MULTI-LLM ADAPTIVE CONFORMAL INFERENCE FOR RELIABLE LLM RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Ensuring factuality is essential for the safe use of Large Language Models (LLMs) in high-stakes domains such as medicine and law. Conformal inference provides distribution-free guarantees, but existing approaches are either overly conservative, discarding many true-claims, or rely on adaptive error rates and simple linear models that fail to capture complex group structures. To address these challenges, we reformulate conformal inference in a multiplicative filtering setting, modeling factuality as a product of claim-level scores. Our method, Multi-LLM Adaptive Conformal Inference (**MACI**), leverages ensembles to produce more accurate factuality scores, which in our experiments led to higher retention, while validity is preserved through group-conditional calibration. Experiments show that MACI consistently achieves user-specified coverage with substantially higher retention and lower time cost than baselines. Our anonymized repository is available at https://github.com/Anonymous2026conf/MACI.git.

1 INTRODUCTION

As the performance of Large Language Models (LLMs) continues to advance, attempts to directly utilize their responses in high-stakes domains such as medicine and law are increasing. However, studies continue to report that LLM responses may contain false information (Wang et al., 2024). Therefore, to use LLMs reliably in these critical fields, guaranteeing the factuality of their responses has emerged as an important challenge.

Various methods have been proposed to guarantee the factuality of LLMs, but some are difficult to apply to black-box models (Meng et al., 2023; Quevedo et al., 2024; Chen et al., 2024a) or require access to large external databases or online databases (Chen et al., 2024b; Lee & Yu, 2025). Sampling-based methods (Manakul et al., 2023; Sawczyn et al., 2025) are relatively free from the constraints, but the process of repeatedly checking for response consistency incurs considerable time and financial costs, and they face difficulties in rigorously providing a statistical guarantee at a user-specified error rate.

Recently, studies applying Conformal Inference (CI) (Papadopoulos et al., 2002; Vovk et al., 2005; Lei et al., 2017; Angelopoulos & Bates, 2022) to guarantee the factuality of LLMs have been proposed. For instance, Mohri & Hashimoto (2024) apply the concept of CI to the existing framework of decomposing LLM responses into independent claims and assigning a factuality-score to each one. Their method proposes filtering out claims that do not pass a predetermined threshold. However, because this single, global threshold is applied uniformly to all data, it provides only marginal coverage and can be overly conservative, resulting in the removal of a lot of true information. To improve upon this, Cherian et al. (2024) introduce conditional conformal inference. Instead of a single static threshold for all data, this method employs a threshold function that allows it to change based on the characteristics of a given sample. But it relies on adaptive error rates that are unsuitable for high-stakes applications requiring a fixed guarantee. Its threshold function also struggles to capture the characteristics of the complex grouping criteria of LLM responses that are separated by their semantic properties.

In this context, we propose a new methodology called Multi-LLM Adaptive Conformal Inference (MACI). The core objective of MACI is to preserve as much factual information as possible while strictly adhering to a high user-specified error rate. To achieve this, we theoretically demonstrate that, assuming an ideal Oracle factuality score, the optimal false-claim filtering method is expressed

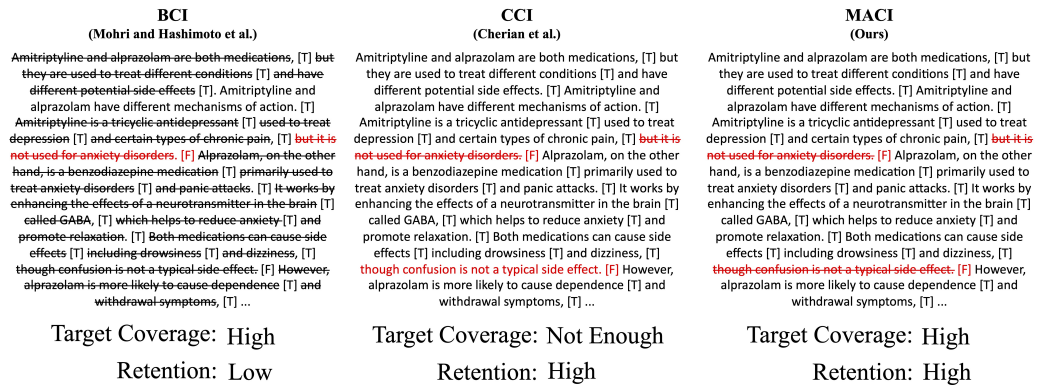


Figure 1: Comparison of Conformal Inference Methods. T (true) and F (false) denote ground-truth labels per claim. Basic Conformal Inference (Mohri & Hashimoto, 2024) attains coverage by aggressive filtering, yielding low retention. Conditional Conformal Inference (Cherian et al., 2024) proposes adaptive thresholds but relaxes guarantees; MACI achieves both high coverage and retention.

as a cumulative probability product. Inspired by this finding, we design a new adaptive CI framework that uses a conformity score in the form of a cumulative probability product. We further theoretically prove that the quality of the factuality-score directly impacts the retention ratio. Accordingly, we adapt our strategy to maximize factuality-score quality through a multi-LLM ensemble. As a result, MACI not only theoretically guarantees group-conditional coverage but also empirically demonstrates robust group-conditional coverage across diverse datasets, all while showing a substantially higher retention ratio than existing methodologies.

Our main contributions are:

- We introduce a multiplicative filtering framework that models factuality as the product of claim-level scores (factuality-score) while preserving finite-sample guarantees.
- We provide the first retention theoretical analysis in conformal inference to our knowledge, linking oracle–estimator deviations to true-claim preservation and motivating ensemble design.
- We extend conformal inference with group-conditional calibration and a multi-LLM ensemble, ensuring group-conditional coverage and showing substantially higher retention than conformal baselines in high-stakes domains.

2 Related Works

2.1 Basic Conformal Inference

Mohri & Hashimoto (2024) generalize CI methods to guarantee the factuality of LLM responses, providing distribution-free, model-free guarantees above user-specified error rates α. Their approach, called Basic Conformal Inference (BCI), applies calibration procedures to decomposed claims, defines factuality-scores, and filters claims likely to be false using thresholds learned from held-out sets. BCI marginally guarantees no false-claims in test samples, but because it only provides marginal coverage, it can under- or over-cover specific subgroups. Moreover, since LLM responses are mostly factual with only slight false information, high target coverage yields conservative thresholds that remove many true-claims.

2.2 Conditional Conformal Inference

Cherian et al. (2024) propose conditional conformal inference (CCI) to address BCI's limits. Extending conditional conformal methods (Gibbs et al., 2024), CCI trains functions that output sample-wise thresholds from calibration sets, ensuring group-conditional rather than marginal coverage. Groups can be defined by prompt or response characteristics, such as the length of the prompt. Beyond

this, CCI preserves more claims through adaptive error rates and conditional boosting, though the practicality of adaptive α in high-stakes settings and the linearity of the threshold function remain open challenges.

3 Background and Preliminaries

Document structure and factuality-scores. Let $\mathcal{P}$ denote the space of prompts and $\mathcal{C}$ the space of claims. Each document $D = (P, C, Y)$ consists of a prompt $P \in \mathcal{P}$, a set of claims $C = \{c_1, \ldots, c_{|C|}\} \subseteq \mathcal{C}$, and labels $Y \in \{0,1\}^{|C|}$ indicating which claims are factual. We assume documents are drawn i.i.d. from a distribution $\mathbf{P}$, which implies exchangeability of calibration and test data. A *factuality-score function* $p : \mathcal{P} \times \mathcal{C} \to [0,1]$ assigns each (P, c) the probability of being factual, with oracle p^* and estimator $\hat{p}$.

Filtering operator. Given a score function p, a threshold $\tau \in [0,1]$, and optional randomization U, the filtering operator $F(p, \tau, U; P, C) \subseteq C$ returns the claims retained under τ. Calibrating τ on held-out data yields the conformal inference rule $F_{n,\alpha}$, the central object of our analysis.

Group-Conditional coverage. Exact instance-level coverage is infeasible in a distribution-free setting (Vovk, 2012; Barber et al., 2020). Instead, we require validity within subgroups, reflecting meaningful categories such as domains, topics, or user populations. Formally, a grouping function $g : \mathcal{P} \times \mathcal{C} \to \{1, \ldots, K\}$ assigns each (P_i, C_i) to one of K groups, and we demand

$$\mathbb{P}\big(\forall c_{n+1,j} \in F_{n,\alpha}(P_{n+1}, C_{n+1}),\ y_{n+1,j} = 1 \mid g(P_{n+1}, C_{n+1}) = k\big) \geq 1 - \alpha, \tag{1}$$

for all $k \in \{1, \ldots, K\}$. This mirrors Mondrian conformal prediction (Vovk et al., 2005) but applies to prompt–claim pairs. In experiments, we instantiate g with high-level dataset-specific categories (e.g., medical question types, entity groups).

4 MACI: Multi-LLM Adaptive Conformal Inference

Building on Section 3, our goal is to design a filtering rule F that satisfies the group-conditional coverage guarantee (1). The baseline BCI method applies one global threshold, which is simple but ignores group heterogeneity and relies on a single predictor.

Following adaptive conformal inference (Romano et al., 2020), we propose *MACI* (Multi-LLM Adaptive Conformal Inference). MACI aggregates scores from multiple LLMs and calibrates group-conditional thresholds, improving retention while preserving coverage guarantees.

