Conformal Prediction in LLMs

A Review of Four Key Papers

Presenter: Haeyoung Lee

September 4, 2025

Seoul National University

Outline

Challenge: How can we improve the trustworthiness of LLM outputs in high-stakes settings?

Key Question: Can we provide statistical reliability guarantees for LLM predictions?

1. Conformal Prediction with LLMs for Multi-Choice Question Answering (ICML 2023 Workshop)
2. Conformal Language Modeling (ICLR 2024)
3. Language Models with Conformal Factuality Guarantees (ICML 2024)
4. LLM validity via enhanced conformal prediction methods (NeurIPS 2024)

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

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, Andrew Beam

ICML 2023 Workshop(Neural Conversational AI Workshop - What's left to TEACH (Trustworthy, Enhanced, Adaptable, Capable and Human-centric) chatbots?)

Conformal Prediction

Key Idea: Output a *prediction set* rather than a single label.

- **(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. ($x \in \mathcal{X}$: input, $y \in \mathcal{Y}$: label)

- **Coverage Guarantee**
  - For a desired error rate $\alpha \in (0, 1)$,

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

    where $(X_{\text{test}}, Y_{\text{test}}) \sim \mathcal{D}_{\text{cal}}$ is drawn from the same distribution as the calibration set.

Conformal Calibration Procedure

- **Score function** (Least Ambiguous Classifier, LAC):
  - Given a data point (x, y), the score function is defined as

    $$S(x, y) = 1 - f(x)_y,$$

    where $f(x)_y$ is the **softmax** probability corresponding to the **true class**.

- **Calibration step**:
  - Compute nonconformity scores $s_i = S(x_i, y_i)$ on a calibration set.
  - Determine threshold $\hat{q}_\alpha$ as the $(1 - \alpha)$ quantile of $\{s_1, \ldots, s_n\}$:

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

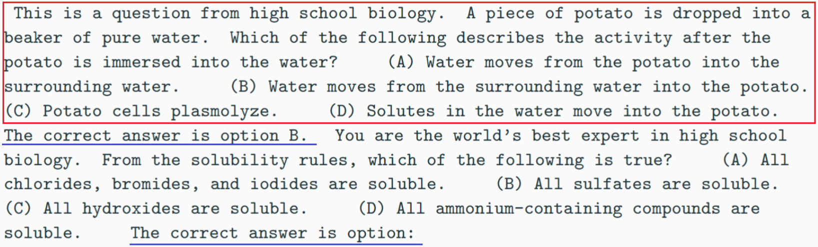
- **At inference time**:
  - given x and α, the calibration prediction set $\mathcal{C}(x)$ is defined as

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

Applying CP to LLMs

- **Task:** Multiple-Choice QA with 4 options (A, B, C, D).
- **Model:** LLaMA-13B
- **Dataset:** MMLU benchmark (MCQA questions from 16 domains - college chemistry, medicine, etc). 50% for calibration and 50% for evaluation.
- **Method:**
  - ▶ Prompt Llama-13B to obtain logits for A–D $\rightarrow$ softmax $f(x)$.
  - ▶ Apply score function $S(x, y)$ to compute nonconformity.
  - ▶ Construct prediction set by including all y with $S(x, y) \leq \hat{q}_\alpha$ (calibration quantile threshold).
- **Measures:** prediction set size (1–4), top-1 accuracy, coverage ($\alpha = 0.1$).

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:

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.

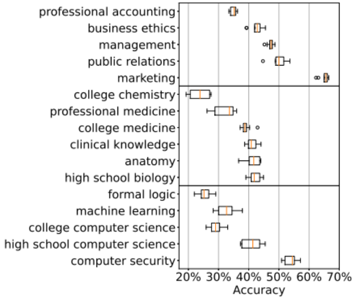


Figure 2: **The accuracy distribution across subjects for ten prompts.** We plot the distribution of accuracy for ten different one-shot prompts.

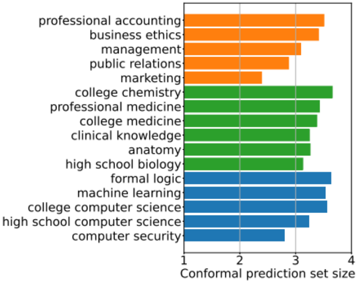


Figure 4: **Uncertainty quantification using prediction set size.** In conformal prediction, a set of predictions is generated for each question. The size of this set indicates how uncertain the model is for a particular question. Larger set sizes denote greater uncertainty, and smaller set sizes denote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

Results

- Negative correlation between set size and top-1 accuracy.

