



Multi-LLM Adaptive Conformal Inference for Reliable LLM Responses



Kangjun Noh¹, Seongchan Lee², Ilmun Kim², and Kyungwoo Song^{1*}

¹Department of Applied Statistics & Data Science, Yonsei University

²Department of Mathematical Sciences, KAIST

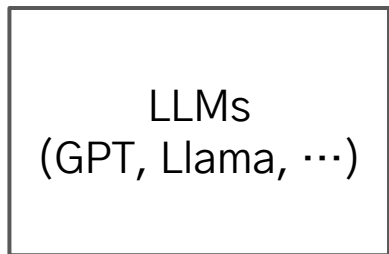
<https://mlai.yonsei.ac.kr/>

One way for Hallucination Detection: Filtering False-Claims

Multi-LLM Adaptive Conformal Inference for Reliable LLM Responses

Prompt

What is the difference between amitriptyline and alprazolam?



LLM output

Amitriptyline and alprazolam are both medications, but they are used to treat different conditions and have different potential side effects. Amitriptyline and alprazolam have different mechanisms of action. Amitriptyline is a tricyclic antidepressant used to treat depression and certain types of chronic pain, **~~but it is not used for anxiety disorders.~~ [False]** Alprazolam, on the other hand, is a benzodiazepine medication primarily used to treat anxiety disorders and panic attacks. It works by enhancing the effects of a neurotransmitter in the brain called GABA, which helps to reduce anxiety and promote relaxation. Both medications can cause side effects including drowsiness and dizziness, **~~though confusion is not a typical side effect. [False]** ~~ However, alprazolam is more likely to cause dependence and withdrawal symptoms, ...

One way for Hallucination Detection: Filtering False-Claims

Multi-LLM Adaptive Conformal Inference for Reliable LLM Responses

LLM output

Amitriptyline and alprazolam are both medications, but they are used to treat different conditions and have different potential side effects. Amitriptyline and alprazolam have different mechanisms of action. Amitriptyline is a tricyclic antidepressant used to treat depression and certain types of chronic pain, ~but it is not used for anxiety disorders. ~ Alprazolam, on the other hand, is a benzodiazepine medication primarily used to treat anxiety disorders and panic attacks. It works by enhancing the effects of a neurotransmitter in the brain called GABA, which helps to reduce anxiety and promote relaxation. Both medications can cause side effects including drowsiness and dizziness, ~though confusion is not a typical side effect. ~ However, alprazolam is more likely to cause dependence and withdrawal symptoms, ...

Hallucination detection

- Whether each subclaim is true or not.
- Filter-out False-Claims.

We want

Close to target Coverage (90%) and High Retention

- **Coverage** is the probability that, after filtering, all retained claims are actually correct.
- **Retention** is the fraction of the original claims that survive the filter, i.e., how much of the text we keep instead of discarding.

Filtering False-Claims: Coverage and Retention

Multi-LLM Adaptive Conformal Inference for Reliable LLM Responses

Good Coverage
Low Retention

LLM output

Amitriptyline and alprazolam are both medications, but they are used to treat different conditions and have different potential side effects. Amitriptyline and alprazolam have different mechanisms of action. Amitriptyline is a tricyclic antidepressant used to treat depression and certain types of chronic pain, ~~but it is not used for anxiety disorders.~~ Alprazolam, on the other hand, is a benzodiazepine medication primarily used to treat anxiety disorders and panic attacks. It works by enhancing the effects of a neurotransmitter in the brain called GABA, which helps to reduce anxiety and promote relaxation. Both medications can cause side effects including drowsiness and dizziness, ~~though confusion is not a typical side effect.~~ However, alprazolam is more likely to cause dependence and withdrawal symptoms, ...