4.1 Oracle Factuality

Building on the definition of a factuality-score function in Section 3, we first consider an idealized regime where the factuality-score coincides with the true conditional probability. In this oracle setting, the model has complete distributional knowledge of claim correctness, so that for any prompt–claim pair (P, c) it can evaluate $\mathbb{P}(y = 1 \mid P, c)$ exactly.

Definition 1 (Oracle factuality-score)**.** For any prompt–claim pair (P, c) with binary factuality label $y \in \{0,1\}$, the oracle factuality-score is defined as $p^*(P, c) := \mathbb{P}(y = 1 \mid P, c)$. For a document $D_i = (P_i, C_i, Y_i)$ and $c_{i,j} \in C_i$, we denote

$$p_i^*(c_{i,j}) := p^*(P_i, c_{i,j}) = \mathbb{P}(y_{i,j} = 1 \mid P_i, c_{i,j}).$$

We assume conditional independence: given (P_i, C_i), the labels $y_{i,j}$ are independent Bernoulli variables with success probabilities $p_i^*(c_{i,j})$. This allows the joint distribution to decompose into marginals, simplifying the analysis of validity. While this assumption may be unrealistic for real LLM outputs, it provides a clean baseline and highlights the role of marginal conditional probabilities in our framework.

Given a document $D_i = (P_i, C_i, Y_i)$ with $n_i = |C_i|$, we use the shorthand $[n] := \{1, \ldots, n\}$ for a positive integer n. Let $\pi_i : [n_i] \to [n_i]$ be a permutation ordering claims by decreasing oracle

scores, $p_i^*(c_{i,\pi_i(1)}) \geq \cdots \geq p_i^*(c_{i,\pi_i(n_i)})$ (ties broken arbitrarily). Define $P_k := \prod_{j=1}^{k} p_i^*(c_{i,\pi_i(j)})$ with the convention $P_0 = 1$ and $P_{n_i+1} = 0$. For a threshold $\tau \in [0,1]$, define the cutoff index and filtered set

$$K_i^*(\tau) := \max\left\{k \in [n_i] : \prod_{j=1}^{k} p_i^*(c_{i,\pi_i(j)}) \geq \tau\right\}, \quad F_\tau^*(P_i, C_i) := \{c_{i,\pi_i(j)} : j \leq K_i^*(\tau)\}.$$

with the convention $\max \emptyset = 0$. Thus F_τ^* ensures coverage $\geq \tau$, is monotone in τ ($\tau_1 \leq \tau_2 \Rightarrow F_{\tau_2}^* \subseteq F_{\tau_1}^*$), But it is conservative since coverage typically exceeds τ. To obtain *exact* coverage, we randomize at the boundary index $K_i^*(\tau)$. Let $P_k = \prod_{j=1}^{k} p_i^*(c_{i,\pi_i(j)})$ and define $\gamma_i(\tau) = \frac{P_{K_i^*(\tau)} - \tau}{P_{K_i^*(\tau)} - P_{K_i^*(\tau)+1}} \in [0,1]$ (with $\gamma_i(\tau) = 0$ if the denominator vanishes). With $U \sim \mathrm{Unif}(0,1)$, the randomized oracle rule is

$$F_\tau^{\mathrm{oracle}}(P_i, C_i) = \begin{cases} \{c_{i,\pi_i(j)} : j \leq K_i^*(\tau)\}, & U > \gamma_i(\tau), \\ \{c_{i,\pi_i(j)} : j \leq K_i^*(\tau) + 1\}, & U \leq \gamma_i(\tau). \end{cases}$$

This randomization balances inclusion and exclusion at the boundary, achieving exact coverage at level τ while maximizing expected retention.

4.2 Adaptive Conformal Inference for false-claim filtering

The oracle procedure in Section 4.1 requires access to the true factuality-score p^* and thus serves as a theoretical benchmark for optimal filtering. In practice, however, p^* is unknown and must be replaced with an estimated score $\hat{p}$, motivating our adaptive conformal inference (ACI) procedure, which mirrors the oracle rule while relying only on estimated scores and preserves coverage guarantees. Concretely, let a black-box classifier (e.g., an LLM) produce estimates $\hat{p}_i(c_j)$ of claim factuality given (P_i, C_i), with any probabilistic classifier applicable. We then replace p^* with $\hat{p}$ and calibrate τ using conformal quantiles on held-out data; with pre-trained LLMs, all available data may be used, but the principle remains the same.

Inspired by the inverse-quantile scores of Romano et al. (2020), we design a conformity score tailored to false-claim filtering. The goal is a score uniformly distributed under the oracle p^*, so that calibration remains valid when p^* is replaced by $\hat{p}$. For each (P_i, C_i, Y_i), let $A_i = c_{i,j} \in C_i : y_{i,j} = 1$ be the set of true-claims and $U_i \sim \mathrm{Unif}(0,1)$. We define $E_i = \inf\left\{\tau \in [0,1] : F(\hat{p}, \tau, U_i; P_i, C_i) \subseteq A_i\right\}$, the smallest threshold at which all retained claims are correct.

Each conformity score E_i reduces the filtering event to a one-dimensional statistic, directly suitable for conformal quantile calibration (Lemma 1 in Appendix A). Applying the standard quantile argument then yields the following guarantee.

Theorem 1 (Marginal coverage guarantee)**.** *If the samples (P_i, C_i, Y_i), for $i \in \{1, \ldots, n+1\}$, are exchangeable, the adaptive conformal inference rule (Algorithm 1) satisfies*

$$\mathbb{P}\big(\forall c_{n+1,j} \in F_{n,\alpha}(P_{n+1}, C_{n+1}),\ y_{n+1,j} = 1\big) \geq 1 - \alpha.$$

Furthermore, if the scores E_i are almost surely distinct, the marginal coverage is nearly tight:

$$\mathbb{P}\big(\forall c_{n+1,j} \in F_{n,\alpha}(P_{n+1}, C_{n+1}),\ y_{n+1,j} = 1\big) \leq 1 - \alpha + \frac{1}{n+1}.$$

Marginal validity ensures that the overall error rate is controlled on average, but this may hide differences between groups, with some subpopulations receiving weaker guarantees. To address this, we extend adaptive conformal inference to a group-conditional setting, so that validity is enforced separately for each group.

This extension follows the Mondrian conformal framework of Vovk et al. (2005). Instead of pooling all scores, calibration is restricted to examples from the same group as the test instance. For group k with calibration set $\mathcal{I}_k = \{i : g(P_i, C_i) = k\}$, the threshold is $\hat{Q}_{1-\alpha}^{(k)} = \mathrm{Quantile}(\{E_i : i \in \mathcal{I}_k\}, 1-\alpha)$. Given a test instance in group k, the filter is $\hat{F}_{n,\alpha}^{(k)}(P_{n+1}, C_{n+1}) = F(\hat{p}, \hat{Q}_{1-\alpha}^{(k)}, U_{n+1}; P_{n+1}, C_{n+1})$.

Theorem 2 (Group-conditional coverage guarantee)**.** *If the samples* $\{(P_i, C_i, Y_i)\}_{i=1}^{n+1}$ *are exchangeable, the group-conditional conformal inference rule satisfies*

$$\mathbb{P}\big(\forall c_{n+1,j} \in F^{(k)}_{n,\alpha}(P_{n+1}, C_{n+1}),\ y_{n+1,j} = 1 \,|\, g(P_{n+1}, C_{n+1}) = k\big) \geq 1 - \alpha,$$

for all $k \in \{1, \ldots, K\}$ *with* $\mathbb{P}(g(P_{n+1}, C_{n+1}) = k) > 0$.

A key implication of Theorem 2 is that it ensures *finite-sample, distribution-free validity within each group*, in contrast to the marginal guarantee of Theorem 1, which holds only in aggregate. Each group achieves a level $1 - \alpha$ based on its own calibration size n_k, ensuring that even small groups are covered, albeit with more conservative thresholds and reduced retention.

In the oracle regime ($\hat{p} = p^*$), conformity scores are exactly uniform on $[0, 1]$ (Lemma 2 in Appendix A). This uniformity implies that Theorem 2 achieves maximal retention efficiency: coverage is attained precisely at the target level, without conservatism, so no true-claims are unnecessarily discarded. To formalize this notion of efficiency, we introduce the *retention ratio*, which measures the proportion of claims that are retained under a given factuality-score function and threshold.

Formally, let $(P, C, Y) \sim \mathbf{P}$ be a random document (cf. Section 3), and let $\rho := \mathbb{P}(y = 1)$ be the marginal probability that a claim is true. For a factuality-score function p and threshold τ, define the retention rate as

$$R(p, \tau) := \mathbb{P}\big(c \in F_\tau(p; P, C)\big). \tag{2}$$

Theorem 3 (Retention gap with MSE)**.** *Let* p^* *denote the oracle factuality-score and* $\hat{p}$ *an estimated score. Fix a threshold* $\tau \in [0, 1]$ *and let*

$$\Delta := \big|R(\hat{p}, \tau) - R(p^*, \tau)\big|.$$

Suppose (i) $\mathbb{E}\big[(\hat{p} - p^*)^2\big] < \infty$ *and (ii) the oracle factuality-scores are not overly concentrated near the threshold* τ*, in the sense that* $\mathbb{P}\big(|p^*(P, c) - \tau| \leq \epsilon\big) \leq C\epsilon^\beta$ *for some* (C, β)*. Then*

$$\Delta \ \leq\ \frac{\mathbb{E}[(\hat{p} - p^*)^2]}{\epsilon^2} + C\epsilon^\beta,$$

and optimizing in ϵ *yields*

$$\Delta \leq C' \left(\mathbb{E}[(\hat{p} - p^*)^2]\right)^{\frac{\beta}{\beta+2}},$$

where C' *depends only on* (C, β).

