Conformal Prediction with Large Language Models for Multi-choice Question Answering

@ ICML 2023 workshop

Kunwoong Kim

2025.2.26.

Department of Statistics, Seoul National University

Preliminaries

Introduction

- LLMs are widely applied to diverse domains and offer powerful performance.
- Issues of LLMs
  - Factually incorrect generation (hallucination)
  - Unpredictable or biased generation
- Hence, uncertainty quantification is crucial for safe deployment of LLMs, especially in high-stakes domains (e.g., healthcare, law).

Contributions

- This study develops a conformal prediction-based uncertainty quantification technique for LLMs, specifically on the MCQA (Multiple Choice Question Answering) task.
  - Multiple Choice Question Answering = 4지선다 객관식
- Experimentally, it shows that the uncertainty provided by conformal prediction can be useful for some downstream tasks (e.g., selective classification).

Conformal Prediction

- **(Calibrated) prediction set**
  - Let $\mathcal{C} : \mathcal{X} \to 2^{|\mathcal{Y}|}$ be a set-valued function that generates a prediction **set** for a given input.
  - Informally, one can define the uncertainty by the set size $|\mathcal{C}(x)|$ for a given input x.
- **Coverage guarantee**
  - For a given desired error rate $\alpha \in (0, 1)$,

$$1 - \alpha \leq \mathbb{P}(Y_{test} \in \mathcal{C}(X_{test}))$$

    where $(X_{test}, Y_{test}) \in \mathcal{D}_{cal}$ is an unseen test data point drawn from the same distribution as the data used to calibrate ($\mathcal{D}_{cal}$).

Conformal Calibration Procedure

- Let $f : \mathcal{X} \to \Delta^{|\mathcal{Y}|}$ be a classifier which provides (softmax) scores, where Δ is a $|\mathcal{Y}|$-dimensional probability simplex.
- Given a data point (x, y), let $S(x, y) = 1 - [f(x)]_y$, where $[f(x)]_y$ is the softmax score at the index of the true class.
- Estimated quantile (using calibration dataset):

$$\hat{q}_\alpha = \text{Quantile}\left(\{s_1, \ldots, s_n\}, \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right),$$

where $s_i := S(x_i, y_i)$.
This is an estimate of the $1 - \alpha$ quantile of the calibration scores.

Conformal Calibration Procedure

- (Calibration) prediction set at inference time: given x (and α), the calibration prediction set is defined as

$$\mathcal{C}(x) = \mathcal{C}(x; \alpha) := \{y \in \mathcal{Y} : S(x, y) \leq \hat{q}_\alpha\}.$$

- Example
  - $\mathcal{Y} = \{0, 1, 2, 3\}$.
  - $f(x) = [0.1, 0.1, 0.4, 0.4]$.
  - $S(x, 0) = S(x, 1) = 0.9$ and $S(x, 2) = S(x, 3) = 0.6$.
  - Given α, if $\hat{q}_\alpha = 0.7$, then $\mathcal{C}(x) = \{2, 3\}$.

Analysis

Experimental setup

- Task
  - MCQA: to predict the correct answer choice out of four possible options.
- Goal
  - To quantify the model uncertainty over the predicted output using conformal prediction.
- Model / Dataset
  - LLaMa-13B: predicting the probabilities for the four options.
  - MMLU benchmark (MCQA questions from 57 domains - college chemistry, medicine, etc). 50% for calibration and 50% for evaluation.
- Measures
  - Conformal prediction set size $\in \{1, 2, 3, 4\}$ with $\alpha = 0.1$
  - Coverage
  - Accuracy

Prompt setup

- One-shot prompt
  - Use 10 different prompts with 10 different one-shot questions, then take average on the predicted probabilities.
  - The one-shot questions are selected among the (1) samples from a real MCQA dataset (MMLU) and (2) generated ones using LLM (GPT-4).