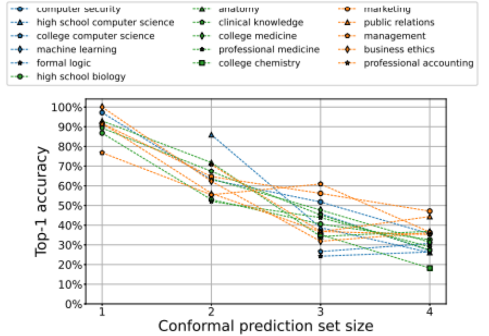


Figure 5: **Top-1 accuracy stratified by prediction set size.** For all subjects, we find a strong correlation between the prediction uncertainty (as measured by set size) and the top-1 accuracy of those predictions. Conformal prediction can be used for selective classification by filtering those predictions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

Conformal Language Modeling

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, Regina Barzilay

ICLR 2024

Introduction

Challenges in applying CP to LMs

- **Infinite output space** — Impossible to enumerate all possible text generations.
- Outputs can be **redundant or hallucinated**, requiring principled filtering.

Proposed Solution: Conformal Language Modeling (CLM)

- **Goal:** Construct prediction sets that contain *at least one admissible response* with high probability.
- **Key idea:** Introduce a parameter set $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ controlling
  - **Rejection Rule** — discard samples failing quality (Q) or diversity (S) thresholds
  - **Stopping Rule** — stop sampling once set confidence $F(C_\lambda)$ exceeds threshold
- **Calibration:** Extend the Learn-Then-Test (LTT) framework to tune λ on calibration data, ensuring that the expected risk of the final prediction set is bounded by ϵ with high probability.

Notation

- X_i : input prompts
- $p_\theta(y \mid x)$: Conditional output distribution defined by the language model.
- $A_i(y)$: a binary random function that measures whether or not a generation y for prompt X_i is good enough (i.e., $A_i(y) = 1$).
- $S(y_k, y_j)$: Text similarity function ($Y \times Y \to \mathbb{R}$) for duplicate detection and diversity.
- $Q(x, y_k)$: Input-conditional quality function ($X \times Y \to \mathbb{R}$) of an individual response.
- $\mathcal{F}$: Set-based confidence function ($2^Y \to \mathbb{R}$) that gives a confidence score for the event $\mathbf{1}\{\exists\, y \in C : A(y) = 1\}$.
- λ_1 : Similarity threshold for filtering redundant samples.
- λ_2 : Quality threshold for rejecting low-quality samples.
- λ_3 : Confidence threshold for stopping criterion.
- n : number of examples in the calibration dataset.
- ϵ : error tolerance
- δ : controls for the sensitivity of algorithm with respect to calibration data.

Conformal Language Modeling

1. **Sampling** — Sample a new candidate response y_k from LM:

$$y_k \sim p_\theta(y \mid x), \quad k = 1, 2, \ldots$$

2. **Accept/Reject** — Update C_k by adding y_k only if it satisfies the quality and diversity criteria:

$$C_k := \begin{cases} C_{k-1} \cup \{y_k\}, & \text{if } Q(x, y_k) \geq \lambda_2 \text{ and } \max_{y_j \in C_{k-1}} S(y_k, y_j) \leq \lambda_1 \\ C_{k-1}, & \text{otherwise} \end{cases}$$

3. **Stopping Rule** — Stop sampling when the set confidence exceeds threshold:

$$F(C_k) \; \geq \; \lambda_3$$

Output: Final prediction set $C_\lambda(X)$ that, with high probability, contains at least one admissible response.

Conformal Language Modeling

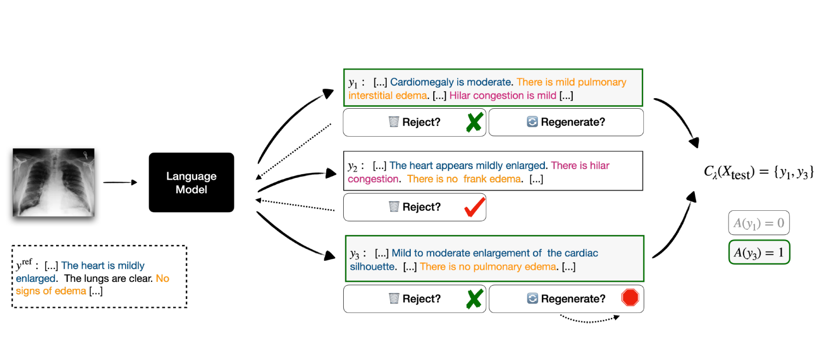

Language Model
y^{ref} : [...] The heart is mildly enlarged. The lungs are clear. No signs of edema [...]
y_1 : [...] Cardiomegaly is moderate. There is mild pulmonary interstitial edema. [...] Hilar congestion is mild [...]
Reject? ✗
Regenerate?
y_2 : [...] The heart appears mildly enlarged. There is hilar congestion. There is no frank edema. [...]
Reject? ✓
y_3 : [...] Mild to moderate enlargement of the cardiac silhouette. [...] There is no pulmonary edema. [...]
Reject? ✗
Regenerate?
$C_\lambda(X_{\text{test}}) = \{y_1, y_3\}$
$A(y_1) = 0$
$A(y_3) = 1$

Figure 1: Illustration of CLM on a test input (radiology report generation). Admissibility labels and the reference report are shown for clarity only.

