \mathcal{Y}_m} \mathbb{1}\{g(x, y) \geq \tau\}\Big] = \sum_{y \in \mathcal{Y}_m} \mathbb{E}_{x \sim \mathcal{X}_m}\left[\mathbb{1}\{g(x, y) \geq \tau\}\right], \tag{7}$$

$$\mathcal{P}(g, \tau) := \mathbb{E}_{x \sim \mathcal{X}_m}\left[\mathbb{1}\{g(x, y^\star) \geq \tau\}\right]. \tag{8}$$

Let $\mathcal{G}$ be a flexible space of scoring functions. If we have an oracle that can give these expected set sizes and coverage for any $g \in \mathcal{G}$ and $\tau \in \mathbb{R}$ we can hope to find an optimal scoring function $g^\star$ and threshold $\tau^\star$ by minimizing the expected set size $S(g, \tau)$ over a flexible space of g and τ, subject to the marginal coverage $\mathcal{P}(g, \tau)$ being at least $1 - \alpha$.

$$g^\star, \tau^\star := \underset{g : \mathcal{X}_m \times \mathcal{Y}_m \mapsto \mathbb{R}, \tau \in \mathbb{R}}{\arg\min} \; S(g, \tau) \quad \text{s.t.} \quad \mathcal{P}(g, \tau) \geq 1 - \alpha. \tag{P1}$$

3.2.2 Practical Version: Differentiable Surrogates and Empirical Estimates

The above problem (P1) gives us a way to characterize optimal scoring functions and thresholds. However, in practice, we do not know the distribution and thus do not have access to the quantities in (7) and (8). Instead we get their estimates using a finite training sample $D_{\text{train}} = \{(x_i, y_i^\star)\}_{i=1}^{n_t}$ drawn independently from the same distribution.

$$\widehat{S}(g, \tau) := \frac{1}{n} \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}_m} \mathbb{1}\{g(x_i, y) \geq \tau\}, \quad \widehat{\mathcal{P}}(g, \tau) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{g(x_i, y_i^\star) \geq \tau\}. \tag{9}$$

Using these plug-in estimators in problem (P1) we get a revised optimization problem. However, it is difficult to solve this problem as the objective and constraints are not differentiable. To make them differentiable, we introduce the following surrogates. Given $g(x, y)$ and τ, define the following sigmoid function with $\beta > 0$, $\sigma(x, y, g, \tau, \beta) := 1/\big(1 + \exp(-\beta\,(g(x, y) - \tau))\big)$. The sigmoid function provides a nice differentiable approximation to the indicator variable for $g(x, y) \geq \tau$. The approximation is tighter with higher β i.e., $\sigma(x, y, g, \tau, \beta) \to \mathbb{1}\{g(x, y) \geq \tau\}$ as $\beta \to \infty$, and $g(x, y) \geq \tau \iff \sigma(x, y, g, \tau) \geq 1/2$. By using these sigmoid surrogates in equation (9) we obtain the following smooth plugin estimates,

$$\widetilde{S}(g, \tau) := \frac{1}{n} \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}_m} \sigma(x_i, y, g, \tau, \beta), \quad \widetilde{\mathcal{P}}(g, \tau) := \frac{1}{n} \sum_{i=1}^{n} \sigma(x_i, y_i^\star, g, \tau, \beta). \tag{10}$$

It is easy to see that as $n, \beta \to \infty$, the surrogate average set size and coverage will converge to their population versions, i.e. $\widetilde{S}(g, \tau) \to S(g, \tau)$ and $\widetilde{\mathcal{P}}(g, \tau) \to \mathcal{P}(g, \tau)$. We replace the expected set size and marginal coverage by these smooth surrogates in (P1) and transform it into an unconstrained problem with the penalty term $\lambda > 0$ and introduce ℓ_2 regularization to encourage low norm solutions. The resulting problem (P2) is differentiable and we solve it using stochastic gradient descent. More details on training are deferred to later sections.

$$\tilde{g}, \tilde{\tau} := \underset{g: \mathcal{X}_m \times \mathcal{Y}_m \mapsto \mathbb{R}, \tau \in \mathbb{R}}{\arg\min} \widetilde{S}(g, \tau) - \lambda \widetilde{\mathcal{P}}(g, \tau) + \lambda_2 ||g||_2. \tag{P2}$$

By solving the above optimization problem we obtain a score function $\tilde{g}$ and use it to compute scores for conformal prediction. While we get the threshold $\tilde{\tau}$ as well here, but since its estimated along with $\tilde{g}$ on the same data so it could be biased and using it on test data may violate the coverage guarantee. We re-estimate it using the split conformal procedure to ensure the coverage guarantee is strictly followed. To train g we use the logits and penultimate layer's representations from the LLM corresponding to the last token as features. We learn g from a function class of $3-$layer neural networks with `tanh` activation. Note that while we made these specific choices here, our framework is flexible to work with any choice of features and function class.

