面向可靠LLM响应的多大语言模型自适应保形推断

Kangjun Noh[1], **Seongchan Lee**[2], **Ilmun Kim**[2],**Kyungwoo Song**[1*]

[1]延世大学应用统计与数据科学系[2]韩国科学技术院数学科学系

arXiv:2602.01285v1 [cs.LG] 1 Feb 2026

摘要

确保事实性对于在医学和法律等高风险领域安全使用大语言模型（LLMs）至关重要。保形推断提供分布无关保证，但现有方法要么过于保守，从而丢弃许多真实声明，要么依赖自适应错误率以及无法捕捉复杂群体结构的简单线性模型。为应对这些挑战，我们在乘法式过滤设定中重新表述保形推断，将事实性建模为主张级分数的乘积。我们的方法，多大语言模型自适应保形推断（**MACI**），利用集成生成更准确的事实性评分；在我们的实验中，这带来了更高的保留率，同时通过组条件校准保持了有效性。实验表明，MACI能够始终达到用户指定覆盖率，并且与基线方法相比，保留率显著更高、时间成本更低。我们的代码仓库位于https://github.com/MLAI-Yonsei/MACI.git。

1 引言

随着大语言模型（LLM）的性能不断提升，尝试直接将其响应应用于医学和法律等高风险领域的做法日益增多。然而，研究仍不断报告称，LLM响应可能包含虚假信息（Wang 等人，2024）。因此，如何确保这些关键领域中模型响应的事实性，已成为可靠使用大语言模型的一项重要挑战。

已有多种方法被提出用于确保大语言模型的事实性，但其中一些难以应用于黑盒模型（Meng 等人，2022；Zhang 等人，2025年；Chen 等人，2024a），或需要访问大型外部数据库或在线数据库（Chen 等人，2024b；Lee 和 Yu，2025年）。基于采样的方法（Manakul 等人，2023；Sawczyn 等人，2025年）相对不受这些限制，但反复检查回答一致性的过程会造成高昂的时间和经济成本，而且难以在用户指定的错误率下严格提供统计保证。

近年来，已有研究提出将保形推断（置信区间）（Papadopoulos 等人，2002；Vovk 等人，2005；Lei 等人，2018；Angelopoulos 和 Bates，2022）应用于保障大语言模型的事实性。例如，Mohri 和 Hashimoto (2024）将置信区间的概念应用于现有框架：将大语言模型的响应拆解为独立主张，并为每项主张赋予事实性评分。他们的方法提出过滤掉未达到预设阈值的主张。然而，由于这一单一的全局阈值被统一应用于所有数据，它只能提供边际覆盖率，而且可能过于保守，从而导致大量真实信息被删除。为改进这一点，Cherian 等人 (2024）提出了条件共形推断。该方法不再对所有数据采用单一的静态阈值，而是使用阈值函数，使阈值能够根据给定样本的特征而变化。但它依赖于不适用于需要固定保证的高风险应用的自适应错误率。其阈值函数也难以捕捉按语义属性划分的大语言模型响应所具有的复杂分组标准的特征。从根本上说，它们通常将符合度分数定义为