- Given an X-ray image X_{test}, LM generates multiple candidate reports (y_1, y_2, y_3).
- Candidates are filtered by quality (Q) and diversity (S).
- CLM guarantees that the final set $C_\lambda(X_{test})$ contains at least one admissible report with high probability.

Calibration with Learn Then Test (LTT)

- **Input:** Calibration dataset $D_{cal} = \{(X_i, A_i)\}_{i=1}^{n}$
  - X_i = input prompt
  - $A_i(y)$ = admission function labeling whether candidate y is admissible.
- **Output:** Calibrated thresholds $\hat{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$

1. **Null Hypothesis:**

   $$H_\lambda : R(\lambda) > \epsilon$$

   where $R(\lambda)$ is the true risk.

2. **Empirical Risk:**

   $$\hat{R}_n(\lambda) := \frac{1}{n} \sum_{i=1}^{n} L_i(\lambda)$$

   - $L_i(\lambda) = \mathbf{1}\{\nexists y \in C_\lambda(X_i) : A_i(y) = 1\}$
   - $A_i(y) = 1$: response y is admissible

3. **Binomial Tail P-value:**

   $$p_{BT\lambda} := P\left(\text{Binom}(n, \epsilon) \leq n\hat{R}_n(\lambda)\right)$$

Calibration with Learn Then Test (LTT)

4. **Construct $\Lambda_{\textbf{valid}}$:** Reject H_λ using an FWER-controlling test (e.g., Pareto Testing) at level δ. Λ_{valid} consists of all λ configurations for which the null hypotheses are rejected.
5. **Select Optimal $\hat{\lambda}$:**

$$\hat{\lambda} = \arg \min_{\lambda \in \Lambda_{\text{valid}}} \frac{1}{n} \sum_{i=1}^{n} \left(\rho_1 |C_\lambda(X_i)| + \rho_2 \frac{[S_\lambda(X_i) - S^*(X_i)]_+}{S_\lambda(X_i)} \right)$$

Goal: Choose the most compact and efficient configuration.

- $|C_\lambda(X_i)|$: prediction set size
- $S_\lambda(X_i)$: total number of samples
- $S^*(X_i)$: index of the first admissible response
- $[S_\lambda(X_i) - S^*(X_i)]_+$: number of excess samples after first admissible
- ρ_1, ρ_2: user-defined weights (e.g., $\rho_1 = \rho_2 = 0.5$).

Experimental Setup

- **Task: Radiology Report Generation** - Generate long-form medical reports from chest X-rays, while filtering hallucinated findings.
- **Dataset:** MIMIC-CXR
- **Model:** ViT (image encoder) + GPT-2 (text decoder)
- **Admission Function** $A : \mathcal{Y} \rightarrow \{0, 1\}$**:** Determines whether a generated report is *admissible*.
  - Criterion: **CheXbert** labels of generated report must *exactly match* labels of the reference report.

Experimental Setup

- **Quality Function (Q):** Input-conditional measure of response quality, based on LM likelihood.

$$Q(x, y) = p_\theta(y \mid x)$$

- **Similarity Function (S):** Prevents redundant candidates in the prediction set
  - Uses ROUGE-L to compare a new sample y_k and existing outputs y_j.
  - Rejects if similarity is too high:

$$\max S(y_k, y_j) \leq \lambda_1$$

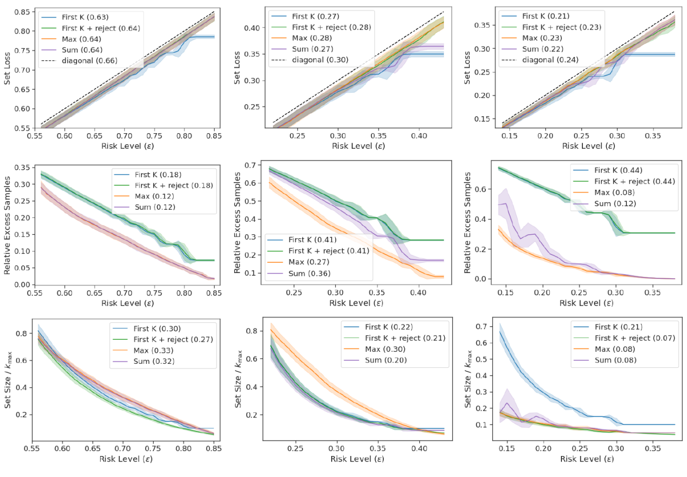
- **Confidence Function (F):** Determines when to stop sampling.