4 Experiments

We conduct experiments on benchmark MCQ and tool usage tasks with open-weight instruction-tuned models to verify the following claims,

C1. Using our CP-OPT scores in conformal prediction on MCQ tasks with LLMs yields a smaller average set size at the same level of coverage in comparison to LLM logits.

C2. Conformal revision of questions and asking LLM again (CROQs) improves the accuracy over the standard inference.

C3. The accuracy improvement with CROQs happens with LLM logits and our CP-OPT scores both. Moreover, the improvement is higher with our scores.

4.1 Experimental Setup

We first describe the setup for the experiments and then discuss the results for the above claims.

4.1.1 Datasets

MMLU [8] is a popular benchmark dataset for multiple choice questions (MCQs) from 57 domains including humanities, math, medicine, etc. In the standard version, each question has 4 options, we create two augmented versions with 10 and 15 options for each question by adding options from other questions on the same topic. We ensure there is no duplication in options. The standard dataset has few training points, we randomly draw 30%, and 10% of the points from the test split and include them in the training and validation sets respectively. Note, that we remove these points from the test set. The resulting splits have 4.5k, 2.9k, and 8.4k points in the train, validation, and test splits.

Dataset	Model	Number of Options	Scores	Avg. Set Size	Coverage
MMLU	Llama 3	4	Logits	2.56	95.8
			CP-OPT	**2.51**	95.4
		10	Logits	5.19	95.57
			CP-OPT	**4.74**	95.02
		15	Logits	7.66	95.3
			CP-OPT	**6.61**	94.6
	Phi 3	4	Logits	2.21	94.6
			CP-OPT	**2.15**	94.6
		10	Logits	4.61	94.7
			CP-OPT	**4.50**	94.5
		15	Logits	**6.46**	93.9
			CP-OPT	6.66	94.0
ToolAlpaca	Llama 3	4	Logits	**1.587**	99.1
			CP-OPT	1.642	98.2
		10	Logits	1.853	94.5
			CP-OPT	**1.550**	94.5
		15	Logits	1.697	92.7
			CP-OPT	**1.413**	89.9
	Phi 3	4	Logits	1.257	99.1
			CP-OPT	**1.138**	95.4
		10	Logits	1.303	94.5
			CP-OPT	**1.266**	94.5
		15	Logits	**1.642**	95.4
			CP-OPT	1.706	94.5

Table 1: Average set sizes and coverage rates for conformal prediction sets on the MMLU and ToolAlpaca datasets using `Llama-3-8B-Instruct` (Llama 3) and `Phi-3-4k-mini-Instruct` (Phi 3). For each dataset, we vary the number of responses. Using CP-OPT for a score function produces smaller average set sizes more frequently compared to using logits.

Model	Score	Set Size	1	2	3	4	Overall
Llama 3	Logits	**Coverage**	98.61	100.00	100.00	100.00	99.08
		Fraction	66.06	17.43	8.26	8.26	100.00
		Acc. Before	98.61	68.42	77.78	55.56	88.07
		Acc. After	**98.61**	**78.95**	**77.78**	55.56	89.91
	Ours	**Coverage**	96.88	100.00	100.00	100.00	98.17
		Fraction	58.72	23.85	11.93	5.50	100.00
		Acc. Before	96.88	80.77	61.54	83.33	88.07
		Acc. After	96.88	**88.46**	**69.23**	**83.33**	**90.83**
Phi 3	Logits	**Coverage**	98.81	100.00	100.00	0.00	99.08
		Fraction	77.06	20.18	2.75	0.00	100.00
		Acc. Before	98.81	**72.73**	0.00	0.00	90.83
		Acc. After	**98.81**	59.09	0.00	0.00	88.07
	Ours	**Coverage**	94.68	100.00	0.00	0.00	95.41
		Fraction	86.24	13.76	0.00	0.00	100.00
		Acc. Before	94.68	66.67	0.00	0.00	90.83
		Acc. After	94.68	**66.67**	0.00	0.00	**90.83**

Table 2: Results for CROQ experiment on ToolAlpaca dataset with 4 response options. Using the CP-OPT (Ours) score function results in a greater accuracy for LLama 3, while accuracy is maintained when using Phi 3. For brevity, we have omitted similar tables for when the number of response options is increased to 10 and 15 since very few prediction sets have size greater than 4.