*通讯作者。

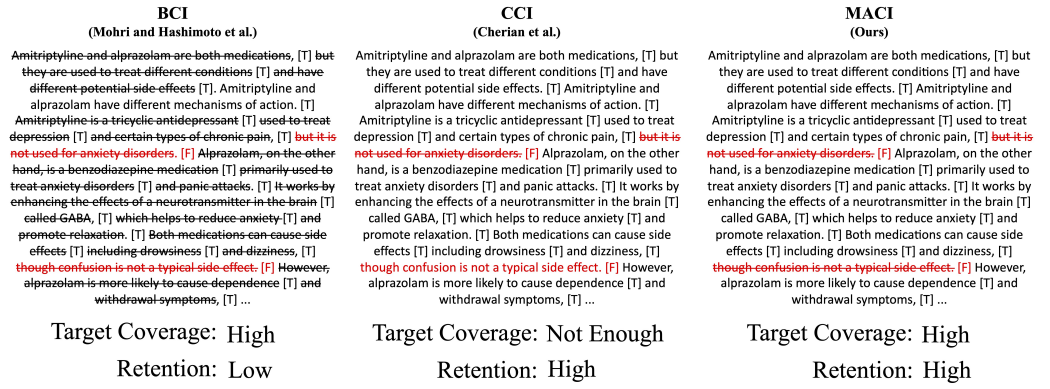


图 1：共形推断方法比较。T（真）和 F（假）表示每条主张的真实标签。基础保序推断（Mohri 和 Hashimoto，2024）通过激进过滤实现覆盖率，但导致保留率较低。条件共形推断（Cherian 等人，2024）提出了自适应阈值，但放宽了保证；MACI 同时实现了高覆盖率和高保留率。

基于单一的最坏情况分数（例如，虚假声明中最高的置信度分数），这忽略了其他主张的整体置信度，使校准流程对最坏情况分数的估计误差高度敏感，进而导致许多真实声明被过滤掉。

在此背景下，我们提出了一种名为多大语言模型自适应保形推断（MACI）的新方法论。MACI 的核心目标是在严格遵守较低用户指定错误率的同时，尽可能保留更多事实信息。假设存在理想的预言机事实性评分，我们推导出一条明确的过滤规则，该规则表示为概率的累积乘积，并在满足目标覆盖约束的条件下保留最多的主张。受这一发现启发，我们设计了一种新的自适应置信区间框架，使用由估计事实性评分的累积乘积构成的符合度分数。与依赖单个最坏情况分数计算符合度分数的既有方法不同，MACI 的符合度分数聚合了主张之间的估计事实性评分，显著提升了校准的稳健性。进一步地，我们从理论上证明，事实性评分的质量会直接影响 MACI 的效率，而效率由保留率量化。因此，我们通过多大语言模型集成调整策略，以最大化事实性评分质量。结果表明，MACI 不仅在理论上保证组条件覆盖率，而且在多样化数据集上实证展现出稳健的组条件覆盖率，同时其保留率显著高于现有方法论。图 1 展示了一个实际示例，说明与现有共形推断方法相比，MACI 在覆盖率和保留率方面的优越性。

我们的主要贡献如下：

- 我们提出一种乘性过滤框架，将事实性建模为主张级分数（事实性评分）的乘积，同时保留有限样本保证。
- 据我们所知，我们首次在保形推断中提供保留率理论分析，将理想值与估计器偏差与真实主张的保留联系起来，并为集成设计提供依据。
- 我们通过组条件校准和多大语言模型集成扩展了保形推断，确保组条件覆盖率，并在高风险领域中展现出显著高于共形基线的保留率。

2 相关工作

现有的基础保序推断（BCI）（Mohri 和 Hashimoto，2024）通过对LLM响应中的各个主张应用基于事实性评分的阈值来执行虚假主张过滤，从而提供错误率低于 α 、覆盖全部数据的无分布保证。然而，由于BCI仅在整个样本分布上提供该保证（边际覆盖率），特定子组可能出现特定于组的覆盖不足或覆盖过度，偏差也可能持续存在。