Function	Definition
FIRST-K	Stop after k samples: $F_{\text{FIRST-K}}(C) = \lvert C \rvert$
FIRST-K+REJECT	Same as FIRST-K, but filters duplicates using S.
MAX	Highest-quality candidate: $F_{\text{MAX}}(C) = \max(Q(y))$
SUM	Total quality score: $F_{\text{SUM}}(C) = \sum_{y \in C} Q(y)$

Experimental Setup - Metrics

- **Set Loss:** Measures the probability that the prediction set fails to contain a correct answer.
- **Excess Samples:** Evaluates unnecessary sampling beyond the first correct response.
- **Final Set Size:** Assesses the size of the final prediction set.
- **Computes Area Under the Curve (AUC) over ϵ or α.**

Experimental Results



(a) MIMIC-CXR

(b) CNN/DM

(c) TriviaQA

Figure 2: Conformal sampling results for C_λ as a function of ϵ. We report the loss, relative excess samples, and overall size (normalized by $k_{\max}$). We also report the AUC over achieved/non-trivial ϵ.

Language Models with Conformal Factuality Guarantees

Christopher Mohri & Tatsunori Hashimoto

ICML 2024

Notation

- $\mathcal{X}$: Input space for LM.
- $\mathcal{Y}$: Output space for LM.
- $x \in \mathcal{X}$: An input query or context.
- $y \in \mathcal{Y}$: An output sequence generated by LM.
- y^*: Ground truth or reference knowledge.
- $\alpha \in (0, 1)$: User-specified error rate.
- n: Number of calibration samples.
- $T \subseteq \mathbb{R}$: Set of possible thresholds.
- $L \colon \mathcal{X} \mapsto \mathcal{Y}$: Base language model.
- $S \colon \mathcal{Y} \mapsto 2^{\mathcal{Y}}$: Sub-claim separator.
- $s \colon 2^{\mathcal{Y}} \times \mathcal{Y} \mapsto \mathbb{R}$: Sub-claim scoring function.

Key Idea: Entailment Set

- **Entailment Set:** Each output y defines an entailment set

  $$E(y) := \{ y' \in \mathcal{Y} \, : \, y' \Rightarrow y \}.$$

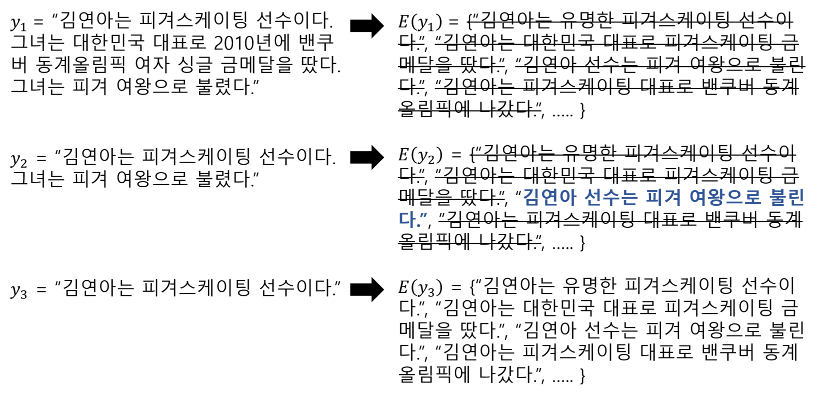
  so if y' is true, then y must also be true.

- **Correctness via Entailment:** An LM output y is **correct** w.r.t. ground truth y^* if

  $$y^* \Rightarrow y \iff y^* \in E(y).$$

- Each y is associated with $E(y)$ containing more specific statements that entail it.

Key Idea: Entailment Set



y_1 = “김연아는 피겨스케이팅 선수이다. 그녀는 대한민국 대표로 2010년에 밴쿠버 동계올림픽 여자 싱글 금메달을 땄다. 그녀는 피겨 여왕으로 불렸다.” ➡ $E(y_1)$ = {~~“김연아는 유명한 피겨스케이팅 선수이다.”~~, ~~“김연아는 대한민국 대표로 피겨스케이팅 금메달을 땄다.”~~, ~~“김연아 선수는 피겨 여왕으로 불린다.”~~, ~~“김연아는 피겨스케이팅 대표로 밴쿠버 동계올림픽에 나갔다.”~~, ... }

y_2 = “김연아는 피겨스케이팅 선수이다. 그녀는 피겨 여왕으로 불렸다.” ➡ $E(y_2)$ = {~~“김연아는 유명한 피겨스케이팅 선수이다.”~~, ~~“김연아는 대한민국 대표로 피겨스케이팅 금메달을 땄다.”~~, **“김연아 선수는 피겨 여왕으로 불린다.”**, ~~“김연아는 피겨스케이팅 대표로 밴쿠버 동계올림픽에 나갔다.”~~, ... }

y_3 = “김연아는 피겨스케이팅 선수이다.” ➡ $E(y_3)$ = {“김연아는 유명한 피겨스케이팅 선수이다.”, “김연아는 대한민국 대표로 피겨스케이팅 금메달을 땄다.”, “김연아 선수는 피겨 여왕으로 불린다.”, “김연아는 피겨스케이팅 대표로 밴쿠버 동계올림픽에 나갔다.”, ... }

- Making LM outputs **less specific** enlarges entailment sets.