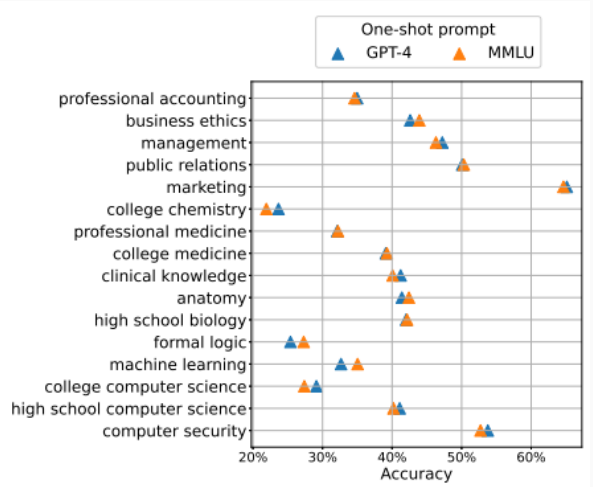

Figure 1: **LLaMA MCQA accuracy is similar for GPT-4 generated questions and real MMLU questions across subjects.** For most MMLU subjects, prediction accuracy using one-shot GPT-4 generated questions is similar to when actual MMLU questions are used in one-shot prompts. Results are averaged over ten randomly selected one-shot GPT-4 and MMLU prompts.

An example of one-shot prompt

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. 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:

Examples of one-shot questions

A.1.1. College Computer Science

GPT-4 Based One-shot Questions

```
Which of the following sorting algorithms has the best average case performance?
    A. Bubble Sort
    B. Quick Sort
    C. Selection Sort
    D. Insertion Sort
The correct answer is option: B

What does the term "Big O Notation" describe in Computer Science?
    A. The speed of a computer
    B. The operating system version
    C. The size of a database
    D. The time complexity of an algorithm
The correct answer is option: D

What does HTTP stand for in terms of web technology?
    A. Hyper Text Transfer Portal
    B. Hyper Transfer Protocol
    C. Hyper Text Transfer Protocol
    D. High Transfer Text Protocol
The correct answer is option: C
```

MMLU Based One-shot Questions

```
An integer c is a common divisor of two integers x and y if
and only if c is a divisor of x and c is a divisor of y.
Which of the following sets of integers could possibly be
the set of all common divisors of two integers?
    A. {-6,-2, -1, 1, 2, 6}
    B. {-6, -2, -1, 0, 1, 2, 6}
    C. {-6, -3, -2, -1, 1, 2, 3, 6}
    D. {-6, -3, -2, -1, 0, 1, 2, 3, 6}
The correct answer is option: C.
```

Examples of one-shot questions

A.1.3. CLINICAL KNOWLEDGE

GPT-4 Based One-shot Questions

```
Which of the following is the most common cause of community-acquired pneumonia?
    A. Streptococcus pneumoniae
    B. Haemophilus influenzae
    C. Klebsiella pneumoniae
    D. Pseudomonas aeruginosa
The correct answer is option: A

Which hormone is primarily responsible for regulating blood calcium levels?
    A. Calcitonin
    B. Parathyroid hormone
    C. Thyroxine
    D. Insulin
The correct answer is option: B

What is the most common cause of acute pancreatitis?
    A. Gallstones
    B. Alcohol
    C. Hypertriglyceridemia
    D. Medications
The correct answer is option: A
```

MMLU Based One-shot Questions

```
The key attribute in successful marathon running is:
    A. strength.
    B. power.
    C. stride length.
    D. stamina.
The correct answer is option D.

Which of the following is NOT a symptom of anaphylaxis?
    A. Stridor.
    B. Bradycardia.
    C. Severe wheeze.
    D. Rash.
The correct answer is option B.
```

Results

- Difference in coverage and set sizes between subjects
  - Subjects with higher accuracies offer lower uncertainties.
  - In contrast, challenging subjects (e.g., formal logic) have higher uncertainties.