置信区间领域的许多研究致力于超越这种边际覆盖率，并提供组条件覆盖保证（Vovk，2012）。H´ebert-Johnson 等人（2018）提出

通过多重校准增强预测保证，使其能够同时满足跨多个可通过计算定义的子组的校准要求。Jung 等人（2023）提出批量多重有效保形预测，引入多重有效覆盖率的概念，确保多个群组和不同阈值水平都能同时达到覆盖率保证。这种方法进一步强化了简单的组条件覆盖率。在实践中，应用此类保证时，虽然群组间的覆盖率偏差会减小，但预测集往往会增大，保留率也会降低。Ding 等人（2023）和Gao 等人（2025年）通过对类别或群组进行聚类，或结合替代信息，分析基于群组的覆盖率与效率（集合大小）之间的权衡，并提出改进方法。Liu 和 Wu（2025年）将这些组条件覆盖方法用于确保大语言模型响应的事实性。他们通过将多重校准和多重有效保形预测应用于主张级分数校准和文档级保形预测，系统地评估不同人口统计学子群体中的事实性保证表现。Detommaso 等人（2024）利用嵌入和自我标注进行组特定多重校准，从而提高大语言模型自身置信度的可靠性。然而，更强的组条件或多重有效保证通常需要更保守的预测集，这可能会损害标准保形预测设置下的效率（例如，集合更大或弃答更多）。将此类方法直接应用于大语言模型响应过滤时，这种保守性可能导致保留率降低，从而引发对其在高吞吐量应用中实用性的担忧。我们的研究延续了组条件覆盖保证这一方向，但重点是在基于特定特征的现实群组定义下确保组条件覆盖率，同时维持较高且切实可行的保留率。Feng 等人（2025年）提出一种框架，通过将检索增强生成（Lewis 等人，2021年）应用于校准和过滤，在确保组条件覆盖率的同时增加保留主张数量。然而，这种方法实质上将保形推断从大语言模型转移到了嵌入模型和检索器等外部组件。

Cherian 等人 (2024) 与 MACI 的目标相同，即在组条件覆盖率下尽可能多地保留真实声明。他们将条件保形方法（Gibbs 等人，2025）扩展为从校准数据中学习样本特定的阈值。他们提出了一个框架，基于提示词和响应特征定义的群组实现条件保证，同时通过自适应错误率和条件提升等技术主动提高真实声明的保留率。这种同时追求组条件覆盖率和保留率这两个实际目标的方法，在确保每个群组中大语言模型回答的事实性的同时保留信息，开辟了新的平衡点。与此同时，在高风险环境中自适应方法的实用性以及阈值函数表达能力有限等局限仍然存在。 α 。

3 背景与预备知识

文档结构与事实性评分。令 $\mathcal{P}$ 表示提示词空间， $\mathcal{C}$ 表示主张空间。每个文档 $D = (P,\ C,\ Y)$ 由提示词 $P \in \mathcal{P}$、一组主张 $C = \{c_1, \ldots, c_{|C|}\} \subseteq \mathcal{C}$ 以及标签 $Y \in \{0,1\}^{|C|}$ 组成，这些标签指明哪些主张具有事实性（真实声明）。我们假设文档独立同分布地取自分布 $\mathbf{P}$，这意味着校准数据和测试数据可交换。事实性评分函数 $p\colon \mathcal{P} \times \mathcal{C} \to [0,1]$ 为每个 $(P,\ c)$ 赋予其具有事实性的概率，其中预言机为 p^*，估计量为 $\hat{}p$。

过滤算子。 给定评分函数 p、阈值 $\tau \in [0,1]$ 和可选的随机化 U，过滤算子 $F(p, \tau, U;\ P,\ C) \subseteq C$返回在 τ 下保留的主张。我们将该算子的数据驱动版本记为 $F_{n,\alpha}$，它通过在留出数据上校准 τ，以达到目标水平 α；这一校准规则是我们分析的核心对象。

组条件覆盖率。 在无分布假设情形下，精确的文档级覆盖率不可实现（Vovk，2012；Foygel Barber 等人，2021年）。因此，我们要求在能够体现领域、主题或用户群体等有意义差异的子组内保持有效性。留出样本是完整文档 $D_{n+1} = (P_{n+1},\ C_{n+1},\ Y_{n+1})$，与校准数据独立同分布地抽取自同一分布。形式上，设 $g\colon \mathcal{P} \times \mathcal{C} \to \{1, \ldots, K\}$ 为一个分组函数，它将每篇文档（$P_i,\ C_i$）分配到 K 个群组之一。每篇文档由提示词 P_i 和一组生成的主张 $C_i = \{c_{i,1}, \ldots, c_{i,N_i}\}$组成，其中 $N_i = |C_i|$ 表示其主张数量。令$A_i = \{c_{i,j} \in C_i :\ y_{i,\ j} = 1\}$ 表示文档 D_i的真实主张集合。于是，我们要求对于每个群组 $k \in \{1, \ldots, K\}$，

$$\mathbb{P}(F_{n,\alpha}(P_{n+1}, C_{n+1}) \subseteq A_{n+1} \mid g(P_{n+1}, C_{n+1}) = k) \geq 1 - \alpha. \tag{1}$$