Hallucination Detection

Amitriptyline and alprazolam are both medications, [T] ~~but they are used to treat different conditions [T] and have different potential side effects. [T] Amitriptyline and alprazolam have different mechanisms of action. [T] Amitriptyline is a tricyclic antidepressant [T] used to treat depression [T] and certain types of chronic pain, [T] ~~but it is not used for anxiety disorders.~~ [F]~~ Alprazolam, on the other hand, is a benzodiazepine medication [T] primarily used to treat anxiety disorders [T] and panic attacks. [T] It works by enhancing the effects of a neurotransmitter in the brain [T] called GABA, [T] which helps to reduce anxiety [T] and promote relaxation. [T] Both medications can cause side effects [T] including drowsiness [T] and dizziness, [T] ~~though confusion is not a typical side effect.~~ [F]~~ However, alprazolam is more likely to cause dependence [T] and withdrawal symptoms, [T] ...~~

Bad Coverage
High Retention

LLM output

Amitriptyline and alprazolam are both medications, but they are used to treat different conditions and have different potential side effects. Amitriptyline and alprazolam have different mechanisms of action. Amitriptyline is a tricyclic antidepressant used to treat depression and certain types of chronic pain, ~~but it is not used for anxiety disorders.~~ Alprazolam, on the other hand, is a benzodiazepine medication primarily used to treat anxiety disorders and panic attacks. It works by enhancing the effects of a neurotransmitter in the brain called GABA, which helps to reduce anxiety and promote relaxation. Both medications can cause side effects including drowsiness and dizziness, ~~though confusion is not a typical side effect.~~ However, alprazolam is more likely to cause dependence and withdrawal symptoms, ...



Hallucination Detection

Amitriptyline and alprazolam are both medications, [T] but they are used to treat different conditions [T] and have different potential side effects. [T] Amitriptyline and alprazolam have different mechanisms of action. [T] Amitriptyline is a tricyclic antidepressant [T] used to treat depression [T] and certain types of chronic pain, [T] **~~but it is not used for anxiety disorders.~~ [F]** Alprazolam, on the other hand, is a benzodiazepine medication [T] primarily used to treat anxiety disorders [T] and panic attacks. [T] It works by enhancing the effects of a neurotransmitter in the brain [T] called GABA, [T] which helps to reduce anxiety [T] and promote relaxation. [T] Both medications can cause side effects [T] including drowsiness [T] and dizziness, [T] **~~though confusion is not a typical side effect.~~ [F]** However, alprazolam is more likely to cause dependence [T] and withdrawal symptoms, [T] ...

LLM output

Amitriptyline and alprazolam are both medications, but they are used to treat different conditions and have different potential side effects. Amitriptyline and alprazolam have different mechanisms of action. Amitriptyline is a tricyclic antidepressant used to treat depression and certain types of chronic pain, ~~but it is not used for anxiety disorders.~~ Alprazolam, on the other hand, is a benzodiazepine medication primarily used to treat anxiety disorders and panic attacks. It works by enhancing the effects of a neurotransmitter in the brain called GABA, which helps to reduce anxiety and promote relaxation. Both medications can cause side effects including drowsiness and dizziness, ~~though confusion is not a typical side effect.~~ However, alprazolam is more likely to cause dependence and withdrawal symptoms, ...



Hallucination Detection

Amitriptyline and alprazolam are both medications, [T] but they are used to treat different conditions [T] and have different potential side effects. [T] Amitriptyline and alprazolam have different mechanisms of action. [T] Amitriptyline is a tricyclic antidepressant [T] used to treat depression [T] and certain types of chronic pain, [T] ~~**but it is not used for anxiety disorders.**~~ [F] Alprazolam, on the other hand, is a benzodiazepine medication [T] primarily used to treat anxiety disorders [T] and panic attacks. [T] It works by enhancing the effects of a neurotransmitter in the brain [T] called GABA, [T] which helps to reduce anxiety [T] and promote relaxation. [T] Both medications can cause side effects [T] including drowsiness [T] and dizziness, [T] ~~**though confusion is not a typical side effect.**~~ [F] However, alprazolam is more likely to cause dependence [T] and withdrawal symptoms, [T] ...