Assumption (i) requires that $\hat{p}$ not deviate too far from the oracle p^* on average, keeping errors manageable. Assumption (ii) requires that oracle scores not cluster near the threshold τ, so small mistakes in $\hat{p}$ rarely alter retention. Under these conditions, Theorem 3 shows that the retention gap decreases polynomially with the MSE of $\hat{p}$, highlighting that improved accuracy directly yields more efficient retention.

4.3 Multi-LLM Ensemble

From a statistical perspective, ensembling multiple predictors reduces variance in the bias–variance tradeoff and lowers MSE, bringing the estimator closer to the oracle benchmark. Yet directly minimizing MSE is not practical because the oracle score is unobservable and binary labels drive predictors toward overconfident outputs, which makes ensembles prone to overfitting.

We therefore use a surrogate objective based on the retention decomposition. By keeping recall above a tolerance and reducing the FPR, we directly improve retention while avoiding overconfidence. This surrogate remains aligned with the oracle goal and, as our figure 3 shows, also reduces MSE in practice.

Recall the retention rate defined in (2), which can be decomposed as follows.

$$R(p, \tau) = \rho \cdot \mathrm{TPR}(p, \tau) + (1 - \rho) \cdot \mathrm{FPR}(p, \tau), \tag{3}$$

where

$$\mathrm{TPR}(p, \tau) = \frac{\mathbb{P}(c \in F_\tau(p; P, C), y = 1)}{\rho}, \quad \mathrm{FPR}(p, \tau) = \frac{\mathbb{P}(c \in F_\tau(p; P, C), y = 0)}{1 - \rho}.$$

Maximizing $R(p, \tau)$ therefore amounts to increasing TPR while decreasing FPR, but the two cannot be optimized simultaneously. To prevent trivial solutions that sacrifice recall, we require the true positive rate to remain above a fixed tolerance, $\text{TPR}(p, \tau) \geq 1 - \delta$ for $\delta \in (0, 1)$. With $\tau_{p,\delta}$ denoting the δ-quantile of conformity scores among true-claims, we thus focus on minimizing the FPR subject to this constraint:

$$p^\star = \arg\min_p \ \mathbb{E}\big[\text{FPR}(p, \tau_{p,\delta})\big]. \tag{4}$$

Since direct fine-tuning toward the oracle p^* is impractical for black-box LLMs, we instead target the surrogate optimum $p^\star$ in (4) by adopting a multi-LLM ensemble strategy. Given base factuality-scorers $\{p_m\}_{m=1}^M$ and weights $w = (w_1, \ldots, w_M)$, the ensemble predictor is

$$p_{\text{ens}}(P, c; w) = \sum_{m=1}^{M} w_m \, p_m(P, c),$$

with w optimized to minimize the empirical FPR under the tolerance constraint (see Appendix B.1 for details). Algorithm 2 summarizes the MACI framework, which combines group-conditional conformal inference with the ensemble to maximize retention while preserving exact coverage.

5 Empirical Results

We empirically validate the superiority of MACI by using three datasets with distinct characteristics. For these datasets, we define a representative grouping criteria for each and a general group criterion of false-claim risk that is commonly applied to the false-claim filtering task. For numerical stability, all conformity computations and thresholding are performed on the transformed $-\ln(1 - \hat{p}(c) + \epsilon)$, where ϵ is a small positive constant. This transform is strictly monotone, so conformal quantiles, selection sets, and coverage guarantees are unchanged, while computations involving probabilities are stabilized by operating in log-space. A detailed explanation of the datasets, group criteria, selecting LLMs, and evaluation metrics can be found in Appendix D.

5.1 Overall Performance

Table 1 compares the group-conditional coverage and retention ratio for three datasets with distinct characteristics against two prominent baselines in the false-claim filtering field: BCI and CCI. MACI demonstrates robust performance in settings where the other two baselines falter, consistently achieving the target coverage across most groups while maintaining the highest retention ratio.

Comparison with BCI. BCI only guarantees marginal coverage via a single threshold, which leads to alternating overcoverage and undercoverage depending on the difficulty differences between groups. This phenomenon is particularly evident in the results for the False-Claim Risk groups across all three datasets and the View Count groups in WikiBio. Furthermore, the task of false-claim filtering typically involves a high proportion of true-claims interspersed with a few false-claims. Consequently, guaranteeing high coverage (e.g., 90%) with BCI requires setting an overly conservative single threshold, which in turn filters out most claims and causes a sharp decline in the retention ratio. By utilizing group-conditional thresholds and a cumulative conformity score, MACI simultaneously achieves stable group-conditional coverage and a high retention ratio.

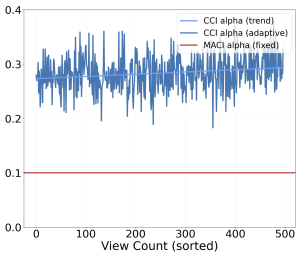
Comparison with CCI. CCI shows an improved retention ratio over BCI, presenting a more advanced result than BCI's overly conservative outputs. The function g_{CCI} (Theorem 3.1. of Cherian et al. (2024)), which operates within a linear feature space framework, presents certain constraints when applied to our grouping scenarios. The grouping criteria for each dataset, such as Medical Content or False-Claim Risk, are complex semantic functions implemented based on prompt and claim parsing. It is therefore difficult to capture such criteria using the simple linear functions and features proposed by CCI, leading to undercoverage or overcoverage. In contrast to g_{CCI}, our grouping function g is an arbitrary measurable function that partitions the space into a finite number of groups, therefore unaffected by the complexity of grouping criteria in threshold calculations. The constraints inherent in g_{CCI} are also reflected in its retention ratio. g_{CCI} risks calculating an overly conservative threshold depending on how well it captures the grouping criteria. This leads to a lower retention ratio compared to MACI, which calculates group-conditional thresholds directly.

Table 1: Group-conditional coverage, marginal coverage, and retention ratio for three different datasets with distinct characteristics. The marginal results are in the row corresponding to the dataset name, followed by two rows showing the results for two representative grouping criteria for that dataset. Coverage within $1-\alpha \pm 0.01$ are marked with a green dot •, while values that fall outside this range, within $1-\alpha \pm 0.02$ (indicating either over or undercoverage), are marked with a red arrow ↓↑. Compared to the two conformal inference baselines, MACI consistently achieves the target coverage in most cases, regardless of the group. Furthermore, its retention ratio is the highest across almost all groups. **Cov.** denotes coverage, **Ret.** denotes the retention ratio. The result with the highest retention ratio, achieved without under-coverage, is marked in **bold**. All reported values are the mean over 30 repeated trials. The performance of CCI is a result of fixing the target coverage (1-α).

	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	
Group	**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**
ExpertQA	0.91↑	0.13	0.85↑	0.18	0.80•	**0.45**	0.91•	0.13	0.85↓	0.17	0.90•	**0.15**	0.91↓	0.13	0.85↓	0.17	0.95•	**0.10**
Question Domain																		
Bio/Med	0.92↑	0.14	0.86↑	0.18	0.82↑	**0.47**	0.92↑	0.14	0.86↓	0.18	0.92↑	**0.22**	0.92↓	0.13	0.86↓	0.18	0.97↑	**0.10**
Tech/Sci	0.91↑	0.14	0.86↑	0.17	0.81•	**0.44**	0.90•	0.13	0.85↓	0.16	0.89•	**0.21**	0.91↓	0.13	0.85↓	0.16	0.94•	**0.10**
Common	0.90↑	0.13	0.84↑	0.18	0.78↓	0.43	0.89•	0.13	0.85↓	0.17	0.89•	**0.21**	0.89↓	0.14	0.84↓	0.17	0.95•	**0.09**
False-Claim Risk																		
Low	0.95↑	0.13	0.85↑	0.31	0.81•	**0.57**	0.94↑	0.13	0.84↓	0.31	0.89•	**0.35**	0.95•	0.13	0.84↓	0.31	0.96•	**0.16**
Medium	0.91↑	0.13	0.87↑	0.18	0.81•	**0.42**	0.91•	0.13	0.86↓	0.18	0.89•	**0.23**	0.91↓	0.13	0.86↓	0.18	0.96•	**0.11**
High	0.87↑	0.13	0.85↑	0.12	0.79•	**0.37**	0.87↓	0.13	0.85↓	0.12	0.90•	**0.15**	0.87↓	0.13	0.85↓	0.12	0.95•	**0.07**

CCI proposes improving retention by applying per-sample adaptive error rates that reflect each sample's characteristics. The method learns α as a function and lowers α for each sample instead of merely exceeding a minimum retention target. However, this adaptive α differs from our objective. Our goal is to design filtering rules that guarantee, with high probability, that the filtered set contains no false-claims, ensuring applicability in real high-stakes domains. Adapting α to raise retention produces filtering rules that are difficult to deploy in such settings. Figure 2 compares CCI with adaptive α and MACI with α = 0.1 on WikiBio. The upper plot sets CCI's target retention to MACI's average retention and outputs a per-sample adaptive α, showing that CCI's α values are generally higher than MACI's fixed small α. The lower table reports actual coverage and retention for both methods. CCI raises retention to nearly match MACI by increasing α overall, but the actual coverage is lower than MACI because the target α is larger. This confirms MACI's superiority in the coverage–retention trade-off.