ToolAlpaca [20] contains 3.9k tool-use instances from a multi-agent simulation environment. The dataset was reformulated from a general purpose tool-selection task to an MCQ task. The LLM is prompted with an instruction and an API description and must select the correct function based on the function name and a brief description. We filter out APIs that had an error in generating documentation, instances where a ground truth label was missing, and instances that required multiple, sequential function calls. After filtering, 2703 MCQ examples remained. The "train" split contains 2503 synthetic examples, "validation" contains 108 synthetic validation examples, and "test" contains 92 real API examples. This split is consistent with the original dataset. The number of functions for each class varies. Since we fixed the number of responses, answers were either downselected, or additional responses from different, random questions were inserted. Arguments were stripped from the function call so that the MCQ task was focused on tool selection, a critical task in tool usage. Example questions and responses for MMLU and ToolAlpaca are provided in Appendix B.

4.1.2 Models and Scores

We use auto-regressive language models based on the transformer architecture. We choose instruction-tuned, open-weight, and small to medium sized models, for reproducibility and reduced computational cost. Specifically, we use `Llama-3-8B-Instruct` [7] by Meta and `Phi-3-4k-mini-Instruct` [1] model by Microsoft.

We use the following scores for conformal prediction, **(i) LLM Logits (Softmax)** are extracted from the LLM as discussed in section 2.1. They were used in prior works [10, 19]. **(ii) CP-OPT (Ours)** are the scores learned by using the score optimization procedure discussed in section 3.2. We use the data from the train split to learn these scores. We omit the self-consistency based heuristic scores proposed by Su et al. [19], as it requires repeated inferences to get good estimates of the scores, and hence has a high computational cost. We use the validation splits of the dataset as D_{cal} for the conformal procedure and calibrate it for the coverage guarantee of 95% i.e. use $\alpha = 0.05$.

4.2 Discussion

C1. Improvement in conformal set sizes with our CP-OPT scores. We run the conformal prediction procedure on logits and CP-OPT scores and obtain conformal sets for points in the test set. We observe the average set size and coverage for each dataset, model, and score combination. The results are in table 1. As expected, we see a drop in the set sizes at a similar coverage level with our scores in most cases. The improvement gets more pronounced with a higher number of options. There are some cases, where we do not see a drop in set sizes or a significant drop in coverage. We notice that this occurs when `Llama-3-8B-Instruct` is presented with fewer options and when `Phi-3-4k-mini-Instruct` is presented with a large number of options.

C2. Test accuracy improvement with conformal revision of questions (CROQs). To verify this, we run the conformal procedure as above with logits and CP-OPT scores to get the conformal sets for the questions in the test data. Then we run the CROQs, as discussed in section 3.1, to get the final predicted labels from the LLM following question revision. We report overall accuracy before and after CROQs in the table 2 for Toolalpaca, and tables 3,4 for MMLU with 4 and 10 options respectively and defer the results for MMLU with 15 options to the Appendix. Across all settings except for MMLU with 4 options, we see improvements in accuracy with CROQs. Specifically, we see relative improvements by up to 4.6% in overall accuracy and up to 15% in non-trivial parts of the input space. The drop in accuracy in MMLU with 4 options is not consistent with the expectations, we defer the diagnosis to future works.

Note that the conformal sets have a coverage guarantee of around 95%, and, for revision on test samples, we do not know if our revised answer set contains the correct response. Thus, these improvements are even more significant given that up to 5% of the revised questions did not have the true choice among the revised answer set. These findings suggest a multi-round CROQs procedure with higher-level coverage guarantees in each round could be more effective. However, the inference cost increases with each round and we may need fresh samples for conformal calibration to avoid biases. We defer this multi-round CROQs exploration to future work.

C3. CROQs gives better final test accuracy when used with our CP-OPT scores, in contrast to logits. This finding follows a similar pattern to *C1*. Since the use of CP-OPT will lead to a smaller set size on average, the number of responses in the revised question is smaller than when using logits. We

posit that this reduction in responses reduces the risk of the LLM losing context due to the presence of additional, unrelated, or distracting answers.