TL;DR: Our Filtering Model, MACI, achieves both Good Coverage & High Retention

Multi-LLM Adaptive Conformal Inference for Reliable LLM Responses

“LLMs occasionally generate responses containing false-claims”

“Covered” means: No False-claim in the filtered output

Prompt: “What is the difference between amitriptyline and alprazolam?”

BCI

(Mohri and Hashimoto et al.)

Amitriptyline and alprazolam are both medications, [T] but they are used to treat different conditions [T] and have different potential side effects [T]. Amitriptyline and alprazolam have different mechanisms of action. [T] Amitriptyline is a tricyclic antidepressant [T] used to treat depression [T] and certain types of chronic pain, [T] **but it is not used for anxiety disorders.** [F] Alprazolam, on the other hand, is a benzodiazepine medication [T] primarily used to treat anxiety disorders [T] and panic attacks. [T] It works by enhancing the effects of a neurotransmitter in the brain [T] called GABA, [T] which helps to reduce anxiety [T] and promote relaxation. [T] Both medications can cause side effects [T] including drowsiness [T] and dizziness, [T] **though confusion is not a typical side effect.** [F] However, alprazolam is more likely to cause dependence [T] and withdrawal symptoms, [T] ...

Covered

CCI

(Cherian et al.)

Amitriptyline and alprazolam are both medications, [T] but they are used to treat different conditions [T] and have different potential side effects. [T] Amitriptyline and alprazolam have different mechanisms of action. [T] Amitriptyline is a tricyclic antidepressant [T] used to treat depression [T] and certain types of chronic pain, [T] **but it is not used for anxiety disorders.** [F] Alprazolam, on the other hand, is a benzodiazepine medication [T] primarily used to treat anxiety disorders [T] and panic attacks. [T] It works by enhancing the effects of a neurotransmitter in the brain [T] called GABA, [T] which helps to reduce anxiety [T] and promote relaxation. [T] Both medications can cause side effects [T] including drowsiness [T] and dizziness, [T] **though confusion is not a typical side effect.** [F] However, alprazolam is more likely to cause dependence [T] and withdrawal symptoms, [T] ...

Miscovered

MACI

(Ours)

Amitriptyline and alprazolam are both medications, [T] but they are used to treat different conditions [T] and have different potential side effects. [T] Amitriptyline and alprazolam have different mechanisms of action. [T] Amitriptyline is a tricyclic antidepressant [T] used to treat depression [T] and certain types of chronic pain, [T] **but it is not used for anxiety disorders.** [F] Alprazolam, on the other hand, is a benzodiazepine medication [T] primarily used to treat anxiety disorders [T] and panic attacks. [T] It works by enhancing the effects of a neurotransmitter in the brain [T] called GABA, [T] which helps to reduce anxiety [T] and promote relaxation. [T] Both medications can cause side effects [T] including drowsiness [T] and dizziness, [T] **though confusion is not a typical side effect.** [F] However, alprazolam is more likely to cause dependence [T] and withdrawal symptoms, [T] ...