Comparison with sampling-based methods. While our primary focus is on CI-based baselines, it is also practical to compare against recent non-CI approaches. We compare MACI's group-conditional coverage, marginal coverage, and retention ratio with sampling-based methods that apply to black-box LLMs and do not rely on retrieval. Brief descriptions of these baselines appear



	CCI (α=adap.)		MACI (α=0.1)	
Group	**Cov.**	**Ret.**	**Cov.**	**Ret.**
WikiBio	0.72	0.26	0.90•	**0.28**
View Count				
Low	0.71	0.24	0.91•	**0.29**
Medium	0.73	0.26	0.90•	**0.27**
High	0.74	0.28	0.90•	**0.31**

Figure 2: Performance comparison of CCI (adaptive α) and MACI measured by View Count on the WikiBio dataset. The horizontal axis of the left graph is the sample index sorted by View Count, and the vertical axis is α. The left graph shows the variation in α when CCI (adaptive α) sets its target retention ratio to MACI's average retention ratio. CCI (adaptive α) trades off higher α to achieve a higher retention ratio, and the table below shows the resulting decrease in coverage.

in Section D.4. Unlike CI-based methods, sampling-based approaches do not provide statistical guarantees; instead, they compute a factuality-score $p \in [0, 1]$, enabling false-claim filtering via a threshold (e.g., 0.5). Table 2 shows that sampling-based methods attain high retention but low coverage. This highlights their limitation in meeting the strict requirement that the filtered set contain no false-claims. Moreover, their target coverage is not user-specified and thus unpredictable. These points underscore the need for MACI in high-stakes settings that require a user-specified high $1 - \alpha$.

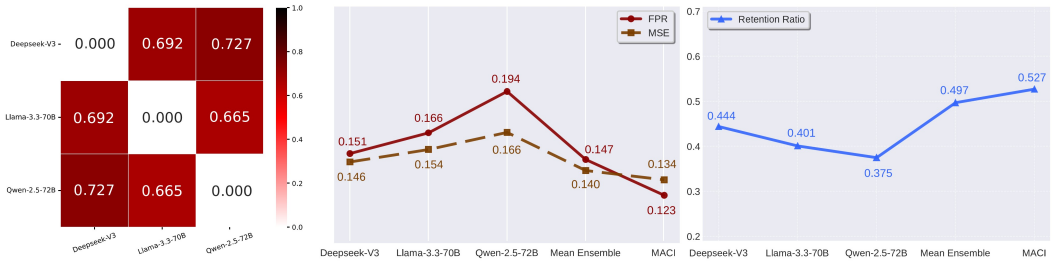
Table 2: Comparison with sampling-based methods, a representative black-box and non-retrieval approach for false-claim filtering. Sampling-based methods generally exhibit very low or unstable coverage and a high retention rate. This suggests they are unsuitable for the strict target that all claims have to be factual (the definition of **Cov.**). In contrast, MACI (α = 0.1) demonstrates the ability to reliably guarantee the user's desired coverage.

	SelfCheck		FSC-text		FSC-KG		MACI	
Group	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**
MedLFQA	0.56	0.97	0.63	0.85	0.64	0.88	**0.90**	0.50
Medical Content								
Info	0.49	0.98	0.59	0.85	0.58	0.87	**0.90**	0.48
Interpret	0.54	0.98	0.59	0.90	0.63	0.93	**0.90**	0.47
Action	0.64	0.95	0.74	0.79	0.73	0.83	**0.90**	0.53
False-Claim Risk								
Low	0.64	0.98	0.70	0.86	0.71	0.89	**0.89**	0.52
Medium	0.56	0.98	0.63	0.88	0.60	0.92	**0.91**	0.46
High	0.41	0.97	0.55	0.82	0.58	0.85	**0.89**	0.41

	SelfCheck		FSC-text		FSC-KG		MACI	
Group	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**
WikiBio	0.12	0.97	0.33	0.77	0.37	0.70	**0.90**	0.25
View Count								
Low	0.11	0.95	0.42	0.69	0.46	0.64	**0.91**	0.21
Medium	0.13	0.97	0.31	0.79	0.27	0.73	**0.91**	0.24
High	0.11	0.98	0.26	0.82	0.39	0.73	**0.91**	0.24
False-Claim Risk								
Low	0.19	0.98	0.41	0.78	0.43	0.72	**0.90**	0.23
Medium	0.08	0.96	0.28	0.74	0.32	0.68	**0.90**	0.25
High	0.08	0.97	0.30	0.78	0.35	0.70	**0.90**	0.28

5.2 Multi-LLM Ensemble

In Sections 4.1, 4.2, and 4.3, we have discussed the importance of the factuality-score $\hat{p}$ and its optimization via multi-LLM ensemble. Consequently, we first seek to verify to what extent our proposed multi-LLM ensemble and optimization method (Section 4.3) improve the performance of $\hat{p}$ compared to a single-LLM, and how the retention ratio is correspondingly improved. We first find that models exhibit significant disagreements in false-claim detection. Figure 3 (a) shows the high Jaccard distance (Jaccard, 1901) between the sets of claims that different LLMs classify as false. The analysis is performed exclusively on the subset of claims from MedLFQA where the ground truth is false. This high distance implies that the models have different patterns for detecting false-claims, suggesting significant potential for performance enhancement through an ensemble. Figure 3 (b) shows that the FPR and MSE are sequentially improved from the single-LLM to the arithmetic mean ensemble, and finally to MACI. Figure 3 (b) also shows that an improvement in FPR is consistently accompanied by an improvement in MSE, and further demonstrates that MACI's $\hat{p}$ is a superior estimator of factuality-score. Figure 3 (c) demonstrates that the corresponding sequential increase in the retention ratio aligns with our objective of maximizing it by enhancing the quality of $\hat{p}$.



(a) Jacard distance between three LLMs

(b) Marginal FPR and MSE

(c) Marginal Retention Ratio

Figure 3: (a) shows the high Jaccard distance between different LLMs' predictions on claims known to be false in MedLFQA, indicating diverse false-claim detection patterns that support using an ensemble. (b) demonstrates the sequential improvement in FPR from a single-LLM and a simple arithmetic mean ensemble to our proposed MACI. It also demonstrates that as the FPR improves (0.147 to 0.123), the MSE also improves in practice (0.140 to 0.134); (c) demonstrates that as the FPR and MSE improve, the retention ratio also increases. (0.497 to 0.527)

5.3 Time Cost

Time cost is a critical factor for real-time LLM response filtering. Our analysis compares MACI with existing methods across two phases: Factuality-Score Generation, where factuality-scores are created for LLM claims, and Calibration, where thresholds and ensemble weights are optimized; the rapid filtering step is excluded. The factuality-scores of sampling-based methods are generated with the Llama-3.3-70B-Instruct. The factuality-score generation time for MACI is the average time of the three models in D.3. Table 3 reports costs on the WikiBio dataset for each phase. Sampling-based methods are slower in score generation since they filter simultaneously and lack a distinct calibration phase; for instance, SelfCheck must generate new responses, and FSC-KG takes over ten times longer than MACI due to knowledge-graph construction and entity extraction. Among conformal methods, CCI inherits SelfCheck's generation times. In calibration, MACI is faster than CCI because it uses simpler ensemble weight optimization instead of adaptive α search and conditional boosting, and its single parallel scoring pass also makes it faster than sampling-based approaches.

Table 3: Time-Cost Comparison of Four Filtering Methods. For sampling-based methods, the factuality-score can be used for filtering at the time of generation, so the time for calibration and filtering is excluded. The factuality-score for CCI is based on the results of SelfCheck, so the same value is listed. The time required for inference can be approximately calculated as: calibration phase time + (factuality-score generation time $\times$ # test samples).

Phase	SelfCheck	FSC-KG	CCI	MACI
Factuality-Score (s)	3.25 ± 0.43	19.30 ± 2.81	3.25 ± 0.43	**1.20 ± 0.13**
Calibration (s)	—	—	10.33 ± 1.18	**3.24 ± 0.65**

6 Conclusions

We reformulate conformal inference through a multiplicative filtering structure, providing a framework for false-claim detection with finite-sample, distribution-free guarantees. Our analysis reveals how deviations from the oracle factuality-score impact retention, motivating the use of ensemble methods to narrow this gap. Building on these insights, we develop MACI, which uses ensemble-based factuality-scores and group-conditional calibration to provide group-conditional coverage guarantees. Experiments demonstrate that MACI achieves user-specified coverage while substantially improving factual claim retention and running more efficiently than existing methods, offering a practical solution for deploying LLMs in high-stakes applications.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL https://arxiv.org/abs/2107.07511.

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction, 2025. URL https://arxiv.org/abs/2411.11824.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 08 2020. ISSN 2049-8772. doi: 10.1093/imaiai/iaaa017. URL https://doi.org/10.1093/imaiai/iaaa017.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=Zj12nzlQbz.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. Complex claim verification with evidence retrieved in the wild, 2024b. URL https://arxiv.org/abs/2305.11859.