Method: Back-off Mechanism F_t

- **Back-off Mechanism**: Make LM outputs progressively **less specific** to enlarge entailment sets.

$$F_t : \mathcal{X} \times \mathcal{Y} \; \mapsto \; \mathcal{Y},$$

$$F_t(x) := F_t(x, L(x)).$$

- The threshold t controls the degree of back-off:
  larger $t \Rightarrow$ less specific output $\Rightarrow$ larger entailment set $E(F_t(x, L(x)))$.
- If the ground truth y^* is in $E(y)$, then y is correct by entailment.
- Define the minimum strictly safe threshold:

$$r(x, y^*) := \inf\{\, t \in \mathcal{T} : y^* \in E(F_t(x, L(x)))\,\}.$$

Calibration: Computing the Conformal Threshold $\hat{q}_\alpha$

Input: $\{(x_i, y_i^*)\}_{i=1}^n$ **Output:** $\hat{q}_\alpha$ (conformal threshold)

1. For each calibration input x_i, generate the LM output y_i.
2. Decompose y_i into sub-claims.
3. Assign a factuality score to each sub-claim (e.g., frequency scoring).
4. List the scores in order across the k sub-claims : $t_1 < t_2 < \cdots < t_K$
5. For each t_k, retain only the sub-claims with score $\geq t_k$.
6. For each t_k, evaluate whether every retained sub-claim is entailed by the ground truth y_i^*; define $r(x_i, y_i^*)$ as the smallest t for which all are entailed.
7. Repeat for all calibration pairs $\{(x_i, y_i^*)\}_{i=1}^n$ to collect $\{r(x_i, y_i^*)\}$.
8. Define the conformal threshold:

$$\hat{q}_\alpha \ = \ \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\text{-th quantile of } \{r(x_i, y_i^*)\}.$$

Test: Generating Conformally Factual Outputs

- For a new input x_{n+1}:
  1. Generate the LM output y_{n+1}.
  2. Decompose y_{n+1} into sub-claims.
  3. Score each sub-claim with s, remove those below threshold $\hat{q}_\alpha$.
  4. Merge remaining sub-claims into a factual sequence.:
- Guarantee:

$$\mathbb{P}(Y^* \in F_{\hat{q}_\alpha}(x, L(x))) \ \geq \ 1 - \alpha.$$

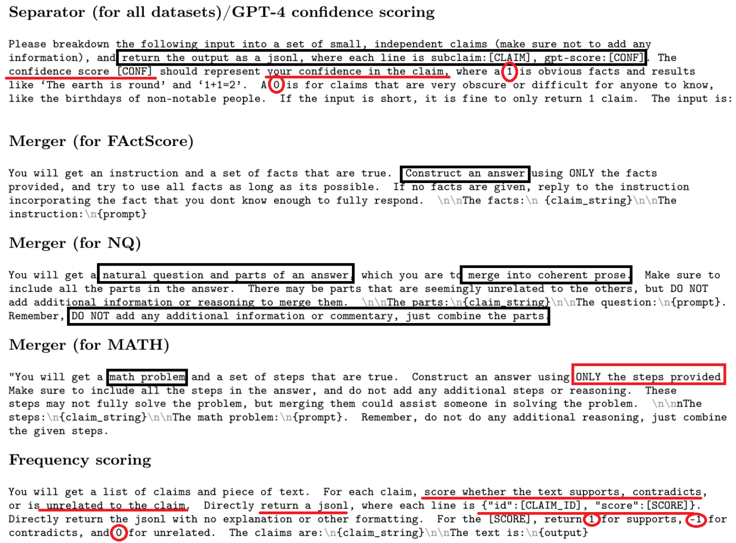
Experimental Setup: Datasets

- **FActScore**: Evaluates factuality in *open-ended biography generation* by decomposing outputs into atomic facts and verifying them against a knowledge source.
- **Natural Questions (NQ)**: Measures factuality in *open-domain question answering*, based on real queries from the Google search engine.
- **MATH**: Demonstrates the framework's applicability to *reasoning tasks* such as mathematical word problems, where solution steps are treated as sub-claims.
- **Annotation**:
  - **50 inputs per dataset** were selected; GPT-4 outputs were manually annotated as *Factual, Subjective, Unverifiable, or False*.
  - **Factual and Subjective** were treated as *entailed*; *Unverifiable and False* were treated as not entailed.
  - All factuality judgments were **verified using Google**.