这与 Mondrian 共形预测框架（Vovk 等人，2005）相似，但此处应用于提示–声明对，而非单个数据点。在我们的实验中，我们使用高层次、特定于数据集的类别（例如，医疗问题类型或实体组）来实例化 g。

4 MACI：多大语言模型自适应保形推断

基于第 3 节，我们的目标是设计一种过滤规则 $F_{n,\ \alpha}$ 使其满足组条件覆盖保证（1）。基线 BCI 方法采用单一全局阈值，虽然简单，却忽略了组异质性，并且依赖单个预测器。受自适应共形预测（Romano 等人，2020）的启发，我们提出 MACI，它聚合多个大语言模型的分数，并校准组条件阈值。这种方法在维持所需覆盖保证的同时提高了保留率。

4.1 ORACLE 事实性

给定第 3 节中事实性评分函数的定义，我们首先考虑一种理想化情形，其中事实性评分与真实条件概率一致。在这种预言机设定下，模型对声明正确性拥有完整的分布知识，因此对于任意提示词—声明对 $(P,\ c)$，它都能精确计算 $\mathbb{P}(y=1 \mid P,\ c)$。

定义 1（预言机过滤规则）。对于任意带有二元事实性标签$y \in \{0,1\}$的提示–声明对（$P,\ c$），预言机事实性评分定义为 $p^*(P,\ c) := \mathbb{P}(y=1 \mid P,\ c)$。对于文档$D_i = (P_i,\ C_i,\ Y_i)$及其声明集 $C_i = \{c_{i,1}, \ldots, c_{i,N_i}\}$，每个主张的预言机分数为

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

令 $[N_i] = \{1, 2, \ldots, N_i\}$ 和 $\pi_i :\ [N_i] \to [N_i]$ 为一个排列，按照预言机分数降序排列各主张，$p_i^*(c_{i,\ \pi_i\,(1)}) \geq \cdots \geq p_i^*(c_{i,\ \pi_i\,(N_i)})$，其中并列项任意排序。定义$\prod_k := \prod_{j=1}^k p_i^*(c_{i,\ \pi_i\,(j)})$，规定 $\prod_0 = 1$ 和 $\prod_{N_i+1} = 0$。对于阈值 $\tau \in [0,1]$，定义截止索引和筛选集

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

采用 max $\emptyset = 0$ 的约定。因此，F_τ^*可确保达到 τ 水平的覆盖率，并且随 τ 单调变化（$\tau_1 \leq \tau_2 \Rightarrow F_{\tau_2}^* \subseteq F_{\tau_1}^*$），但该方法较为保守，因为覆盖率通常会超过 τ。为获得精确的覆盖率，我们在边界索引 K_i^*（τ）处进行随机化。定义 γ_i（τ）$= \frac{\Pi_{K_i^*(} \in [0,1]}{\Pi_{K^*\tau} - \Pi_{K^*}}$（若分母为零，则令 γ_i（τ）为 $=0$）。令 $U_i \sim$服从 Unif(0，1)，随机化预言机规则为

$$F(p^*, \tau, U_i; P_i, C_i) = \begin{cases} \{c_{i,\pi_i(j)} : j \leq K_i^*(\tau)\}, & U_i > \gamma_i(\tau), \\ \{c_{i,\pi_i(j)} : j \leq K_i^*(\tau) + 1\}, & U_i \leq \gamma_i(\tau). \end{cases}$$