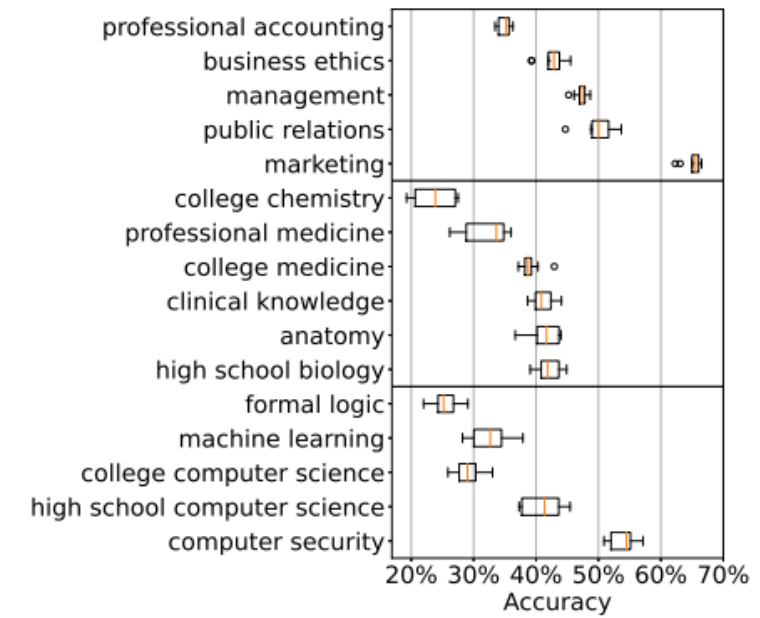

Figure 2: **The accuracy distribution across subjects for ten prompts.** We plot the distribution of accuracy for ten different one-shot prompts.

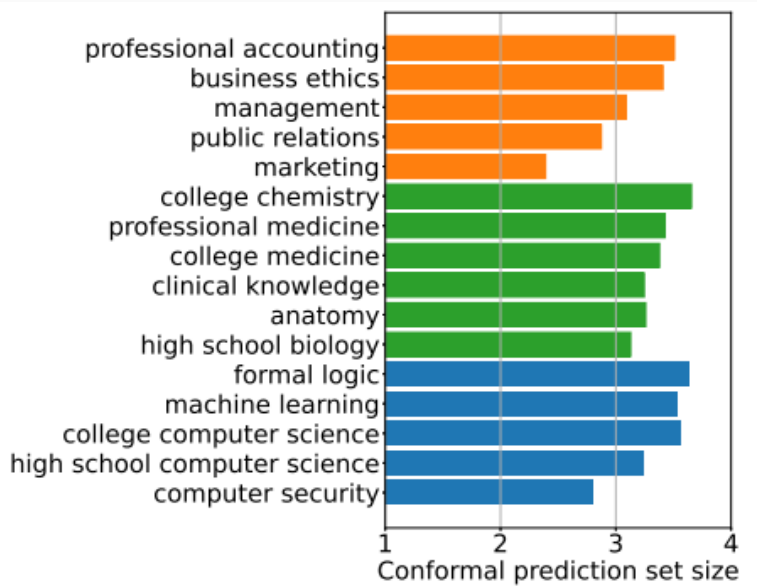

Figure 4: **Uncertainty quantification using prediction set size.** In conformal prediction, a set of predictions is generated for each question. The size of this set indicates how uncertain the model is for a particular question. Larger set sizes denote greater uncertainty, and smaller set sizes denote less uncertainty. The colors denote the three categories of questions.

Results

- Selective classification
  - Negative correlation between set size and top-1 accuracy (Figure 5).
  - Thus, the set size obtained from conformal prediction procedure can filter low-quality predictions in downstream applications for LLMs.