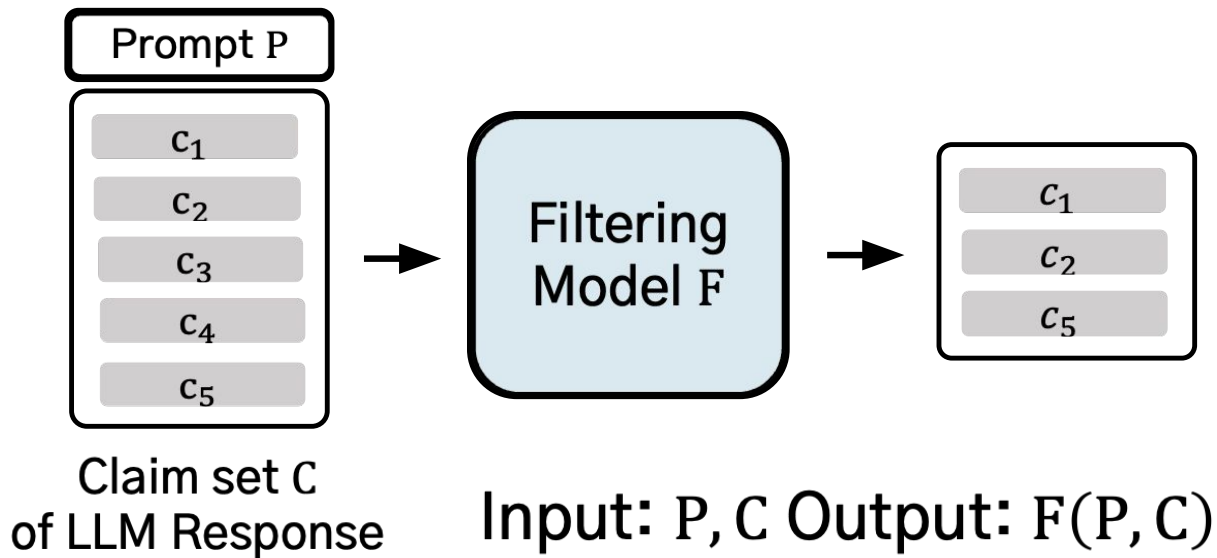
Covered

Coverage Constraint: $\mathbb{P}(\text{no false-claim in filtered set } F) \geq 1 - \alpha$

Problem Setting

Multi-LLM Adaptive Conformal Inference for Reliable LLM Responses

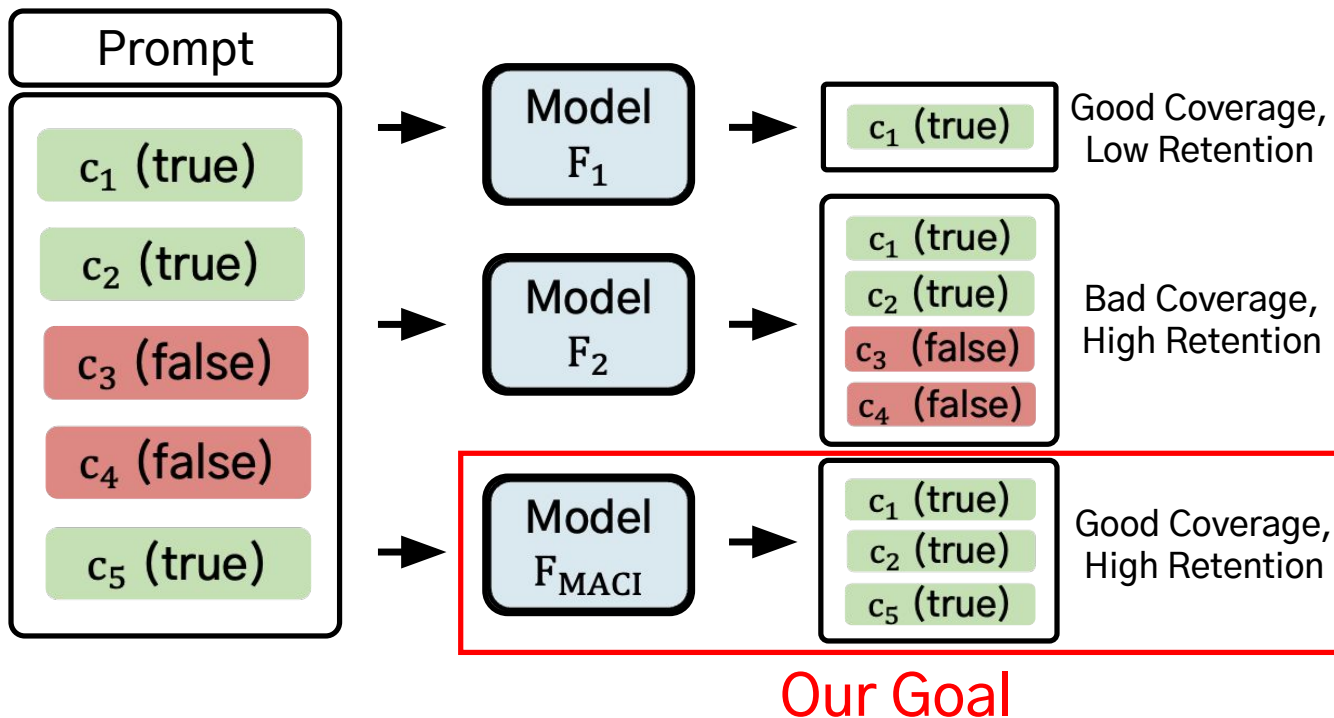
Filtering **false claims** to guarantee the factuality of LLM Response