这种随机化在边界处平衡纳入与排除，从而在 τ 水平实现精确覆盖率，同时使预期保留率最大化。

4.2 面向虚假声明过滤的自适应保形推断

在介绍我们的自适应保形推断（ACI）框架之前，我们先阐明本节所采用的视角。保形推断旨在通过设计一种过滤规则，确保达到目标覆盖率，从而保证有效性。原则上，可以通过丢弃大多数主张来满足这一要求，但这种策略会过于保守。因此，在确保有效性后，我们将考察使用估算的事实性评分而非 oracle 分数时，效率会受到怎样的影响，以及在简单的理想化假设下，这种表现与预言机基准相比如何。

基于这一区分，本节余下部分将首先说明我们的方法如何确保**有效性**，然后分析其在生成信息丰富的过滤集方面的**效率**。

有效性。 如第 4.1 节所述，预言机过滤规则需要访问预言机事实性评分 p^*，因此仅作为理论基准。在实践中， p^* 是未知的，应以从黑盒验证器（例如，大语言模型）获得的估计分数 $\hat{}p$ 替代。由此得到的过滤算子 $F(\hat{}p,\ \tau,\ U;\ P,\ C)$ 沿用了预言机的结构，但其效果关键取决于阈值 τ 的选取。目标是以数据驱动方式校准 τ，从而以高概率确保所有保留主张都具有事实性，进而实现有效的有限样本覆盖率。我们使用在留出校准数据上计算的保形分位数来估计 τ。即使 $\hat{}p$ 只是对预言机分数的不完美近似，这一校准步骤仍能保证覆盖率 p^*。

为了进行这种校准，我们需要一个能够以标量形式捕捉文档级过滤事件的符合度分数。回顾一下，对于每个文档 (P_i, C_i, Y_i)， $A_i = \{c_{i,j} \in C_i : y_{i,j} = 1\}$表示真实声明的集合，并令 $U_i \sim \mathrm{Unif}(0,1)$为随机化变量。我们定义符合度分数 $E_i = \inf\{\tau \in [0,1]:\ F(\hat{}p, \tau, U_i;\ P_i, C_i) \supseteq A_i\}$，它表示使所有保留主张均为真实声明的最小阈值。每个 E_i 都将文档级过滤要求压缩为单个标量，因而可直接用于保形分位数校准（附录 A 中的引理 1）。应用标准的保形分位数论证，即可得到以下有限样本保证。

定理 1 (边际覆盖保证)。 如果样本 $(P_i,\ C_i,\ Y_i)$，对于 $i \in \{1, \ldots, n+1\}$，是可交换的，则 *ACI*（算法*1*）满足

$$\mathbb{P}\big(F_{n,\alpha}(P_{n+1}, C_{n+1}) \supseteq A_{n+1}\big) \geq 1 - \alpha$$

此外，如果分数 E_i 几乎必然互异，则边际覆盖率几乎达到紧致界：

$$\mathbb{P}\big(F_{n,\alpha}(P_{n+1}, C_{n+1}) \supseteq A_{n+1}\big) \leq 1 - \alpha + \frac{1}{n+1}.$$

边际有效性确保总体错误率在平均意义上受到控制，但这可能掩盖群组之间的差异，导致某些子群体获得较弱的保证。为解决这一问题，我们将自适应保形推断扩展到组条件设定，从而分别对每个群组强制实施有效性。该扩展遵循 Vovk 等人（2005）提出的 Mondrian 保形框架。我们不再汇集所有分数，而是将校准限制在与测试实例属于同一群组的样本上。对于群组 k 及其校准集 $\mathcal{I}_k = \{i :\ g(P_i,\ C_i) = k\}$，阈值为 $\hat{}Q^{(k)}_{1-\alpha} = \mathrm{Quantile}(\{E_i :\ i \in \mathcal{I}_k\},\ 1-\alpha)$。给定一个属于群组 k的测试实例，过滤器为 $F^{(k)}_{n,\alpha}(P_{n+1},\ C_{n+1}) = F(\hat{}p,\ \hat{}Q^{(k)}_{1-\alpha},\ U_{n+1};\ P_{n+1}, C_{n+1})$。