John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. Large language model validity via enhanced conformal prediction methods, 2024. URL https://arxiv.org/abs/2406.09714.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, and Dejian Yang et al. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees, 2024. URL https://arxiv.org/abs/2305.12616.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, and Archi Mitra et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. Language models hallucinate, but may excel at fact verification. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1090–1111, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.62. URL https://aclanthology.org/2024.naacl-long.62/.

Paul Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:241–72, 01 1901. doi: 10.5169/seals-266440.

Minbyul Jeong, Hyeon Hwang, Chanwoong Yoon, Taewhoo Lee, and Jaewoo Kang. Olaph: Improving factuality in biomedical long-form question answering, 2024. URL https://arxiv.org/abs/2405.12701.

DongGeon Lee and Hwanjo Yu. Refind at semeval-2025 task 3: Retrieval-augmented factuality hallucination detection in large language models, 2025. URL https://arxiv.org/abs/2502.13622.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression, 2017. URL https://arxiv.org/abs/1604.04173.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. Expertqa: Expert-curated questions and attributed answers, 2024. URL https://arxiv.org/abs/2309.07852.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023. URL https://arxiv.org/abs/2303.08896.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023. URL https://arxiv.org/abs/2202.05262.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation, 2023. URL https://arxiv.org/abs/2305.14251.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees, 2024. URL https://arxiv.org/abs/2402.10978.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, pp. 345–356. Springer, 2002.

Ernesto Quevedo, Jorge Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. Detecting hallucinations in large language model generation: A token probability approach, 2024. URL https://arxiv.org/abs/2405.19648.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage, 2020. URL https://arxiv.org/abs/2006.02544.

Albert Sawczyn, Jakub Binkowski, Denis Janiak, Bogdan Gabrys, and Tomasz Kajdanowicz. Fact-selfcheck: Fact-level black-box hallucination detection for llms, 2025. URL https://arxiv.org/abs/2503.17229.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback, 2023. URL https://arxiv.org/abs/2305.14975.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In Steven C. H. Hoi and Wray Buntine (eds.), *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR. URL https://proceedings.mlr.press/v25/vovk12.html.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Jyoti Das, and Preslav Nakov. Factuality of large language models: A survey, 2024. URL https://arxiv.org/abs/2402.02420.

APPENDIX

Overview of Appendices Appendix A contains the proofs of the main theoretical results that were omitted from the paper. Appendix B provides additional methodological details, including precise definitions of the ensemble objective and empirical quantities used in our framework. Appendix C reviews background material on conformal inference and its adaptation to false-claim filtering. Appendix D reports implementation details, datasets, and evaluation metrics for our numerical experiments, together with supplementary results. Appendix E reports the use of a Large Language Model for our research.

A PROOFS OF MAIN RESULTS

Lemma 1. *For each $i \in \{1, \dots, n\}$, each threshold $\tau \in [0, 1]$, and each auxiliary randomization $U_i \sim \mathrm{Unif}(0, 1)$, we have*

$$\{E_i \le \tau\} \iff \{F(\hat{p}, \tau, U_i; P_i, C_i) \subseteq A_i\}.$$

Proof. Fix $i \in \{1, \dots, n\}$, a threshold $\tau \in [0, 1]$, and a randomization variable $U_i \sim \mathrm{Unif}(0, 1)$. By the definition of E_i,

$$E_i = \inf \big\{t \in [0, 1] : F(\hat{p}, t, U_i; P_i, C_i) \subseteq A_i\big\}.$$

($\Rightarrow$) Suppose $E_i \le \tau$. Then, by the definition of the infimum, there exists $\tau^* \le \tau$ such that

$$F(\hat{p}, \tau^*, U_i; P_i, C_i) \subseteq A_i.$$

Since the retained set $F(\hat{p}, t, U_i; P_i, C_i)$ is non-increasing in monotone in τ, it follows that

$$F(\hat{p}, \tau, U_i; P_i, C_i) \subseteq A_i.$$

($\Leftarrow$) Conversely, suppose that

$$F(\hat{p}, \tau, U_i; P_i, C_i) \subseteq A_i.$$

Then τ belongs to the set

$$\{\tau \in [0, 1] : F(\hat{p}, \tau, U_i; P_i, C_i) \subseteq A_i\}.$$

Hence, by the definition of the infimum, we obtain $E_i \le \tau$. Combining the two directions establishes the desired equivalence. That is,

$$\{E_i \le \tau\} \iff \{F(\hat{p}, \tau, U_i; P_i, C_i) \subseteq A_i\}.$$

□

A.1 PROOF OF THEOREM 1

The proof follows the standard argument for marginal coverage in conformal prediction and is restated here in our setting.

Lower bound. By Algorithm 1 and Lemma 1, the event that all retained claims are factual can be written as

$$\{\forall c_{n+1,j} \in F_{n,\alpha}(P_{n+1}, C_{n+1}),\ y_{n+1,j} = 1\} \iff \{E_{n+1} \le \hat{Q}_{1-\alpha}(\{E_i\}_{i=1}^n)\}.$$

Since the samples (P_i, C_i, Y_i) are exchangeable and the randomizations U_i are i.i.d., the conformity scores $\{E_1, \dots, E_{n+1}\}$ are themselves exchangeable. Therefore, for any $\tau \in [0, 1]$,

$$\mathbb{P}\big(E_{n+1} \le \mathrm{Quantile}(\{E_i\}_{i=1}^{n+1}; \tau)\big) \ \ge\ \tau.$$

(see Fact 2.15 of Angelopoulos et al. (2025)). Choosing $\tau = 1 - \alpha$ and using the equivalence above, we obtain

$$\mathbb{P}\big(\forall c_{n+1,j} \in F_{n,\alpha}(P_{n+1}, C_{n+1}),\ y_{n+1,j} = 1\big) \ \ge\ 1 - \alpha,$$

which proves the marginal coverage lower bound.

Upper bound. Assume the conformity scores $\{E_i\}_{i=1}^{n+1}$ are distinct with probability one, eliminating the possibility of ties. Denote their order statistics by $E_{(1)}, \ldots, E_{(n+1)}$, which under this condition form a strictly increasing sequence almost surely. Let $k = \lceil (1-\alpha)(n+1) \rceil$. By construction,

$$\{\forall c_{n+1,j} \in F_{n,\alpha}(P_{n+1}, C_{n+1}),\ y_{n+1,j} = 1\} \iff \{E_{n+1} \le E_{(k)}\}.$$

Since the conformity scores are exchangeable and distinct, the rank of E_{n+1} is uniformly distributed on $\{1, \ldots, n+1\}$. It follows that

$$\mathbb{P}(E_{n+1} \le E_{(k)}) = \frac{k}{n+1} = \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1}.$$

Finally, since

$$\frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \le 1 - \alpha + \frac{1}{n+1},$$

We conclude that

$$\mathbb{P}\big(\forall c_{n+1,j} \in F_{n,\alpha}(P_{n+1}, C_{n+1}),\ y_{n+1,j} = 1\big) \ \le\ 1 - \alpha + \tfrac{1}{n+1},$$

which establishes the upper bound.

A.2 Proof of Theorem 2

Fix $k \in \{1, \ldots, K\}$ with $\mathbb{P}(g(P_{n+1}, C_{n+1}) = k) > 0$. By Algorithm 1 and Lemma 1,

$$\{\forall c_{n+1,j} \in F^{(k)}_{n,\alpha}(P_{n+1}, C_{n+1}),\ y_{n+1,j} = 1\} \iff \{E^{(k)}_{n+1} \le \hat{Q}^{(k)}_{1-\alpha}(\{E^{(k)}_i\}_{i \in \mathcal{I}_k})\},$$

where $\mathcal{I}_k = \{i \in [n] : g(P_i, C_i) = k\}$.

Condition on $g(P_{n+1}, C_{n+1}) = k$. Then the conformity scores $\{E^{(k)}_i : i \in \mathcal{I}_k\} \cup \{E^{(k)}_{n+1}\}$ are exchangeable. Let $m = |\mathcal{I}_k|$ and set $r = \lceil (1-\alpha)(m+1) \rceil$. If $E^{(k)}_{(1)} \le \cdots \le E^{(k)}_{(m+1)}$ are the order statistics, then

$$\{E^{(k)}_{n+1} \le \hat{Q}^{(k)}_{1-\alpha}(\{E^{(k)}_i\})\} \iff \{E^{(k)}_{n+1} \le E^{(k)}_{(r)}\}.$$

By exchangeability, $E^{(k)}_{n+1}$ is equally likely to occupy any of the $m+1$ ranks. If ties occur at the cutoff, the event $\{E^{(k)}_{n+1} \le E^{(k)}_{(r)}\}$ only becomes more likely. Therefore,

$$\mathbb{P}(E^{(k)}_{n+1} \le E^{(k)}_{(r)} \mid g(P_{n+1}, C_{n+1}) = k) \ \ge\ \frac{r}{m+1} \ \ge\ 1 - \alpha.$$

This completes the proof.

