Large language model validity via Enhanced conformal prediction methods (NeurIPS 2024)

Sungeun Lee

Feburary 26, 2025

Seoul National University

Contents

1. Introduction

2. Methods

3. Conclusion

Introduction

Notations

- X : Features given as input.
- P: Prompt provided to the LLM.
- C : Set of sub-claims extracted from the prompt.
- $S(X_i, Y_i)$: A score measures the consistency between the input X_i and the output Y_i.
- $1(\cdot)$: An indicator function that returns 1 if the condition is ture and 0 if it is false.

Introduction

Large language models such as GPT achieve human-level or superior performance in natural language processing tasks.

Current Problems

1. **Hallucination**
   Generating non-existent information as if it were ture.
2. **Inaccuracy**
   Including errors in model output(e.g., numerical values, years, names).
3. **Toxic**
   Socially harmful content such as discrimination, violence, and hate speech.

Introduction

To overcome the three aforementioned issues, *Mohri & Hashimoto (2021)* proposed the following methodology:

1. **Sub-Claims**
   The response generated by the LLM is decomposed into individual claims, each of which is evaluated separately for reliability.
2. **Conformal Prediction**
   Any claims with a confidence level below a certain threshold (e.g., 90%) are discarded.
3. **Score Function**
   A function that quantifies and assesses the factual consistency and reliability of each subclaim generated by a large language model (LLM).
   In *Mohri & Hashimoto (2021)*, a Frequency Score function is used.

Introduction

Challenges in Mohri & Hashimoto (2021)

1. **Marginal Guarantee**

   Performance may vary depending on the domain or prompt.
   - "What are the symptoms of a common flu?"
   - "What are the symptoms of rare genetic disorder X?"

   Due to insufficient guarantees on specific topics, it is difficult to expect consistent performance.

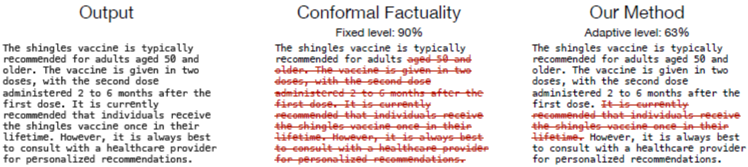
2. **Incomplete Score Function**

   The score function is not fully optimized, which may lead to excessive filtering of even valid claims.

3. **Limitations of Fixed α Value**

   The α value is fixed and does not adjust dynamically depending on the situation. As a result, it may lack confidence in critical safety cases while being overly conservative in general cases.

Introduction



Output

The shingles vaccine is typically recommended for adults aged 50 and older. The vaccine is given in two doses, with the second dose administered 2 to 6 months after the first dose. It is currently recommended that individuals receive the shingles vaccine once in their lifetime. However, it is always best to consult with a healthcare provider for personalized recommendations.

Conformal Factuality

Fixed level: 90%

The shingles vaccine is typically recommended for adults ~~aged 50 and older. The vaccine is given in two doses, with the second dose administered 2 to 6 months after the first dose. It is currently recommended that individuals receive the shingles vaccine once in their lifetime. However, it is always best to consult with a healthcare provider for personalized recommendations.~~

Our Method

Adaptive level: 63%

The shingles vaccine is typically recommended for adults aged 50 and older. The vaccine is given in two doses, with the second dose administered 2 to 6 months after the first dose. ~~It is currently recommended that individuals receive the shingles vaccine once in their lifetime.~~ However, it is always best to consult with a healthcare provider for personalized recommendations.

Figure 1: Left: Original, **Center**: Mohri & Hashimoto Method , **Right**: Proposed Method

- The proposed method (right panel) demonstrates that it maintains high accuracy without excessively removing unnecessary information, in contrast to the existing Conformal Factuality method (center panel).
- The conventional approach tends to eliminate too many claims to ensure reliability, which may reduce its practical applicability.

Methods

Existing Methodologies

The ultimate goal is to obtain highly reliable responses from LLMs.

1. **Optimal Score Function, *Gibbs (2023)***
   Conformal Prediction with Conditional Guarantees, demonstrated that the optimal score function can be defined through differentiation as follows:

   $$g_S = \operatorname*{argmin}_{g \in \mathcal{F}} \frac{1}{n+1} \sum_{i=1}^{n} \ell_\alpha(S(X_i, Y_i) - g(X_i)) + \frac{1}{n+1} \ell_\alpha(S - g(X_{n+1}))$$

   , where $\ell_\alpha(\cdot)$ is the pinball loss function at level α, defined as

   $$\ell_\alpha(r) = (1-\alpha)[r]_+ + \alpha[r]_-,$$

   and $\mathcal{F} = \{\Phi(X)^\top \beta : \beta \in \mathbb{R}^d\}$, $S(X_{n+1}, Y_{n+1}) = S$.
2. **Minimize Prediction Set, *Stutz (2021)***
   Learning Optimal Conformal Classifiers, demonstrated that it is possible to minimize the size of the prediction set while maintaining reliability.

Methods

Conditional Boosting and **Level-Adaptive Conformal Prediction** are techniques designed to maximize the retention of valid information while ensuring statistical reliability.

1. **Conditional Boosting**
   A method that refines the scoring function by differentiating through the conformal filtering process. It dynamically adjusts the filtering threshold based on the characteristics of the input prompt, thereby optimizing the balance between retaining useful sub-claims and removing unreliable ones.