5 Related Works

Conformal Prediction for Uncertainty Quantification with LLMs Recently there has been growing interest in using conformal prediction to quantify and control uncertainty in LLM-related tasks. In the context of multi-choice question answering (MCQ), previous works have investigated a variety of conformal score functions, including (the softmax of) the LLM logits corresponding to the response options [10, 16] or functions thereof [26], confidence scores generated by the LLM itself, or "self-consistency" scores derived by repeated querying of the LLM [19]. We build on this work by aiming to learn a conformal score function that yields small conformal sets, rather than taking the score function as given.

In addition to the MCQ setting, there has been recent work utilizing conformal prediction in the context of open-ended response generation [15, 12, 5]. This setting differs in that there is not necessarily a unique correct response, so the notion of coverage must be redefined around *acceptability* or *factuality* rather than correctness. When factuality is the target, the goal is to calibrate a pruning procedure that removes a minimal number of claims from an LLM-generated open response, such that the remaining claims are all factual with high probability; that is, the goal is to retain as large a set as possible, rather than to generate a set with the smallest number of responses possible as in MCQ. Conformal prediction has also been used to capture token-level uncertainty [6, 22].

Optimizing conformal prediction procedures Several recent works have considered how to learn good conformal score functions from data, primarily in the context of supervised learning models [4, 18, 25, 24]. With LLMs, Cherian et al. [5] consider how to learn a good score function to achieve factuality guarantees; their optimization problem differs from ours due to the difference in setting as well as the addition of conditional coverage constraints (ensuring that coverage holds in different parts of the feature space). Kiyani et al. [9] design a framework to minimize the size ("length," in their terminology) of conformal sets, which they apply to MCQ as well as to supervised learning problems. However, their framework is concerned with how to generate sets given a model and a conformity score, rather than how to learn a conformity score.

The works mentioned above all aim to produce (small) conformal sets that satisfy coverage guarantees. Among these, only Ren et al. [16] consider how conformal sets may be used downstream, in their case to improve the efficiency and autonomy of robot behavior. To our knowledge, our work is the first to investigate whether conformal prediction can be used to increase the accuracy of LLMs on MCQ tasks.

6 Conclusion and Future Work

We investigated how conformal prediction (CP) can improve uncertainty quantification in open-world decision-making scenarios, such as multiple-choice question answering (MCQ) and tool usage tasks with large language models (LLMs). We introduced CP-OPT, an optimization framework that minimizes prediction set sizes while maintaining coverage guarantees, and demonstrated its superiority over baseline LLM logits across various models and datasets. Additionally, we proposed CROQs, a method that re-prompts LLMs with the answer options from the conformal set, significantly improving accuracy by narrowing down the available choices. Future work will explore the conditions under which CROQs enhances performance, multi-round CROQs, and the calibration of CP thresholds for tasks with varying numbers of answer options, such as tool usage, where the set of APIs may differ across queries.

Acknowledgments

We are grateful to Sujay Bhatt for pointing out the connection between CROQ and the Monty Hall problem. We thank Alec Koppel and Udari Madhushani Sehwag for fruitful discussions and feedback. We also appreciate the valuable inputs provided by the anonymous reviewers.

Disclaimer

This paper was prepared for informational purposes in part by the Artificial Intelligence Research group of JPMorganChase and its affiliates ("J.P. Morgan") and is not a product of the Research Department of JPMorganChase. J.P. Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

[1] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[2] A. N. Angelopoulos and S. Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, Dec. 2022. URL http://arxiv.org/abs/2107.07511.

[3] A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei. Learn then test: Calibrating predictive algorithms to achieve risk control, 2022.

[4] Y. Bai, S. Mei, H. Wang, Y. Zhou, and C. Xiong. EFFICIENT AND DIFFERENTIABLE CONFORMAL PREDICTION WITH GENERAL FUNCTION CLASSES. In *The Tenth International Conference on Learning Representations*, 2022.

[5] J. J. Cherian, I. Gibbs, and E. J. Candès. Large language model validity via enhanced conformal prediction methods, 2024.

[6] N. Deutschmann, M. Alberts, and M. R. Martínez. Conformal autoregressive generation: Beam search with coverage guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11775–11783, 2024.

[7] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[8] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

[9] S. Kiyani, G. Pappas, and H. Hassani. Length Optimization in Conformal Prediction, June 2024. URL http://arxiv.org/abs/2406.18814.