Lemma 2 (Uniformity under oracle factuality-score)**.** *If $\hat{p} = p^*$, then conditionally on (P_i, C_i) the conformity score E_i is uniformly distributed on $[0, 1]$.*

Proof. Recall from Section 4.1 that $F^{\text{oracle}}_\tau(P_i, C_i)$ denotes the set of retained claims at threshold τ. The conformity score is defined as the smallest threshold at which the retained set is entirely factual:

$$E_i := \inf\{\tau \in [0, 1] : F^{\text{oracle}}_\tau(P_i, C_i) \subseteq A_i\}.$$

By construction of the randomized oracle filter, the retention rule is calibrated to satisfy

$$\mathbb{P}\big(F^{\text{oracle}}_\tau(P_i, C_i) \subseteq A_i \,\big|\, P_i, C_i\big) = \tau.$$

This equality holds for every $\tau \in [0, 1]$. Consequently,

$$\mathbb{P}(E_i \le \tau \mid P_i, C_i) = \tau.$$

Equivalently, the conditional distribution function of E_i is

$$G_{E_i|(P_i,C_i)}(\tau) = \tau.$$

Thus, conditional on (P_i, C_i), we have $E_i \sim \text{Unif}(0, 1)$. □

A.3 Proof of Theorem 3

For notational simplicity, we suppress the dependence on (P, c) and write $\hat{p} := \hat{p}(P, c)$ and $p^* := p^*(P, c)$ throughout the proof.

Recall from equation 2 that the retention rate can be written as

$$R(p, \tau) = \mathbb{P}\big(c \in F_\tau(p; P, C)\big).$$

In the thresholding case where $F_\tau(p; P, C) = \{c : p \geq \tau\}$, this simplifies to $R(p, \tau) = \mathbb{E}[h_p]$ with $h_p := \mathbb{1}\{p \geq \tau\}$. Therefore, the retention gap is

$$\begin{aligned}\Delta &= |R(\hat{p}, \tau) - R(p^*, \tau)| \\ &= \big|\mathbb{E}[h_{\hat{p}}] - \mathbb{E}[h_{p^*}]\big| \\ &= \big|\mathbb{E}[h_{\hat{p}} - h_{p^*}]\big| \;\leq\; \mathbb{E}\big[|h_{\hat{p}} - h_{p^*}|\big],\end{aligned}$$

where we used the inequality $|\mathbb{E}[Z]| \leq \mathbb{E}[|Z|]$.

Since $h_{\hat{p}}, h_{p^*} \in \{0, 1\}$, their absolute difference equals 1 precisely when the two thresholding decisions disagree. Hence

$$\begin{aligned}\Delta &\leq \mathbb{P}\big(h_{\hat{p}} \neq h_{p^*}\big) \\ &= \mathbb{P}\big((\hat{p} - \tau)(p^* - \tau) < 0\big).\end{aligned}$$

Fix $\epsilon > 0$. If $h_{\hat{p}} \neq h_{p^*}$, then one score is above τ and the other below. This can only happen in two cases:

1. p^* lies within ϵ of τ, i.e. $|p^* - \tau| \leq \epsilon$.
2. p^* is farther than ϵ from τ but $\hat{p}$ crosses the threshold, which forces $|\hat{p} - p^*| > \epsilon$.

Therefore,

$$\{h_{\hat{p}} \neq h_{p^*}\} \subseteq \{|p^* - \tau| \leq \epsilon\} \cup \{|\hat{p} - p^*| > \epsilon\}.$$

Taking probabilities and applying the union bound gives

$$\Delta \leq \underbrace{\mathbb{P}\big(|\hat{p} - p^*| > \epsilon\big)}_{(\mathbb{I})} + \underbrace{\mathbb{P}\big(|p^* - \tau| \leq \epsilon\big)}_{(\mathbb{III})}.$$

For $(\mathbb{I})$, by Markov's inequality,

$$\mathbb{P}\big(|\hat{p} - p^*| > \epsilon\big) \leq \frac{\mathbb{E}[(\hat{p} - p^*)^2]}{\epsilon^2}.$$

For $(\mathbb{III})$, Assumption (ii) ensures

$$\mathbb{P}\big(|p^* - \tau| \leq \epsilon\big) \leq C\epsilon^\beta.$$

Hence for every $\epsilon > 0$,

$$\Delta \leq \frac{\mathbb{E}[(\hat{p} - p^*)^2]}{\epsilon^2} + C\epsilon^\beta.$$

Let $V := \mathbb{E}[(\hat{p} - p^*)^2]$. The inequality

$$\Delta \leq \frac{V}{\epsilon^2} + C\epsilon^\beta$$

holds for any $\epsilon > 0$. Hence, we may minimize the right-hand side over ϵ. Balancing the two contributions by setting $\epsilon = V^{1/(\beta+2)}$ (up to constant factors) yields

$$\Delta \leq C'\, V^{\frac{\beta}{\beta+2}},$$

where C' depends only on (C, β). This completes the proof.

Algorithm 1 Adaptive Conformal Inference (ACI)

1: **Input:** Calibration dataset $\mathcal{D}_{\text{cal}} = \{(P_i, C_i, Y_i)\}_{i=1}^{n_{\text{cal}}}$ of size n_{cal}, a new instance (P_{n+1}, C_{n+1}), a black-box classifier $\hat{p}$, and an error level $\alpha \in (0, 1)$.

2: **Output:** Filtered set $F_{n,\alpha}(P_{n+1}, C_{n+1})$ that satisfies marginal coverage.

— Calibration Phase —

3: **for** $i = 1, \ldots, n_{\text{cal}}$ **do**

4: Sample $U_i \sim \text{Unif}(0, 1)$.

5: Let $A_i = \{\, c_{i,j} \in C_i : y_{i,j} = 1 \,\}$ be the set of factual claims.

6: Compute conformity score

$$E_i = \inf\{\tau \in [0, 1] : F(\hat{p}, \tau, U_i; P_i, C_i) \subseteq A_i\}.$$

7: **end for**

8: Compute empirical quantile

$$\hat{Q}_{1-\alpha} = \inf\left\{q \in [0, 1] : \frac{1}{n_{\text{cal}}} \sum_{i=1}^{n_{\text{cal}}} \mathbb{1}\{E_i \leq q\} \geq 1 - \alpha\right\}.$$

— Filtering Phase —

9: Sample $U_{n+1} \sim \text{Unif}(0, 1)$.

10: Construct the conformal filter

$$F_{n,\alpha}(P_{n+1}, C_{n+1}) = F(\hat{p}, \hat{Q}_{1-\alpha}, U_{n+1}; P_{n+1}, C_{n+1}).$$

11: **Return** $F_{n,\alpha}(P_{n+1}, C_{n+1})$.

B Methodological Details

B.1 Details of Multi-LLM Ensemble Objective

This appendix provides the formal definitions of the proxy objective, empirical quantities, and optimization procedure underlying our multi-LLM ensemble (MACI), complementing the description in Section 4.3.

Recall from (3) that

$$R(p, \tau) = \rho \cdot \text{TPR}(p, \tau) + (1 - \rho) \cdot \text{FPR}(p, \tau).$$

Because TPR and FPR cannot be optimized simultaneously, we enforce a tolerance $\delta \in (0, 1)$ such that $\text{TPR}(p, \tau) \geq 1 - \delta$. Let $\tau_{p,\delta}$ denote the δ-quantile of factuality-scores among true-claims. The population-level objective is then

$$p^\star = \arg\min_p \ \mathbb{E}\big[\,\text{FPR}(p, \tau_{p,\delta})\big],$$

where FPR is the false positive rate at threshold $\tau_{p,\delta}$.

Since the distribution $\mathbf{P}$ is unknown, we approximate the objective using a hold-out set $\mathcal{D}_{\text{opt}} = \{(P_\ell, C_\ell, Y_\ell)\}_{\ell=1}^{n_{\text{opt}}}$. Let $N_1 = \sum_{\ell=1}^{n_{\text{opt}}} |\{c_{\ell,j} \in C_\ell : y_{\ell,j} = 1\}|$ be the total number of true-claims. The empirical δ-quantile among true-claims is

$$\widehat{\tau}_{p,\delta} = \inf\left\{t : \tfrac{1}{N_1} \sum_{\ell=1}^{n_{\text{opt}}} \sum_{c \in C_\ell : y=1} \mathbb{1}\{p(P_\ell, c) \leq t\} \geq \delta\right\}.$$

For document (P_ℓ, C_ℓ, Y_ℓ), the empirical FPR is defined as

$$\widehat{\text{FPR}}_\ell(p, \tau) = \frac{|\{c \in F_\tau(p; P_\ell, C_\ell) : y = 0\}|}{1 \vee |\{c \in C_\ell : y = 0\}|},$$

where $a \vee b = \max(a, b)$. The empirical optimization problem is then

$$\hat{p} = \operatorname*{argmin}_p \ \frac{1}{n_{\text{opt}}} \sum_{\ell=1}^{n_{\text{opt}}} \widehat{\text{FPR}}_\ell(p, \widehat{\tau}_{p,\delta}).$$

Algorithm 2 Multi-LLM Adaptive Conformal Inference (MACI)

Input: Data $\mathcal{D}_{\text{opt}} = \{(P_i, C_i, Y_i)\}_{i=1}^{n_{\text{opt}}}$ and $\mathcal{D}_{\text{cal}} = \{(P_i, C_i, Y_i)\}_{i=1}^{n_{\text{cal}}}$, of sizes n_{opt} and n_{cal} respectively, a new instance (P_{n+1}, C_{n+1}), a collection of base classifiers $\{\hat{p}_m\}_{m=1}^{M}$, a grouping function g, an error level $\alpha \in (0,1)$, and a TPR tolerance $\delta \in (0,1)$.