定理 2 （组条件覆盖保证）。 如果样本 $\{(P_i、\ C_i、\ Y_i)\}_{i=1}^{n+1}$ 可交换，则组条件保形推断规则满足

$$\mathbb{P}\big(F^{(k)}_{n,\alpha}(P_{n+1}, C_{n+1}) \supseteq A_{n+1} \,|\, g(P_{n+1}, C_{n+1}) = k\big) \geq 1 - \alpha,$$

对于所有 $k \in \{1, \ldots, K\}$，满足 $\mathbb{P}(g(P_{n+1}、\ C_{n+1}) = k) > 0$。

定理 2的一个关键推论是，它确保了每个组内有限样本、分布无关的有效性，这与定理 1的边际保证形成对比，后者仅在总体上成立。每个组都基于自身的校准集大小 $1-\alpha$ 达到相应水平 n_k，从而确保即使是小组也能获得覆盖率，但代价是采用更保守的阈值并降低保留率。

效率。 对于任何估计分数 $\hat{}p$，组条件有效性都成立，但仅凭有效性并不能确定保留多少主张。过滤规则可能满足覆盖要求，却只接纳少量主张，从而产生虽有效但不具备实际用途的结果。因此，有必要在分析有效性的同时考察效率。

为了刻画在有效性约束下可达到的效率，我们考虑预言机情形 $\hat{}p = p^*$，并假设给定 $(P_i,\ C_i)$ 时，标签 $y_{i,\ j}$ 相互独立，且服从均值为 $p^*_i(c_{i,\ j})$ 的伯努利分布。这个理想化假设仅用于获得一个易于处理的基准：（1）它将各项主张之间的条件联合分布分解开来；（2）将覆盖率约束转化为逐项主张概率的乘积；因此（3）得出一种最优有效规则，该规则可简化为对预言机得分进行逐声明阈值化。在这一简化结构下，预言机过滤器具有闭式表达式，可在满足覆盖率约束的条件下最大化保留率，并作为我们框架的效率基准。

在此预言机设定下，所得的一致性分数在 $[0,1]$（附录2 中的引理A）上恰好服从均匀分布。均匀性保证定理2 实现最大保留效率：覆盖率

恰好达到目标水平，且不存在保守性，因此不会不必要地舍弃任何真实主张。为形式化定义这一效率概念，我们引入保留率，用于衡量在给定事实性评分函数和阈值下保留的主张比例。形式化地，设 $(P,\ C,\ Y) \sim \mathbf{P}$ 为一个随机文档（参见第3节）。对于事实性评分函数 p 和阈值 τ，将保留率定义为

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

定理 3(均方误差下的保留率差距)。 令 p^* 表示预言机事实性评分，$\hat{}p$ 表示估计分数。固定阈值 $\tau \in [0,1]$，并令 $\Delta := |R(\hat{}p,\tau) - R(p^*,\tau)|$. 假设预言机事实性评分在 τ附近满足边界条件：存在常数 $\mathfrak{C} > 0$ 和 $\beta \geq 0$，使得

$$\mathbb{P}\big(|p^*(P,c) - \tau| \leq \epsilon\big) \leq \mathfrak{C}\epsilon^\beta, \qquad \forall \epsilon > 0.$$

那么，对于任意 $\epsilon > 0$,

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

对右侧关于 ϵ 进行优化可得

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

其中，$\mathfrak{C}'$ 仅取决于$(\mathfrak{C}, \beta)$。

定理 3中的假设与统计学习中的经典间隔假设相呼应（Audibert 和 Tsybakov，2007，式 (1.7)）。指数 β 刻画了预言机事实性评分在阈值附近区分样本的清晰程度。当 $\beta > 0$时，该病症反映了真实的间隔性质；而 $\beta = 0$ 则对应平凡的无间隔情形，仅为统一记号而纳入。根据这一假设，定理 3表明，保留率差距会随估计误差 $\mathbb{E}[(\hat{}p - p^*)^2]$以多项式速率缩小。常数 $\mathfrak{C}$只影响界的整体尺度。只要该常数有限，核心结论便保持不变，因为更小的估计误差仍能保证更小的保留率差距。