[10] B. Kumar, C. Lu, G. Gupta, A. Palepu, D. Bellamy, R. Raskar, and A. Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

[11] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[12] C. Mohri and T. Hashimoto. Language models with conformal factuality guarantees. *arXiv preprint arXiv:2402.10978*, 2024.

[13] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pages 345–356. Springer, 2002.

[14] C. Qu, S. Dai, X. Wei, H. Cai, S. Wang, D. Yin, J. Xu, and J.-R. Wen. Tool learning with large language models: A survey. *arXiv preprint arXiv:2405.17935*, 2024.

[15] V. Quach, A. Fisch, T. Schuster, A. Yala, J. H. Sohn, T. S. Jaakkola, and R. Barzilay. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pzUhfQ74c5.

[16] A. Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley, Z. Xu, D. Sadigh, A. Zeng, and A. Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=4ZK8ODNyFXx.

[17] J. S. Rosenthal. Monty hall, monty fall, monty crawl. *Math Horizons*, 16(1):5–7, 2008. ISSN 10724117, 19476213. URL http://www.jstor.org/stable/25678763.

[18] D. Stutz, K. D. Dvijotham, A. T. Cemgil, and A. Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=t8O-4LKFVx.

[19] J. Su, J. Luo, H. Wang, and L. Cheng. Api is enough: Conformal prediction for large language models without logit-access, 2024.

[20] Q. Tang, Z. Deng, H. Lin, X. Han, Q. Liang, B. Cao, and L. Sun. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases. *arXiv preprint arXiv:2306.05301*, 2023.

[21] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[22] D. Ulmer, C. Zerva, and A. Martins. Non-exchangeable conformal language generation with nearest neighbors. In Y. Graham and M. Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1909–1929, St. Julian's, Malta, Mar. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.129.

[23] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[24] R. Xie, R. F. Barber, and E. J. Candès. Boosted Conformal Prediction Intervals, June 2024. URL http://arxiv.org/abs/2406.07449.

[25] Y. Yang and A. K. Kuchibhotla. Selection and Aggregation of Conformal Prediction Sets. *Journal of the American Statistical Association*, pages 1–13, May 2024. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2024.2344700. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2024.2344700.

[26] F. Ye, M. Yang, J. Pang, L. Wang, D. F. Wong, E. Yilmaz, S. Shi, and Z. Tu. Benchmarking llms via uncertainty quantification, 2024.

[27] C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

A Additional Results

This appendix contains results for the CROQ experiment on the MMLU dataset with 15 response options.

Model	Score	Set Size	1	2	3	4	Overall
Llama 3	Logits	**Coverage**	93.58	92.8	93.03	100.00	95.81
		Fraction	33.87	13.93	14.66	37.54	100.00
		Acc. Before	93.58	70.18	**49.39**	40.30	**63.84**
		Acc. After	93.58	**70.44**	48.09	40.30	63.69
	Ours	**Coverage**	93.55	90.71	93.71	100.00	95.37
		Fraction	33.68	15.35	16.81	34.17	100.0
		Acc. Before	93.55	**69.14**	**53.53**	37.27	**63.84**
		Acc. After	93.55	68.14	52.47	37.27	63.52
Phi 3	Logits	**Coverage**	94.75	91.48	93.17	100.0	94.64
		Fraction	37.30	22.86	21.20	18.64	100.0
		Acc. Before	94.75	**70.25**	**52.68**	41.31	**70.27**
		Acc. After	94.75	66.92	50.67	41.31	69.08
	Ours	**Coverage**	93.88	91.55	94.37	100.0	94.64
		Fraction	41.32	21.55	17.73	19.39	100.0
		Acc. Before	93.88	**67.78**	**52.14**	39.29	**70.27**
		Acc. After	93.88	64.81	50.87	39.29	69.41

Table 3: Results for CROQ experiment on MMLU dataset with 4 response options. In this setting, the CROQ procedure leads to a lower accuracy after revision.