2. **Level-Adaptive Conformal Prediction**
   A technique that adapts the confidence level (or error rate) for each prompt according to its specific properties. This ensures that the reported confidence in the output closely matches the actual reliability of the LLM's response, even when prompt difficulty varies.

Splitting LLM Responses into Sub-Claims

Splitting LLM responses into sub-claims

Example

Sungeun was born in 1999 and is currently pursuing a master's degree in statistics at Seoul National University.

This sentence can be divided into two sub-claims:

- Sungeun was born in 1999.
- Sungeun is currently pursuing a master's degree in statistics at Seoul National University.

A process is needed to assess the reliability of each sub-claim.

Initial Conformity Scoring Function

In this study, the following methods were used to assign a confidence score $p_\theta(P_i, C_{ij})$ to each sub-claim.

1. **Frequency Score**: Evaluates confidence based on the proportion of times a specific sub-claim appears when generating the same question multiple times.
2. **Self-evaluation Score**: The LLM assesses the reliability of the sub-claim itself and returns a probability score.
3. **Log-Probability Score**: Evaluates confidence based on the probability of the LLM generating specific words in the sub-claim.
4. **Ordinal Score**: Assigns confidence based on the order in which the sub-claim appears in the original response.

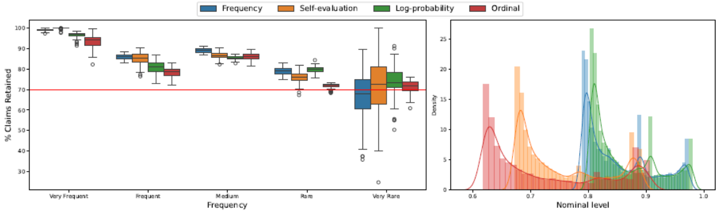


Figure 2: Claim retention rate and score distribution by scoring method

Threshold Calibration via Conformal Prediction

The conventional conformal prediction approach sets a fixed **threshold (τ)** to filter out low-confidence sub-claims.

Conventional Filtering Method

$$F_\tau(C_i) = \{C_{ij} \mid p_\theta(P_i, C_{ij}) \geq \tau\}$$

Limitations:

- Useful sub-claims may be excessively removed due to conservative filtering.
- A uniform threshold ignores variations in confidence across different prompts.

Level-Adaptive Calibration (Overview)

Rather than applying a fixed threshold, we adapt the error level based on the characteristics of each prompt. **Adaptive Calibration Mechanism:**

1. Split the data into a calibration set and a training set.
2. Learn an optimal error level function $\alpha(P)$ for each prompt P using the calibration set.
3. Adjust the filtering threshold so that the desired confidence level $1 - \alpha(P)$ is achieved.

This mechanism ensures that the actual confidence levels match the target levels for similar prompts.

Conditional Boosting

Objective: Maximize the number of retained sub-claims while ensuring reliability.

Optimization Problem

$$\theta^* = \underset{\theta}{\text{argmax}} \sum_{i=1}^{m} \sum_{j=1}^{k_{n+i}} 1\{p_\theta(P_{n+i}, C_{(n+i)j}) \geq \hat{\tau}_i(\theta)\}$$

Key Components:

- $p_\theta(\cdot)$: A parameterized scoring function assigning a confidence score to each sub-claim.
- $\hat{\tau}_i(\theta)$: A filtering threshold determined via quantile regression on the calibration set.
- k_{n+i}: Total number of sub-claims generated for prompt P_{n+i}.

Conditional Boosting

Differentiability of the Filtering Threshold:

Assuming the augmented quantile regression yields a unique, non-degenerate solution for all $S > \hat{\tau}_i(\theta)$, we have:

$$\partial_\theta \hat{\tau}_i(\theta) = \Phi(X_{n+i})^\top \left(\Phi(X)_B^{-1} \, \partial_\theta S_B(\theta)\right),$$

where $S_B(\theta)$ is the score vector computed over the optimal basis B. This differentiability enables gradient descent to optimize θ while maximizing the retention of valid sub-claims.

Final Level-Adaptive Calibration

Calibration Equation

$$P(F(C_{n+1}) \text{ is factual} \mid \alpha_{n+1} \in I) = \mathbb{E}[\alpha_{n+1} \mid \alpha_{n+1} \in I]$$

This equation guarantees that by using the data-driven adaptive error level $\alpha(P)$, the gap between the actual and expected confidence levels is minimized. Consequently, the realized confidence level aligns with the target level $1 - \alpha(P)$ for each prompt.

Conclusion

Conclusion

In summary, this paper makes the following contributions:

- It introduces **Conditional Boosting**, a method that optimizes the filtering threshold (τ) by making it differentiable. This enables dynamic adjustment of the scoring function via gradient descent, allowing for improved retention of valid sub-claims while filtering out unreliable ones.
- It proposes **Level-Adaptive Conformal Prediction**, which adjusts the error level α for each prompt based on its characteristics. This adaptive approach ensures that the reported confidence level of the LLM output closely matches its actual reliability.

However, there are some limitations:

- The proposed methods rely on the i.i.d. assumption for the prompt-response pairs, which may not hold in all real-world applications.
- The performance of the system is sensitive to the quality of the underlying scoring function; further research is needed to develop more robust and accurate score estimation methods.