Output: A filtered subset $\hat{F}_{n,\alpha}^{(k_{\text{test}})}(P_{n+1}, C_{n+1})$ that satisfies group-conditional coverage.

— Optimization and Calibration Phase —

for each group $k \in \{1, \dots, K\}$ **do**

Define the optimization indices $\mathcal{I}_{\text{opt},k} = \{\, i \in \mathcal{D}_{\text{opt}} : g(P_i, C_i) = k \,\}$.

For any candidate weights w, compute the empirical threshold

$$\widehat{\tau}_{\hat{p}_{\text{ens}}(w),\delta} := \inf\Big\{ t \in \mathbb{R} : \frac{1}{N_{1,k}} \sum_{i \in \mathcal{I}_{\text{opt},k}} \sum_{c \in C_i : y=1} \mathbb{1}\{\hat{p}_{\text{ens}}(P_i, c; w) \le t\} \;\ge\; \delta \Big\},$$

where $N_{1,k} = \sum_{i \in \mathcal{I}_{\text{opt},k}} |\{c_{i,j} \in C_i : y_{i,j} = 1\}|$ is the number of true-claims in group k.

Compute the optimal ensemble weights w_k^* by solving

$$w_k^* = \arg\min_{w} \; \frac{1}{|\mathcal{I}_{\text{opt},k}|} \sum_{i \in \mathcal{I}_{\text{opt},k}} \widehat{\text{FPR}}_i\big(\hat{p}_{\text{ens}}(w), \widehat{\tau}_{\hat{p}_{\text{ens}}(w),\delta}\big)$$

$$\text{subject to} \quad \frac{1}{|\mathcal{I}_{\text{opt},k}|} \sum_{i \in \mathcal{I}_{\text{opt},k}} \widehat{\text{TPR}}_i\big(\hat{p}_{\text{ens}}(w), \widehat{\tau}_{\hat{p}_{\text{ens}}(w),\delta}\big) \;\ge\; 1 - \delta.$$

end for

for each group $k \in \{1, \dots, K\}$ **do**

Define the calibration indices $\mathcal{I}_{\text{cal},k} = \{\, i \in \mathcal{D}_{\text{cal}} : g(P_i, C_i) = k \,\}$.

Define the group-conditional ensemble classifier $\hat{p}_k^*(c) = \hat{p}_{\text{ens}}(c; w_k^*)$.

for each $i \in \mathcal{I}_{\text{cal},k}$ **do**

Sample $U_i \sim \text{Unif}(0,1)$.

Let $A_i = \{\, c_{i,j} \in C_i : y_{i,j} = 1 \,\}$ be the set of factual claims.

Compute the conformity score

$$E_i = \inf\{\tau \in [0,1] : F(\hat{p}_k^*, \tau, U_i; P_i, C_i) \subseteq A_i\}.$$

end for

Compute the group-conditional empirical quantile

$$\hat{Q}_{1-\alpha}^{(k)} = \inf\Big\{ q \in [0,1] : \frac{1}{|\mathcal{I}_{\text{cal},k}|} \sum_{i \in \mathcal{I}_{\text{cal},k}} \mathbb{1}\{E_i \le q\} \ge 1 - \alpha \Big\}.$$

end for

— Filtering Phase —

Determine the group of the new instance: $k_{\text{test}} = g(P_{n+1}, C_{n+1})$.

Retrieve the corresponding optimal weights $w_{k_{\text{test}}}^*$ and threshold $\hat{Q}_{1-\alpha}^{(k_{\text{test}})}$.

Define the group-conditional ensemble classifier $\hat{p}_{k_{\text{test}}}^*(c) = \hat{p}_{\text{ens}}(c; w_{k_{\text{test}}}^*)$.

Sample $U_{n+1} \sim \text{Unif}(0,1)$.

Construct the adaptive conformal filter:

$$\hat{F}_{n,\alpha}^{(k_{\text{test}})}(P_{n+1}, C_{n+1}) = F(\hat{p}_{k_{\text{test}}}^*, \hat{Q}_{1-\alpha}^{(k_{\text{test}})}, U_{n+1}; P_{n+1}, C_{n+1}).$$

Return $\hat{F}_{n,\alpha}^{(k_{\text{test}})}(P_{n+1}, C_{n+1})$.

Direct fine-tuning toward p^* is infeasible in black-box LLMs. Instead, let $\{p_m\}_{m=1}^{M}$ denote base factuality-scores and $w = (w_1, \dots, w_M)$ a non-negative weight vector summing to one. The en-

semble predictor is

$$p_{\text{ens}}(P, c; w) = \sum_{m=1}^{M} w_m \, p_m(P, c),$$

and the weights are optimized by

$$w^\star = \operatorname*{argmin}_{w} \frac{1}{n_{\text{opt}}} \sum_{\ell=1}^{n_{\text{opt}}} \widehat{\text{FPR}}_\ell(p_{\text{ens}}(\cdot; w), \, \widehat{\tau}_{p_{\text{ens}}(\cdot;w),\delta}).$$

C Background

C.1 Conformal Inference

Conformal Inference (CI) (Papadopoulos et al., 2002; Vovk et al., 2005; Lei et al., 2017; Angelopoulos & Bates, 2022) is a statistical framework that provides distribution-free uncertainty quantification for any machine learning model. Under the sole assumption that the data is exchangeable, a condition satisfied by i.i.d. data, CI generates a prediction set $C(X_{n+1})$ for a new test point X_{n+1} that contains the true label Y_{n+1} with a user-specified probability of at least $1 - \alpha$. This is achieved through a calibration process using a hold-out calibration dataset, D_{calib}. The core mechanism involves defining a non-conformity score function, $S(\cdot, \cdot)$, which measures how poorly a data point (X_i, Y_i) conforms to a model's predictions. For instance, a common score for a probabilistic classifier with a score function $\hat{p}$ is $S(X_i, Y_i) = 1 - \hat{p}(Y_i \mid X_i)$, where a higher score indicates that the true label was assigned a lower probability. These scores are computed for each sample in D_{calib}, and a threshold $\hat{\tau}$ is determined by taking the value at the $\lceil (|D_{\text{calib}}| + 1)(1 - \alpha) \rceil$-th position in the sorted list of scores. For a new test point X_{n+1}, the prediction set is constructed by including all possible labels $y \in \mathcal{Y}$ whose non-conformity score does not exceed this threshold, i.e., $C(X_{n+1}) = \{y \in \mathcal{Y} \mid S(X_{n+1}, y) \leq \hat{\tau}\}$. This construction provides the powerful finite-sample marginal coverage guarantee, $\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$, offering a robust foundation for building reliable machine learning systems.

C.2 False-Claim Filtering with Conformal Inference

Mohri & Hashimoto (2024) adapt the CI framework to filter false-claims from Large Language Model (LLM) outputs, proposing a foundational method we refer to as Basic Conformal Inference (BCI). The process begins with a set of n prompts, $\{P_i\}_{i=1}^n$. For each prompt P_i, an LLM generates a response R_i, which is then segmented into a collection of independent claims, $C_i = \{c_{i,1}, \ldots, c_{i,n_i}\}$. Each claim $c_{i,j}$ is associated with a ground-truth binary label $y_{i,j} \in \{0, 1\}$, where $y_{i,j} = 1$ denotes a true-claim and $y_{i,j} = 0$ denotes a false-claim. Thus, each data point is a tuple $D_i = (P_i, C_i, Y_i)$, and the dataset $\{D_i\}_{i=1}^n$ is assumed to be drawn i.i.d. from an unknown joint distribution $\mathbf{P}$. A score function, p, assigns a confidence-score $p(c_{i,j})$ to each claim. This score function can be constructed in various ways, such as by directly querying an LLM (Tian et al., 2023; Guan et al., 2024) or by capturing frequency (Wang et al., 2023; Manakul et al., 2023). Their formal goal is to output a filtered set of claims, $F_{n,\alpha}(P_i, C_i) \subseteq C_i$, that contains no false-claim with user-specified error rate, i.e.,

$$\mathbb{P}\left(\exists c_{n+1,j} \in F_{n,\alpha}(P_{n+1}, C_{n+1}) \text{ such that } y_{n+1,j} = 0\right) \leq \alpha.$$

They define the filtered set as all claims whose scores exceed the calibrated global threshold $\hat{\tau}$, that is, $F_{\hat{\tau}}(P_i, C_i) := \{\, c_{i,j} \in C_i : p(P_i, c_{i,j}) \geq \hat{\tau} \,\}$. The threshold $\hat{\tau}$ is determined by the conformal procedure. Specifically, they define a non-conformity score for each sample (C_i, Y_i) as the lowest possible confidence-score threshold τ that ensures all retained claims are true :

$$S(C_i, Y_i) := \inf \left\{ \tau \in [0, 1] : \forall c_{i,j} \in F_\tau(P_i, C_i), \; y_{i,j} = 1 \right\}.$$

This non-conformity score is computed for all samples in the calibration set D_{calib}. The global threshold $\hat{\tau}$ is then set to the $(1 - \alpha)$ quantile of these non-conformity scores, as detailed in Section C.1. They show that if the data samples are exchangeable, this procedure satisfies the desired probability guarantee.