Model	Score	Set Size	1	2	3	4	5	6	7	8	9	10	Overall
Llama 3	Logits	**Coverage**	94.73	91.44	91.47	94.96	95.29	96.44	96.88	97.18	98.01	100.00	95.57
		Fraction	16.67	11.51	9.04	8.24	7.81	8.01	8.75	9.67	8.96	11.33	100.00
		Acc. Before	94.73	**78.14**	62.99	52.88	50.00	40.74	39.76	34.85	33.91	30.47	55.35
		Acc. After	94.73	77.22	**65.22**	**57.35**	**51.37**	**45.04**	**40.71**	**37.55**	**34.70**	30.47	**56.68**
	Ours	**Coverage**	94.61	92.23	90.39	92.82	95.85	96.66	96.88	97.46	99.60	100.00	95.02
		Fraction	14.76	12.38	11.49	10.57	11.43	11.00	9.52	7.95	5.95	4.95	100.00
		Acc. Before	94.61	80.54	63.22	50.06	47.04	39.48	37.53	31.19	**29.14**	27.34	55.35
		Acc. After	94.61	**81.02**	**63.95**	**54.32**	**52.54**	**42.72**	**40.15**	**32.99**	28.14	27.34	**57.26**
Phi 3	Logits	**Coverage**	95.25	92.20	91.24	92.83	95.32	96.40	96.21	94.89	97.74	100.00	94.74
		Fraction	17.23	13.24	10.56	10.92	10.40	9.57	9.09	7.67	5.77	5.55	100.00
		Acc. Before	95.25	79.48	62.36	55.43	46.92	45.78	41.64	33.75	31.89	27.78	58.59
		Acc. After	95.25	**81.81**	**67.42**	**61.30**	**53.42**	**48.39**	**42.95**	**34.83**	**32.72**	27.78	**61.25**
	Ours	**Coverage**	94.81	90.79	91.27	92.95	94.73	95.58	95.77	97.28	98.52	100.00	94.53
		Fraction	19.19	13.02	10.47	10.43	10.59	9.39	8.41	7.43	6.40	4.68	100.00
		Acc. Before	94.81	77.12	64.17	54.38	47.31	44.50	39.21	32.11	29.87	25.38	58.59
		Acc. After	94.81	**79.40**	**68.14**	**60.41**	**54.71**	**48.93**	**39.77**	**32.43**	**31.35**	25.38	**61.30**

Table 4: Results for CROQ experiment on MMLU dataset with 10 response options. The CROQ procedure consistently increases accuracy. This effect is more pronounced with using CP-OPT scores (Ours) in comparison to Logits.

Set Size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Overall
Coverage	95.26	91.54	92.37	93.37	92.81	93.87	95.25	95.06	95.73	95.18	97.86	96.70	98.12	98.79	100.00	95.36
Fraction	9.02	7.86	7.46	6.80	5.95	6.01	5.74	6.49	6.94	7.14	6.10	5.76	5.68	5.87	7.17	100.00
Accuracy Before	95.26	83.38	73.45	65.27	58.88	52.77	44.42	42.78	44.27	**35.22**	39.49	32.37	33.19	27.88	**26.49**	52.35
Accuracy After	95.26	**83.99**	**77.11**	**70.86**	**61.08**	**58.10**	**47.73**	**46.80**	**46.84**	34.88	**40.66**	**33.40**	**33.61**	27.88	25.83	**54.21**

Table 5: Results for CROQ experiment on MMLU dataset with 15 response options, using `Llama-3-8B-Instruct` and logits.

Set Size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Overall
Coverage	94.14	93.17	91.12	93.39	93.37	94.72	94.26	95.76	95.75	94.43	96.82	97.13	96.89	98.75	100.00	94.60
Fraction	8.11	8.69	8.69	8.97	8.41	8.77	8.27	8.12	7.54	6.61	5.59	4.14	3.43	2.85	1.82	100.00
Accuracy Before	94.14	84.02	71.04	63.76	54.44	50.34	41.32	41.52	35.59	34.11	**29.30**	28.65	**29.07**	21.25	20.92	52.35
Accuracy After	94.14	**84.29**	**73.77**	**67.20**	**56.28**	**54.40**	**48.49**	**44.44**	**39.06**	**36.62**	28.87	**32.09**	**27.68**	20.83	20.26	**54.74**

Table 6: Results for CROQ experiment on MMLU dataset with 15 response options, using `Llama-3-8B-Instruct` and CP-OPT.

Set Size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Overall
Coverage	95.81	93.16	91.01	91.88	93.16	93.11	93.76	94.11	93.69	94.74	96.08	95.81	93.58	95.91	100.00	93.96
Fraction	11.33	10.23	8.71	8.47	8.15	7.93	7.42	6.44	5.83	5.19	5.15	4.82	4.25	3.19	2.88	100.00
Accuracy Before	95.81	81.55	67.30	58.40	53.86	50.45	41.76	40.52	40.12	38.67	35.02	28.82	**24.30**	21.56	20.58	53.96
Accuracy After	95.81	**85.03**	**73.84**	**66.53**	**62.30**	**54.64**	**48.80**	**44.01**	**44.60**	**42.11**	**37.79**	**30.54**	23.18	**22.68**	20.58	**58.00**

Table 7: Results for CROQ experiment on MMLU dataset with 15 response options, using `Phi-3-4k-mini-Instruct` and logits.