Experimental Setup

- **Separator / Confidence Scoring**:
  - Decomposes an LM output into independent sub-claims.
  - Simultaneously asks GPT-4 to provide a self-reported confidence score ranging from **0(very obscure) to 1 (obvious fact)**.
- **Frequency Scoring**:
  - Sample K=5 alternate outputs (temp.=1.0). For each sub-claim, GPT-4 evaluates whether the alternate text **supports (+1), contradicts (-1), or is unrelated (0)**. The frequency score is the **sum** of these evaluations.
- **Merger Function**:
  - Recombines selected sub-claims into a coherent sequence.

GPT-4 Prompts



Separator (for all datasets)/GPT-4 confidence scoring

Please breakdown the following input into a set of small, independent claims (make sure not to add any information), and return the output as a jsonl, where each line is subclaim:[CLAIM], gpt-score:[CONF]. The confidence score [CONF] should represent your confidence in the claim, where a 1 is obvious facts and results like ‘The earth is round’ and ‘1+1=2’. A 0 is for claims that are very obscure or difficult for anyone to know, like the birthdays of non-notable people. If the input is short, it is fine to only return 1 claim. The input is:

Merger (for FActScore)

You will get an instruction and a set of facts that are true. Construct an answer using ONLY the facts provided, and try to use all facts as long as its possible. If no facts are given, reply to the instruction incorporating the fact that you dont know enough to fully respond. \n\nThe facts:\n {claim_string}\n\nThe instruction:\n{prompt}

Merger (for NQ)

You will get a natural question and parts of an answer, which you are to merge into coherent prose. Make sure to include all the parts in the answer. There may be parts that are seemingly unrelated to the others, but DO NOT add additional information or reasoning to merge them. \n\nThe parts:\n{claim_string}\n\nThe question:\n{prompt}. Remember, DO NOT add any additional information or commentary, just combine the parts.

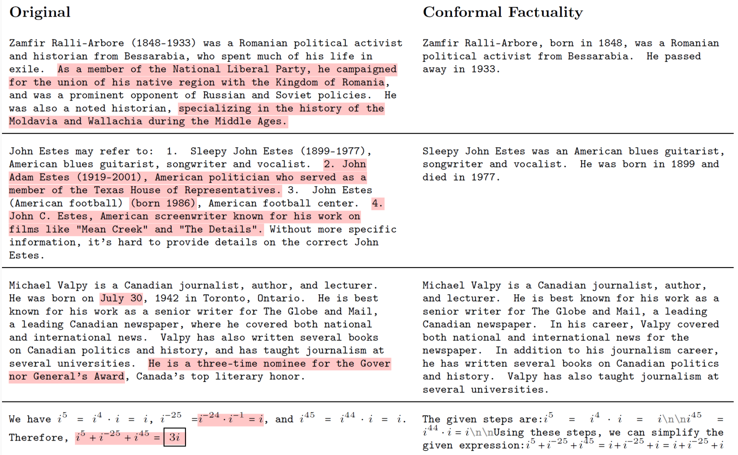
Merger (for MATH)

"You will get a math problem and a set of steps that are true. Construct an answer using ONLY the steps provided. Make sure to include all the steps in the answer, and do not add any additional steps or reasoning. These steps may not fully solve the problem, but merging them could assist someone in solving the problem. \n\nThe steps:\n{claim_string}\n\nThe math problem:\n{prompt}. Remember, do not do any additional reasoning, just combine the given steps.

Frequency scoring

You will get a list of claims and piece of text. For each claim, score whether the text supports, contradicts, or is unrelated to the claim. Directly return a jsonl, where each line is {"id":[CLAIM_ID], "score":[SCORE]}. Directly return the jsonl with no explanation or other formatting. For the [SCORE], return 1 for supports, -1 for contradicts, and 0 for unrelated. The claims are:\n{claim_string}\n\nThe text is:\n{output}