$$\mathbb{P}(\{\forall c_i \in F(P, C) : c_i \text{ is true}\}) \geq 1 - \alpha$$

Our Goal: Good Coverage & High Retention

Multi-LLM Adaptive Conformal Inference for Reliable LLM Responses

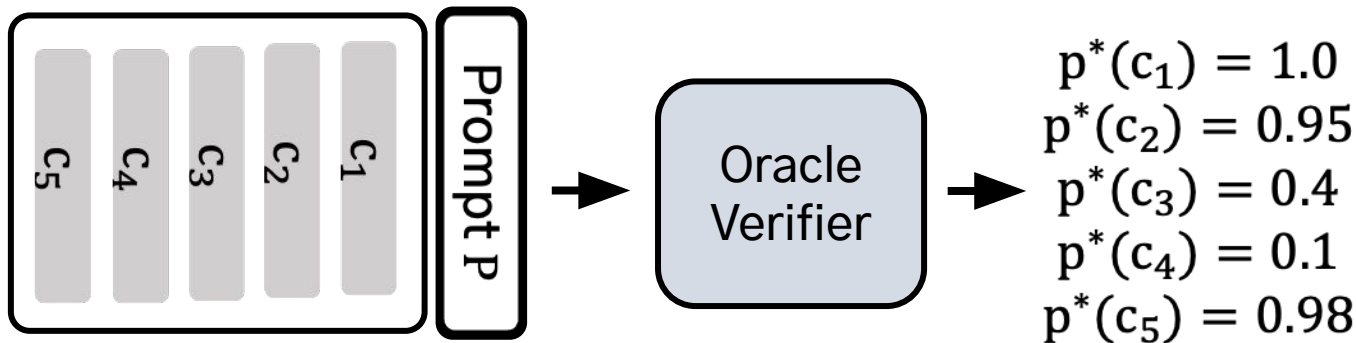


Motivation: Oracle Filtering Algorithm (Oracle Verifier)

Multi-LLM Adaptive Conformal Inference for Reliable LLM Responses

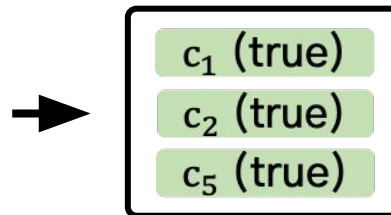
Oracle verifier knows probability that each claim is true, given prompt & claim set.

Oracle verifier can allocate $p^*(c) := \mathbb{P}(c \text{ is true} \mid P, C)$



Algorithm

$$p^*(c_1) \times p^*(c_5) \times p^*(c_2) > 0.90$$
$$p^*(c_1) \times p^*(c_5) \times p^*(c_2) \times p^*(c_3) < 0.90$$



**Optimal Coverage,
Optimal Retention!**

Our Method 1: Mimicking Oracle Factuality-Score by Multi-LLM Ensemble

Multi-LLM Adaptive Conformal Inference for Reliable LLM Responses

Step 1: Gathering high-quality Factuality-Scores by Multi-LLM Ensemble

Multi-LLM Verifier

Ensemble each LLM's output probability for each claim to be true!



Model 1



Model 2

...