Set Size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Overall
Coverage	95.25	92.64	92.09	92.69	92.22	93.96	94.05	95.06	93.05	96.35	94.20	94.80	93.58	97.15	100.00	94.02
Fraction	10.50	8.87	8.11	7.63	7.32	8.25	8.38	7.45	7.00	6.18	5.73	5.25	4.44	3.33	1.55	100.00
Accuracy Before	95.25	85.27	73.65	66.10	56.24	52.37	46.18	38.54	36.44	34.36	33.75	30.77	23.53	20.28	16.79	53.96
Accuracy After	95.25	**86.08**	**78.18**	**72.94**	**62.40**	**59.28**	**53.40**	**46.50**	**42.03**	**35.70**	**36.02**	**31.67**	23.53	**21.35**	16.79	**57.83**

Table 8: Results for CROQ experiment on MMLU dataset with 15 response options, using `Phi-3-4k-mini-Instruct` and CP-OPT.

B Example Questions and Prompts

B.1 MMLU

The following is an example of an MCQ prompt in the CP-OPT format.

Llama 3 Prompt:

> This question refers to the following information.
> In order to make the title of this discourse generally intelligible, I have translated the term "Protoplasm," which is the scientific name of the substance of which I am about to speak, by the words "the physical basis of life." I suppose that, to many, the idea that there is such a thing as a physical basis, or matter, of life may be novel-so widely spread is the conception of life as something which works through matter. ... Thus the matter of life, so far as we know it (and we have no right to speculate on any other), breaks up, in consequence of that continual death which is the condition of its manifesting vitality, into carbonic acid, water, and nitrogenous compounds, which certainly possess no properties but those of ordinary matter.

Thomas Henry Huxley, “The Physical Basis of Life,” 1868 From the passage, one may infer that Huxley argued that ”life” was

A. essentially a philosophical notion

B. a force that works through matter

C. merely a property of a certain kind of matter

D. a supernatural phenomenon

the correct answer is

Phi 3 Prompt:

<|user|>

This question refers to the following information.
In order to make the title of this discourse generally intelligible, I have translated the term “Protoplasm,” which is the scientific name of the substance of which I am about to speak, by the words “the physical basis of life.” I suppose that, to many, the idea that there is such a thing as a physical basis, or matter, of life may be novel-so widely spread is the conception of life as something which works through matter. ... Thus the matter of life, so far as we know it (and we have no right to speculate on any other), breaks up, in consequence of that continual death which is the condition of its manifesting vitality, into carbonic acid, water, and nitrogenous compounds, which certainly possess no properties but those of ordinary matter.

Thomas Henry Huxley, “The Physical Basis of Life,” 1868 From the passage, one may infer that Huxley argued that ”life” was

A. essentially a philosophical notion

B. a force that works through matter

C. merely a property of a certain kind of matter

D. a supernatural phenomenon

<|end|>
<|assistant|>
the correct answer is

Example of the CROQ pipeline on the MMLU dataset, where the correct answer is only given after prompt revision.

Initial Prompt:

Each of the following are aspects of the McDonaldization of Society EXCEPT:

A. Spatial discrimination
B. Bureaucratic organization that formalizes well-establish division of labor and impersonal structures
C. oxidative phosphorylation.
D. about 1 minute.
E. Competitive inhibition
F. DNA polymerase I
G. A dissolution of hierarchical modes of authority into collaborative teambased decision protocols

H. 1-butene rearranges to 2-butene in solution
I. Rationalization of decisions into cost/benefit analysis structures and away from traditional modes of thinking
J. An intense effort on achieving sameness across diverse markets
the correct answer is

Output:
Prediction: C. oxidative phosphorylation.
Prediction Set: {A, C, D, E, F, G, H}

Revised Prompt:

Each of the following are aspects of the McDonaldization of Society EXCEPT:

A. Spatial discrimination
B. oxidative phosphorylation.
C. about 1 minute.
D. Competitive inhibition
E. DNA polymerase I
F. A dissolution of hierarchical modes of authority into collaborative teambased decision protocols
G. 1-butene rearranges to 2-butene in solution
the correct answer is