Results



Original	Conformal Factuality
Zamfir Ralli-Arbore (1848-1933) was a Romanian political activist and historian from Bessarabia, who spent much of his life in exile. As a member of the National Liberal Party, he campaigned for the union of his native region with the Kingdom of Romania, and was a prominent opponent of Russian and Soviet policies. He was also a noted historian, specializing in the history of the Moldavia and Wallachia during the Middle Ages.	Zamfir Ralli-Arbore, born in 1848, was a Romanian political activist from Bessarabia. He passed away in 1933.
John Estes may refer to: 1. Sleepy John Estes (1899-1977), American blues guitarist, songwriter and vocalist. 2. John Adam Estes (1919-2001), American politician who served as a member of the Texas House of Representatives. 3. John Estes (American football) (born 1986), American football center. 4. John C. Estes, American screenwriter known for his work on films like "Mean Creek" and "The Details". Without more specific information, it's hard to provide details on the correct John Estes.	Sleepy John Estes was an American blues guitarist, songwriter and vocalist. He was born in 1899 and died in 1977.
Michael Valpy is a Canadian journalist, author, and lecturer. He was born on July 30, 1942 in Toronto, Ontario. He is best known for his work as a senior writer for The Globe and Mail, a leading Canadian newspaper, where he covered both national and international news. Valpy has also written several books on Canadian politics and history, and has taught journalism at several universities. He is a three-time nominee for the Governor General's Award, Canada's top literary honor.	Michael Valpy is a Canadian journalist, author, and lecturer. He is best known for his work as a senior writer for The Globe and Mail, a leading Canadian newspaper. In his career, Valpy covered both national and international news for the newspaper. In addition to his journalism career, he has written several books on Canadian politics and history. Valpy has also taught journalism at several universities.
We have $i^5 = i^4 \cdot i = i$, $i^{-25} = i^{-24} \cdot i^{-1} = i$, and $i^{45} = i^{44} \cdot i = i$. Therefore, $i^5 + i^{-25} + i^{45} = \boxed{3i}$	The given steps are:$i^5 = i^4 \cdot i = i$\n\n$i^{45} = i^{44} \cdot i = i$\n\nUsing these steps, we can simplify the given expression:$i^5 + i^{-25} + i^{45} = i + i^{-25} + i = i + i^{-25} + i$

Table 1: Original GPT-4 vs. Conformal Factuality outputs. False claims (red) are removed.

- Removes hallucinated facts in biographies (FActScore).
- In reasoning (MATH), eliminates incorrect steps, preserving a correct partial proof.

Limitations

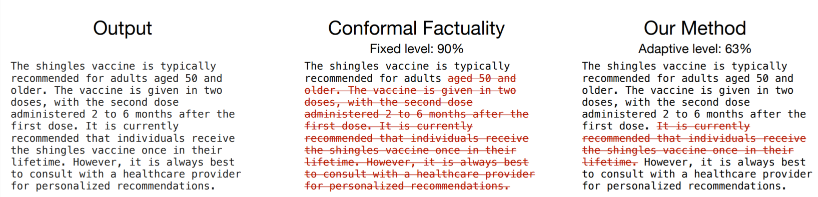
- **Marginal Coverage Guarantee**: The current guarantee is marginal (on average over inputs), not conditional (for every specific input). Factuality may vary across different tasks or user groups.
- **Marginal over Calibration Set Draw**: The guarantee is also marginal over the selection of the calibration set.

Large Language Model Validity via Enhanced Conformal Prediction Methods

John J. Cherian, Isaac Gibbs, Emmanuel J. Candès

NeurIPS 2024

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.

Fig 1. Filtered responses to the question “How often is a shingles vaccine required?” using the marginally valid conformal factuality method of Mohri & Hashimoto (90%) and the proposed adaptive boosted CP method (63%).

- The conventional approach tends to eliminate too many claims to ensure reliability, which may reduce its practical applicability.
- The proposed method demonstrates that it maintains high accuracy without excessively removing unnecessary information.

Notation

- P_i: Prompt text.
- R_i: Raw LM response.
- C_i : Set of parsed sub-claims from R_i.
- W_{ij}: Ground-truth factuality label for sub-claim j of example i ($1 =$ true, $0 =$ false)
- $p(P_i, C_{ij}) \in \mathbb{R}$: Score measuring the quality of claim C_{ij}.
- $\hat{F}(C_i) = F(C_i; \hat{\tau}) := \{ C_{ij} : p(P_i, C_{ij}) \geq \hat{\tau} \}$: Filtered set of claims at cutoff $\hat{\tau}$.
- $S(C_i, W_i) := \inf \big\{ \tau : \ \forall \, C_{ij} \in F(C_i; \tau), \ W_{ij} = 1 \big\}$.
- $X_i = X(P_i, R_i)$: A set of features computed using the prompt and response.
- $\alpha \in (0, 1)$: User-specified error level.
- $\alpha(X_i)$: Prompt-specific target error level, learned via auxiliary model
- $\ell_\alpha(u) = \alpha \cdot \max(u, 0) + (1 - \alpha) \cdot \max(-u, 0)$: Pinball loss.
- n: Number of calibration examples.

Level-Adaptive CP: Error Level Function $\alpha(X)$

Goal: Learn a prompt-dependent error level function $\alpha(X)$ while satisfying a quality criterion (e.g., $\geq 70\%$ claim retention).