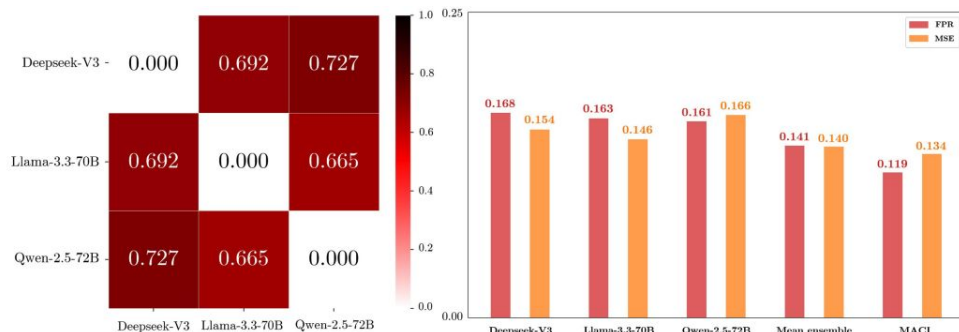


Model m

$$\hat{p}(c) := \sum_{i=1}^m w_m \mathbb{P}_i(c \text{ is true} \mid P, C), \quad \sum w = 1$$

$\langle \hat{p}(c) \rangle$: Factuality-Score for claim c

Why Multi-LLM Ensemble? High disagreement for False-Claims!!



Multi-LLM Ensemble boosts the quality of Factuality-Scores

Our Method 2: Mimicking Oracle Conformity-Score by Adaptive Conformity Score

Multi-LLM Adaptive Conformal Inference for Reliable LLM Responses

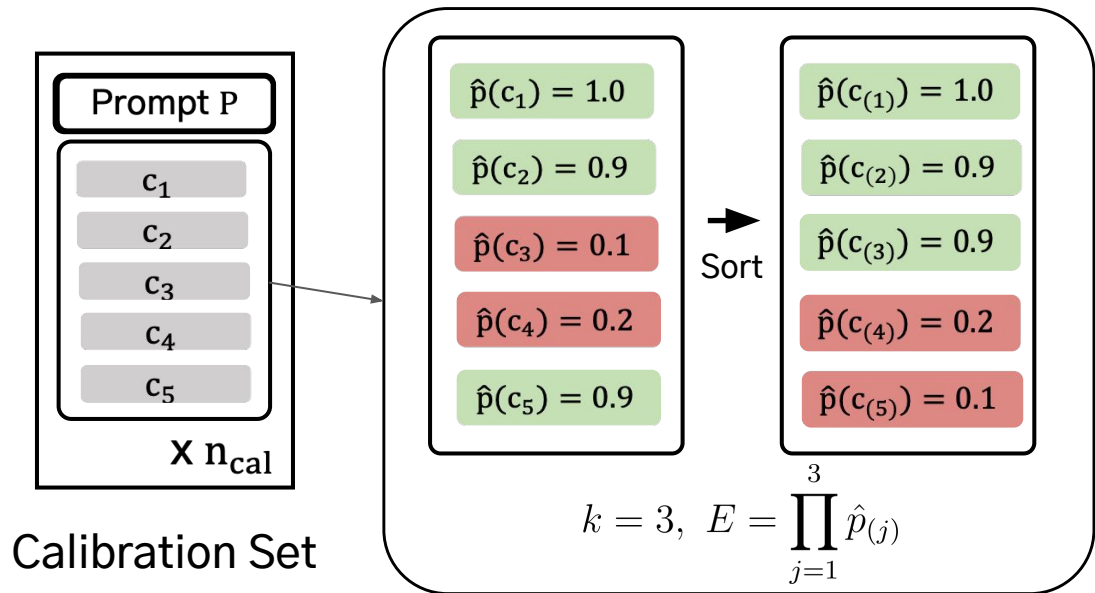
Step 2: Calculate Adaptive Conformity Scores

$$E = \prod_{j=1}^k \hat{p}_{(j)}$$

Definition of conformity score (new!)