Output: Prediction: F. A dissolution of hierarchical modes of authority into collaborative teambased decision protocols

Initial Prompt:

At trial, during the plaintiff's case-in-chief, the plaintiff called as a witness the managing agent of the defendant corporation, who was then sworn in and testified. Defense counsel objected to the plaintiff's questions either as leading or as impeaching the witness. In ruling on the objections, the trial court should

A. sustain all the objections and require the plaintiff to pursue this type of interrogation only during the plaintiff's cross-examination of this witness during the defendant's case-in-chief.
B. Yes, the court will grant it because the plaintiff is not a member of the second class that he set up.
C. The student only, because his conduct was the legal cause of the other driver's death.
D. No, because the common law doctrine of negligence per se does not abrogate the defendant's right to apportion fault under the comparative negligence statute.
E. sustain the leading question objections but overrule the other objections because a party is not permitted to ask leading questions of his own witness at trial.
F. sustain the impeachment questions but overrule the other objections because a party is not permitted to impeach his own witness at trial.', G. when the nephew dies.
H. overrule all the objections because the witness is adverse to the plaintiff and therefore may be interrogated by leading questions and subjected to impeachment.
I. The statute violates the establishment clause of the First Amendment, as incorporated into the Fourteenth Amendment, by adopting the controversial views of particular churches on abortion.
J. The company will prevail because the provision notifying her of the contract is in bold and the contract is easily accessible.
the correct answer is

Output:
Prediction: E. sustain the leading question objections but overrule the other objections because a party is not permitted to ask leading questions of his own witness at trial.

Prediction Set: {E, H}

Revised Prompt:

At trial, during the plaintiff's case-in-chief, the plaintiff called as a witness the managing agent of the defendant corporation, who was then sworn in and testified. Defense counsel objected to the plaintiff's questions either as leading or as impeaching the witness. In ruling on the objections, the trial court should

A. sustain the leading question objections but overrule the other objections because a party is not permitted to ask leading questions of his own witness at trial.
B. overrule all the objections because the witness is adverse to the plaintiff and therefore may be interrogated by leading questions and subjected to impeachment.
the correct answer is

Output: Prediction: B. overrule all the objections because the witness is adverse to the plaintiff and therefore may be interrogated by leading questions and subjected to impeachment.

B.2 ToolAlpaca

Initial Prompt:

Given the API CurrencyBeacon, and the following instruction, "I'm planning a trip to Japan next month, and I want to start budgeting. Can you tell me the current exchange rate from US dollars to Japanese yen, and also provide the average exchange rate for July?" Which of the following functions should you call?

A. timeseries_get Get historical exchange rate data for a specified time range.
B. latest_get Get real-time exchange rates for all supported currencies.
C. historical_get Get historical exchange rate data for a specific date.
D. convert_get Convert an amount from one currency to another.

the correct answer is

Output:
Prediction: A. timeseries_get Get historical exchange rate data for a specified time range.
Prediction Set: {A,B,C}

Revised Prompt:

Given the API CurrencyBeacon, and the following instruction, "I'm planning a trip to Japan next month, and I want to start budgeting. Can you tell me the current exchange rate from US dollars to Japanese yen, and also provide the average exchange rate for July?" Which of the following functions should you call?

A. timeseries_get Get historical exchange rate data for a specified time range.
B. latest_get Get real-time exchange rates for all supported currencies.
C. historical_get Get historical exchange rate data for a specific date.

the correct answer is

Output: Prediction: B. latest_get Get real-time exchange rates for all supported currencies.

Initial Prompt:

Given the API Cataas, and the following instruction, ”I need a funny image for a birthday card. Can you find me a picture of a cat with the text ’Happy Birthday’ on it?” Which of the following functions should you call?

A. findCatByTag Get random cat by tag
B. findCatWithText Get random cat saying text
C. api Will return all cats
D. count Count how many cats
the correct answer is

Output:
Prediction: A. findCatByTag Get random cat by tag
Prediction Set: {A,B}

Revised Prompt:

Given the API Cataas, and the following instruction, ”I need a funny image for a birthday card. Can you find me a picture of a cat with the text ’Happy Birthday’ on it?” Which of the following functions should you call?

A. findCatByTag Get random cat by tag
B. findCatWithText Get random cat saying text

the correct answer is

Output: Prediction: B. findCatWithText Get random cat saying text