D EXPERIMENT DETAILS

D.1 DATASETS

MedLFQA For the medical question-answering task, Cherian et al. (2024) create an experimental dataset using prompts from the MedLFQA benchmark (Jeong et al., 2024). To generate the data, they first prompt GPT-3.5-Turbo to produce new responses, which are then parsed into atomic claims by GPT-4o. For the crucial step of ground-truth annotation, they employ an automated verification procedure. For each generated claim, they prompt GPT-3.5-Turbo to verify whether it is substantiated by the reference answer provided in the original MedLFQA benchmark, effectively treating the reference answer as the ground-truth source text. From this dataset, we randomly extracted 2,000 samples, comprising 33,833 claims, for our experiments.

WikiBio Cherian et al. (2024) follow the principles of the FACTSCORE (Min et al., 2023) dataset to construct a new, large-scale benchmark for evaluating the factuality of LLM output. To generate the data, they prompt GPT-3.5-Turbo to write short biographies for 8,516 names sampled from Wikipedia. To circumvent the high cost of manual annotation, they then employ a variant of the FACTSCORE procedure for fact-checking. For each generated claim, they use the BM25 algorithm (Robertson & Zaragoza, 2009) to retrieve ground-truth passages from Wikipedia and subsequently prompt GPT-3.5-Turbo to verify whether the claim is supported by the retrieved text. This completed dataset is referred to as WikiBio in our paper for convenience. We randomly extracted 2,000 samples, comprising 53,804 claims, for our experiments.

ExpertQA Malaviya et al. (2024) construct the ExpertQA dataset, a large-scale benchmark for evaluating the factuality and attribution of LLM output. To generate the data, they first asked 484 qualified experts across 32 fields to formulate challenging, information-seeking questions from their professional lives. To ensure high-quality annotations, they then employed an expert-in-the-loop evaluation procedure where the same experts validated sentence-level claims in responses generated by six representative LLMs. Using the rich, human-annotated information provided in this dataset, we construct a binary ground truth for each claim. Among the datasets in our study, ExpertQA was the most challenging and also the most rigorously labeled, due to its direct validation by domain experts. We randomly extracted 2,000 samples, comprising 11,538 claims, for our experiments.

Of the 2,000 samples, 1,500 were used for the calibration phase, and 500 were used for the filtering (test) phase. Figure 4 shows an actual format of the data sample we use.

D.2 GROUPING CRITERIA

We employ complex grouping criteria that are likely to occur in reality, yet simultaneously require prompt and response parsing along with numeric values. We create one grouping criteria applicable to all three datasets and three classification criteria reflecting the characteristics of each dataset. The criteria are as follows:

Common: False-Claim Risk This is a composite risk index calculated by analyzing features of the prompt and response texts. The risk score increases with longer response lengths, a higher frequency of lists or numbers, and the inclusion of absolute or definitive expressions like 'always', 'never', or 'cure'. Conversely, the risk score decreases when expressions citing sources or evidence, such as 'according to' or 'research shows', are present. This index estimates the potential risk of containing false information based solely on textual characteristics.

MedLFQA: Medical Content Medical-related questions are classified into three groups based on the 'intent' of the user's prompt:
Information-Seeking (Info): Cases that ask for factual information about a specific disease or drug, using keywords like "what is," "symptom," or "treatment."
Interpretation-Seeking (Interpret): Cases that request an interpretation of what a specific symptom or condition means, using phrases like "what does it mean" or "should I worry."
Action-Seeking (Action): Cases that ask for specific guidance on actions or treatment, using phrases like "should I," "can I take," or "how to."

WikiBio: View Count Groups are divided based on the cumulative number of page views for each person's Wikipedia page in the WikiBio dataset. This is used as an indicator of public interest in or awareness of the person.

ExpertQA: Question Domain Questions (prompts) from the ExpertQA dataset are classified into three high-level domains based on the academic field specified in the official metadata:
Biology/Medicine (Bio/Med): Life science and health-related fields such as Healthcare, Medicine, Biology, Chemistry, and Psychology.
Technology/Science (Tech/Sci): Engineering and physics-related fields such as Engineering and Technology, Physics, and Astronomy.
Common: All other academic fields that do not fall into the two categories above.

D.3 Selecting LLMs

Although our methodology is model-free and works with any large language model, we choose to use Llama-3.3-70B-Instruct (Grattafiori et al., 2024), Qwen-2.5-72B-Instruct (Qwen et al., 2025), and DeepSeek-V3 (DeepSeek-AI et al., 2025). We selected these three white-box models for their high transparency and reproducibility. Their public availability and stable serving allow us to openly share and control all settings, such as decoding parameters, logs, and the calibration pipeline, making our work easily replicable. This open nature also reduces our dependence on unseen changes to policies or filters that come with version updates, which is a key advantage in fields where reproducibility and auditing are crucial.

D.4 Sampling-based Methods

SelfCheck Manakul et al. (2023) proposes a black-box, zero-resource method that detects hallucinations by sampling multiple responses from a large language model for the same query and quantifying the content consistency between the original response and the samples. Specifically, the method generates multiple stochastic responses for a single prompt and calculates response reliability by aggregating mutual consistency at the sentence/passage level, using metrics such as semantic similarity, NLI-based contradiction signals, and question-answering agreement. In our experiments, we use Llama-3.3-70B-Instruct as the model to generate multiple samples for the same prompt when applying the SelfCheck procedure.

FactSelfCheck (FSC) Sawczyn et al. (2025) proposes a method for detecting fact-level hallucinations by extracting factual units from a response and multiple samples to construct a fact graph, then aggregating supporting and contradictory signals for each fact across all samples. The procedure involves extracting facts (e.g., entity-relation-entity triplets) from an initial response and multiple samples, calculating the degree of consensus among these facts to aggregate them into fact, sentence, or passage-level scores, and finally performing threshold-based filtering. In our experiments, we also use Llama-3.3-70B-Instruct to generate the multiple samples required for the FSC procedure.

D.5 Evaluation Metrics

To evaluate our proposed method, we assess two key aspects: the quality of our oracle-approximating factuality-score function and the performance of the final filtering procedure.

Coverage Coverage is the primary metric for verifying the theoretical guarantee of our conformal inference procedure. A sample D_i is considered "covered" if its filtered set $F(C_i)$ contains no hallucinatory claims. The empirical coverage is the fraction of samples in the test set that are successfully covered. For a given error rate α, a valid conformal procedure is expected to yield an empirical coverage rate approaching or exceeding $1 - \alpha$.

$$\textbf{Cov.} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{D_i \in \mathcal{D}_{\text{test}}} \mathbb{1}\left[\forall c_{i,j} \in F(C_i), y_{i,j} = 1\right]$$

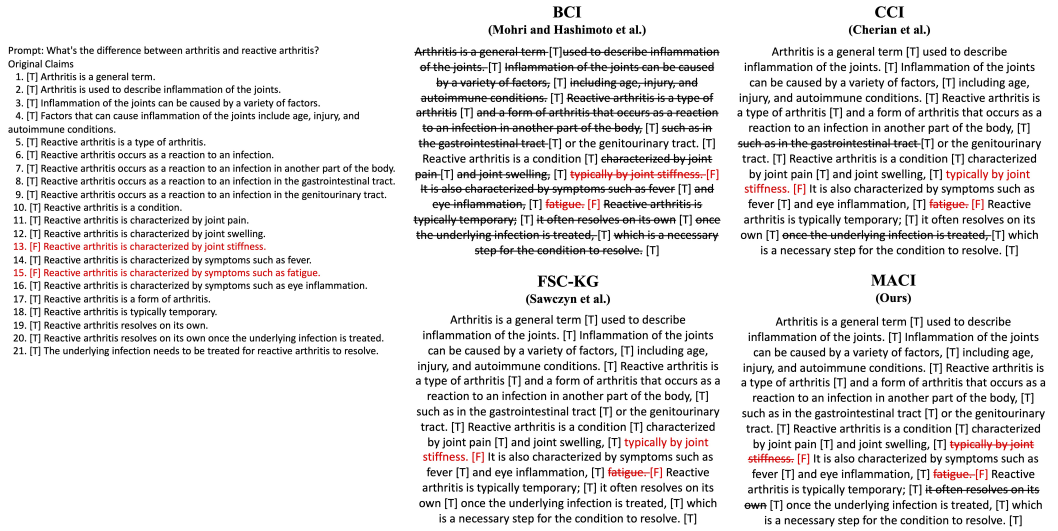


Figure 4: An example of independently decomposed claims in MedLFQA and the aggregated results of four methods that filter the false-claims of those claims. BCI yields conservative results, while CCI and FSC-KG show high retention but fail to filter out all false-claims, whereas MACI successfully filters out all false-claims.

Retention Ratio While coverage measures the safety of the filter, retention ratios measure its utility. Retention measures the average fraction of total claims remaining after filtering, indicating how much of the original text volume is preserved:

$$\textbf{Ret.} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i \in \mathcal{D}_{\text{test}}} \frac{|F(C_i)|}{|C_i|}$$

E Statement on the Use of Large Language Models

We used an LLM for minor editing and scripting automation only; core ideas, experiments, and analyses were conducted by the authors.