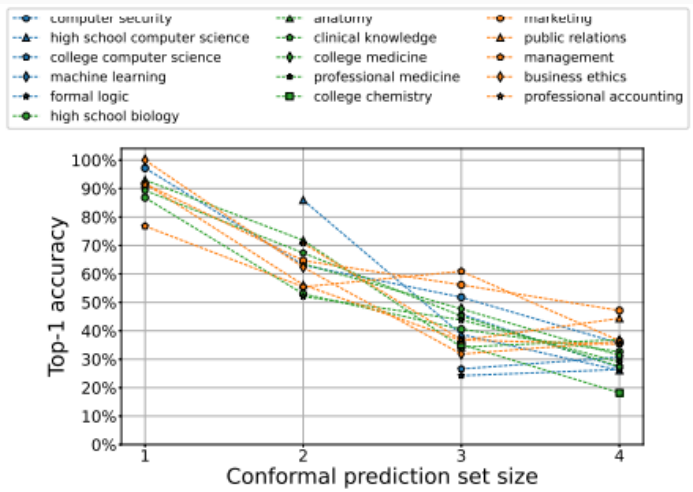

Figure 5: **Top-1 accuracy stratified by prediction set size.** For all subjects, we find a strong correlation between the prediction uncertainty (as measured by set size) and the top-1 accuracy of those predictions. Conformal prediction can be used for selective classification by filtering those predictions in which the model is highly uncertain.

Results

- Size-stratified coverage
  - Top-k highest softmax probabilities are commonly used (Figure 7), however the coverage over the fixed prediction sets of size k is inconsistent.
  - Conformal prediction produces 'adaptive' prediction sets (Figure 6): it tries to attain the proper level of coverage (depending on the chosen error rate α).

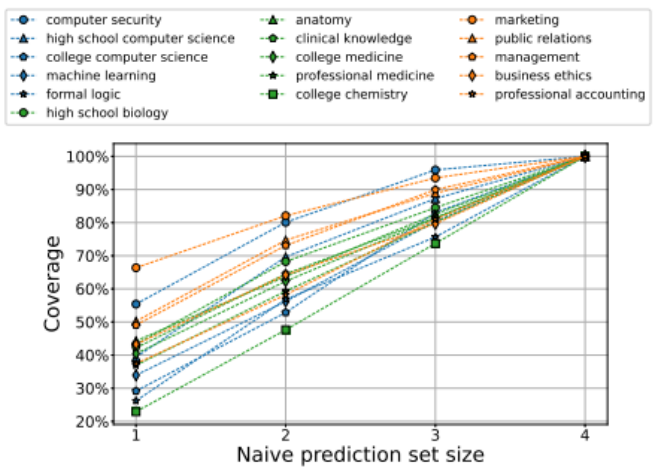

Figure 7: **Coverage of naive top-k prediction sets.** Coverage sharply falls off at smaller set sizes for naive prediction sets constructed by simply taking the top-k softmax scores for all predictions.

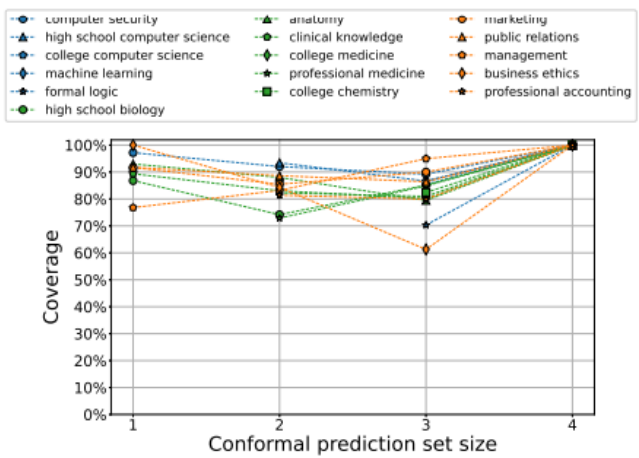

Figure 6: **Stratified coverage at each size of prediction set.** For most subjects, coverage is fairly consistent at all set sizes for prediction sets constructed with the conformal prediction procedure at $\alpha = 0.1$. This means that the true answer is one of the items in the predicted set on average about 90% of the time.

Conclusion

Conclusion

- LLM developers should estimate uncertainty to enhance trustworthiness.
- Uncertainty quantification helps filter low-quality outputs in downstream tasks.
- Conformal prediction can be a promising approach to uncertainty quantification.


THANK YOU