What is different compared to existing approaches?
([Hashimoto et al.](#), [Cherian et al.](#))

1. Same procedure with **Oracle Filtering Algorithm**
2. Utilize **global uncertainty information**, relationship of claims



< Example: Calculating Adaptive Conformity Score >

Our Method 2: Mimicking Oracle Conformity–Score by Adaptive Conformity Score

Multi-LLM Adaptive Conformal Inference for Reliable LLM Responses

Step 3: Calculate Conformal Cutoff and Filter-out False-like Claims

$$E_{(1)} \leq E_{(2)} \leq \dots \leq E_{(n_{cal})}$$

$$\hat{Q}_{1-\alpha} = E_{(\lceil (n_{cal}+1)(1-\alpha) \rceil)}$$

and conduct filtering with the rule below (same with Oracle Filtering)

$$F(P, R) = \arg \max_{C' \subseteq C} |C'| \quad \text{s.t.} \quad \prod_{c_i \in C'} \hat{p}(c_i) \geq \hat{Q}_{1-\alpha}$$

with Exchangeability assumption,

$$\mathbb{P}(\{\forall c_i \in F(P, C) : c_i \text{ is true}\}) \geq 1 - \alpha$$

Empirical Result: MACI achieves both Best Coverage & Highest Retention!!

Multi-LLM Adaptive Conformal Inference for Reliable LLM Responses

Group	Target Coverage: 80% ($\alpha = 0.2$)						Target Coverage: 90% ($\alpha = 0.1$)						Target Coverage: 95% ($\alpha = 0.05$)					
	BCI		CCI		MACI		BCI		CCI		MACI		BCI		CCI		MACI	
	Cov.	Ret.	Cov.	Ret.	Cov.	Ret.	Cov.	Ret.	Cov.	Ret.	Cov.	Ret.	Cov.	Ret.	Cov.	Ret.	Cov.	Ret.
MedLFQA	0.80●	0.06	0.81●	0.56	0.80●	0.71	0.90●	0.02	0.90●	0.31	0.90●	0.50	0.95●	0.01	0.95●	0.18	0.95●	0.30
Medical Content																		
Info	0.81●	0.06	0.76↓	0.54	0.80●	0.70	0.91●	0.02	0.86↓	0.30	0.90●	0.48	0.96●	0.01	0.93↓	0.18	0.95●	0.30
Interpret	0.80●	0.07	0.84↑	0.58	0.79●	0.69	0.89●	0.03	0.93↑	0.33	0.90●	0.47	0.94●	0.01	0.96●	0.21	0.96●	0.26
Action	0.79●	0.06	0.85↑	0.49	0.80●	0.73	0.90●	0.02	0.92↑	0.27	0.90●	0.53	0.96●	0.01	0.96●	0.16	0.95●	0.33
False-Claim Risk																		
Low	0.84↑	0.07	0.83↑	0.68	0.79●	0.78	0.94↑	0.03	0.91●	0.41	0.89●	0.52	0.97↑	0.01	0.95●	0.28	0.95●	0.37
Medium	0.83↑	0.06	0.81●	0.66	0.79●	0.70	0.89●	0.03	0.90●	0.39	0.91●	0.46	0.94●	0.01	0.95●	0.25	0.95●	0.31
High	0.73↓	0.06	0.78↓	0.43	0.80●	0.64	0.88↓	0.01	0.89●	0.22	0.89●	0.41	0.94●	0.01	0.94●	0.12	0.95●	0.26
WikiBio	0.81●	0.02	0.79●	0.19	0.81●	0.43	0.90●	0.01	0.89●	0.11	0.90●	0.25	0.95●	0.01	0.93↓	0.06	0.95●	0.13
View Count																		
Low	0.74↓	0.03	0.79●	0.18	0.81●	0.36	0.87↓	0.01	0.88↓	0.11	0.91●	0.21	0.94●	0.01	0.92↓	0.06	0.96●	0.11
Medium	0.84↑	0.02	0.78↓	0.19	0.81●	0.46	0.91●	0.01	0.88↓	0.11	0.91●	0.24	0.95●	0.01	0.92↓	0.06	0.95●	0.12
High	0.85↑	0.02	0.81●	0.20	0.81●	0.51	0.91●	0.01	0.92↑	0.12	0.91●	0.24	0.95●	0.01	0.95●	0.07	0.96●	0.12
False-Claim Risk																		
Low	0.81●	0.03	0.80●	0.21	0.82↑	0.40	0.90●	0.01	0.90●	0.11	0.90●	0.23	0.95●	0.01	0.93↓	0.07	0.94●	0.17
Medium	0.81●	0.02	0.78↓	0.19	0.81●	0.42	0.91●	0.01	0.89●	0.11	0.90●	0.25	0.95●	0.01	0.93↓	0.06	0.95●	0.12
High	0.81●	0.02	0.79●	0.18	0.81●	0.45	0.89●	0.01	0.88↓	0.11	0.90●	0.28	0.94●	0.01	0.92↓	0.06	0.96●	0.09

