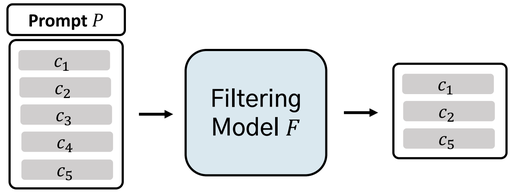
Multi-LLM Adaptive Conformal Inference for Reliable LLM Responses

Kangjun Noh[1], Seongchan Lee[2], Ilmun Kim[2], Kyungwoo Song[1] [1] Yonsei University [2] KAIST

ICLR International Conference On Learning Representations · 연세대학교 YONSEI UNIVERSITY · KAIST

1. Problem Setting

Filtering false claims to gaurantee the factuality of LLM Response



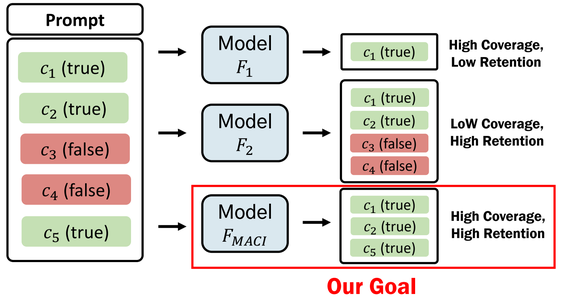
Claim set C of LLM Response

Input: P, C **Output:** $F(P, C)$

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

2. Our Goal

High Coverage, High Retention

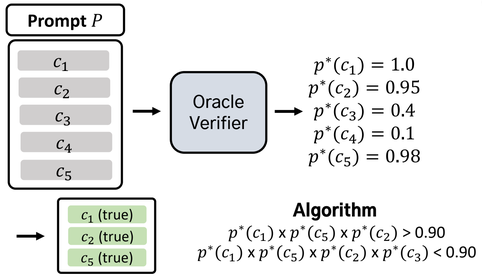


Our Goal

3. Our Motivation

The Oracle Verifier

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



Algorithm

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

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] ...

Target Coverage: High
Retention: Low

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] ...

Target Coverage: Not Enough
Retention: High

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] ...

Target Coverage: High
Retention: High

4. Our Method: *Multi-LLM Adaptive Conformal Inference*

Step 1: Mimicing the Oracle verifier by Multi-LLM Ensemble

Multi-LLM Verifier

"Given the prompt {P} and the set of claims {C}, what is the probability that c_i is true?"

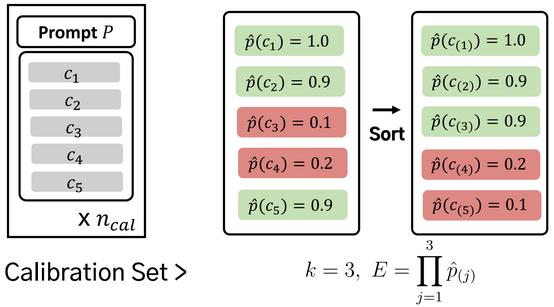
$$\hat{p}(c) := \sum_{i=1}^{m} w_m \mathbb{P}_i(c \text{ is true} \mid P, C), \sum w = 1$$

Step 2: Adaptive Conformal Inference

2-1: Define and Get the Comformity score E

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

Algorithm is same with the **Oracle verifier**!
(k is the largest index of true-claims. please see the example below!)



2-2: Get the Quantile and Coverage

Calculate E for each calibration samples, and get the $1-\alpha$ quantile

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

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

and conduct filtering with the rule below

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

with Exchangeability assumption,

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

Group	Target Coverage: 80% ($\alpha = 0.2$) BCI Cov.	BCI Ret.	CCI Cov.	CCI Ret.	MACI Cov.	MACI Ret.	Target Coverage: 90% ($\alpha = 0.1$) BCI Cov.	BCI Ret.	CCI Cov.	CCI Ret.	MACI Cov.	MACI Ret.	Target Coverage: 95% ($\alpha = 0.05$) BCI Cov.	BCI Ret.	CCI Cov.	CCI Ret.	MACI Cov.	MACI 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**

Paper QR

Email: shwndnjs58@yonsei.ac.kr
Code: https://github.com/MLAI-Yonsei/MACI
MLAI webpage: https://mlai.yonsei.ac.kr/
About Kangjun Noh: https://naneunno-kangjun.github.io/