1. **Split** the calibration data Z into two parts: Z_1 and Z_2.
2. **From Z_1:** Further divide Z_1 into two folds $Z_1^{(1)}$ and $Z_1^{(2)}$.
3. **From $Z_1^{(1)}$:** Choose a grid $\mathcal{A} \subset (0,1)$ (e.g., $\{0.01, \ldots, 0.99\}$). For each $\alpha \in \mathcal{A}$, run conformal calibration to obtain the level-specific cutoff $\hat{\tau}_\alpha$.
4. **From $Z_1^{(2)}$:** For each example i and each $\alpha \in \mathcal{A}$, compute
$$Q(C_i, \hat{\tau}_\alpha) := \mathbf{1}\left\{ \frac{|F(C_i; \hat{\tau}_\alpha)|}{|C_i|} \geq 0.7 \right\}.$$
5. **Minimal level satisfying quality:**
$$\alpha_i^\star := \inf\{\alpha \in \mathcal{A} : Q(C_i, \hat{\tau}_\alpha(X_i)) = 1\}.$$
6. **Train a regression model:** learn $\widehat{\alpha}(X)$ from inputs X_i with targets $\alpha_i^\star$.

Level-Adaptive CP: Quantile Function g_S

Goal: Learn $g_S : \mathcal{X} \to \mathbb{R}$ that returns the $(1 - \alpha(X))$ conditional quantile of the conformity score S.

- **Conformity scores on Z_2:**

$$S(C_i, W_i) = \inf\{\tau : \ \forall\, C_{ij} \in F(C_i; \tau), \ W_{ij} = 1\}.$$

- **Adaptive Quantile Regression:**

$$g_S = \arg\min_{g \in \mathcal{F}} \ \frac{1}{n+1} \sum_{i=1}^{n} \ell_{\hat{\alpha}(X_i)}\big(S(C_i, W_i) - g(X_i)\big) + \frac{1}{n+1} \ell_{\hat{\alpha}(X_{n+1})}\big(S - g(X_{n+1})\big),$$

where *pinball loss:* $\ell_\alpha(u) = (1-\alpha)[u]_+ + \alpha[u]_-$

- **Test-time cutoff:**

$$\hat{\tau}_{\hat{\alpha}(X_{n+1})}(X_{n+1}) := \sup\{\, S \ : \ S \le g_S(X_{n+1})\}.$$

Inference Example

Prompt: How often is a shingles vaccine required?

Generated sub-claims:

- C_1: The shingles vaccine is recommended for adults aged 50 and older. (score: 0.80)
- C_2: The vaccine is given in two doses, second one 2 months after the first. (score: 0.70)
- C_3: Individuals receive the shingles vaccine once in their lifetime. (score: 0.40)
- C_4: It is best to consult a healthcare provider for personal recommendations. (score: 0.90)

Learned error level: $\hat{\alpha}(X) = 0.37 \quad (1 - \hat{\alpha} = 0.63)$
Estimated threshold: $\hat{\tau}_{\hat{\alpha}(X)} = 0.50$

Filtering decision using $p(X, C_i)$:

- $p(X, C_1) = 0.80 > 0.50 \Rightarrow$ retain
- $p(X, C_2) = 0.70 > 0.50 \Rightarrow$ retain
- $p(X, C_3) = 0.40 < 0.50 \Rightarrow$ filter out
- $p(X, C_4) = 0.90 > 0.50 \Rightarrow$ retain

Conditional Boosting

- Combine multiple simple scorers into a single scorer to keep more *true* sub-claims.
  - *Frequency:* how often other generations support/contradict the claim
  - *Self-evaluation:* the model's own probability/confidence for the claim
  - *Token log-probability:* internal likelihood of True or False
  - *Ordinal:* position/order the claim appears in the answer

Results

Compared to fixed- CP and prior methods:

- Level-adaptive method retains more correct claims at nominal coverage.
- Conditional boosting further improves recall without loss of validity.

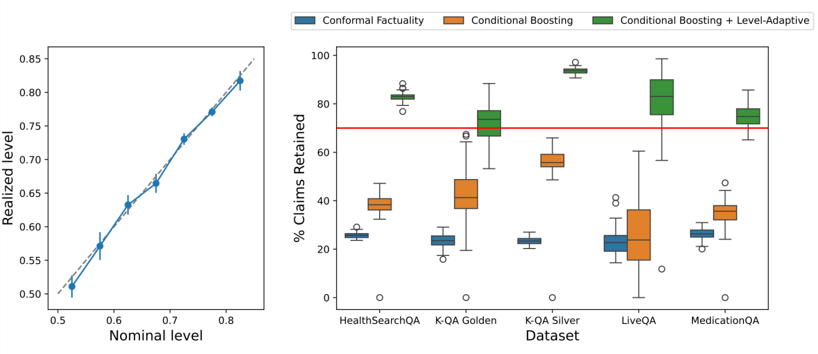


Fig 1. Performance on the Wikipedia Biographies dataset

Thank you!