- MACI achieves both **Best Coverage & Highest Retention** among the baselines throughout Medical, CommonSense, Expert Knowledge datasets.

- Conformal inference can be effectively utilized to detect hallucinated text.
- MACI excels in both coverage (validity) and retention ratio (efficiency).

BCI

(Mohri and Hashimoto et al.)

Amitriptyline and alprazolam are both medications, [T] but they are used to treat different conditions [T] and have different potential side effects [T]. Amitriptyline and alprazolam have different mechanisms of action. [T] Amitriptyline is a tricyclic antidepressant [T] used to treat depression [T] and certain types of chronic pain, [T] ~~but it is not used for anxiety disorders.~~ [F] Alprazolam, on the other hand, is a benzodiazepine medication [T] primarily used to treat anxiety disorders [T] and panic attacks. [T] It works by enhancing the effects of a neurotransmitter in the brain [T] called GABA, [T] which helps to reduce anxiety [T] and promote relaxation. [T] Both medications can cause side effects [T] including drowsiness [T] and dizziness, [T] ~~though confusion is not a typical side effect.~~ [F] However, alprazolam is more likely to cause dependence [T] and withdrawal symptoms, [T] ...

CCI

(Cherian et al.)

Amitriptyline and alprazolam are both medications, [T] but they are used to treat different conditions [T] and have different potential side effects. [T] Amitriptyline and alprazolam have different mechanisms of action. [T] Amitriptyline is a tricyclic antidepressant [T] used to treat depression [T] and certain types of chronic pain, [T] ~~but it is not used for anxiety disorders.~~ [F] Alprazolam, on the other hand, is a benzodiazepine medication [T] primarily used to treat anxiety disorders [T] and panic attacks. [T] It works by enhancing the effects of a neurotransmitter in the brain [T] called GABA, [T] which helps to reduce anxiety [T] and promote relaxation. [T] Both medications can cause side effects [T] including drowsiness [T] and dizziness, [T] ~~though confusion is not a typical side effect.~~ [F] However, alprazolam is more likely to cause dependence [T] and withdrawal symptoms, [T] ...

MACI

(Ours)

Amitriptyline and alprazolam are both medications, [T] but they are used to treat different conditions [T] and have different potential side effects. [T] Amitriptyline and alprazolam have different mechanisms of action. [T] Amitriptyline is a tricyclic antidepressant [T] used to treat depression [T] and certain types of chronic pain, [T] ~~but it is not used for anxiety disorders.~~ [F] Alprazolam, on the other hand, is a benzodiazepine medication [T] primarily used to treat anxiety disorders [T] and panic attacks. [T] It works by enhancing the effects of a neurotransmitter in the brain [T] called GABA, [T] which helps to reduce anxiety [T] and promote relaxation. [T] Both medications can cause side effects [T] including drowsiness [T] and dizziness, [T] ~~though confusion is not a typical side effect.~~ [F] However, alprazolam is more likely to cause dependence [T] and withdrawal symptoms, [T] ...