边际型条件在统计学习理论中应用广泛，尤其是在二元分类和经验风险最小化中。这类假设在推导有意义的统计保证方面发挥着核心作用，因此在文献中被广泛采用。

4.3 多大语言模型集成

从统计学角度看，对多个预测器进行集成可以减少偏差—方差权衡中的方差，并降低均方误差，使估计量更接近预言机基准。然而，直接最小化均方误差并不可行，因为预言机分数无法观测，而且二元标签会使预测器倾向于产生过度自信的输出，导致集成容易过拟合。因此，我们采用基于保留率分解的替代目标。通过将召回率维持在容差以上并降低假阳性率，我们可以直接提高保留率，同时避免过度自信。该替代目标仍与预言机目标一致；正如图 3 所示，它在实践中也能降低均方误差。

Let $\rho := \mathbb{P}(y = 1)$ 表示某项主张为真的边缘概率。保留率可以 b e dec组成如下。

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

其中

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

最大化 $R(p,\tau)$因此相当于提高真正例率并降低假阳性率，但这两者无法同时优化。为防止牺牲召回率的平凡解，我们要求真正例率保持在固定容差之上，即真正例率$(p,\tau) \geq 1-\delta$，对于 $\delta \in (0,1)$。以$\tau_{p,\delta}$表示真实声明中事实性评分的 δ-分位数，我们因此着重于在该约束下最小化假阳性率:

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

由于直接针对预言机进行微调 p^* 对黑盒大语言模型并不可行，我们转而针对替代最优解 $p^\star$ 在（4）中，通过采用多大语言模型集成策略。给定基础事实性评分器 $\{p_m\}_{m=1}^M$ 与权重 $w = (w_1, \ldots, w_M)$，集成预测器为

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

采用 w 优化方法，在容差约束下尽可能降低经验假正率（详见附录 B.1）。算法 2 概述了 MACI 框架，该框架将组条件保形推断与集成相结合，在保持精确覆盖率的同时最大化保留率。

5 实证结果

数据集结构与评估协议。我们使用三个特征各异的数据集 MedLFQA（Jeong 等人，2024；Cherian 等人，2024）、WikiBio（Min 等人，2023；Cherian 等人，2024）和 ExpertQA（Malaviya 等人，2024），通过实证验证 MACI 的优越性。遵循先前关于虚假声明过滤的研究（Mohri 和 Hashimoto，2024；Cherian 等人，2024），每个样本表示为四元组 {(提示词、回答、声明集、真实标签)}，其中回答被分解为原子声明，且每项声明均标注一个二元事实性标签。对于这些数据集，我们为每个基准定义一种具有代表性的分组标准，并定义一种通用的虚假声明风险分组。在所有实验中，底层回答均固定为各数据集发布的回答（主要由 GPT-4 和 GPT-3.5-turbo 生成），并且包括 CCI、BCI 和 MACI 在内的所有方法都应用于完全相同的回答集合。

估计事实性评分。为了估计事实性评分，我们使用验证指令查询一组 M 大语言模型，要求它们为每个 $[0, 1]$ 的{（提示词，声明）} 组合输出以自然语言表述的事实性评分。在主要实验中，我们使用 $M = 3$ 个模型：Llama-3.3-70B-Instruct、Qwen-2.5-72B-Instruct 和 DeepSeek-V3。这些事实性评分作为基础不确定性信号。MACI 通过算法 M 中描述的基于优化的集成方法聚合这些2分数，从而为每项声明得到一个事实性评分。由于 MACI 只需获取每项声明对应的标量分数，因此它可以作为即插即用过滤器，用于任意生成器，而不限于生成基准响应的大语言模型。

关于数据集、转换到对数空间、事实核查提示（用于生成自然语言表述的事实性评分）、组标准、大语言模型选择以及评估指标的详细说明见附录D。此外，附录E 还包含大量额外实验结果，包括与 MultiValid 保形推断和组聚类方法的比较、符合度分数的变体、联合概率建模，以及对 MACI 在协变量偏移下运行情况的讨论。

5.1 总体性能

表 1比较了三个特征各异的数据集在组条件覆盖率和保留率方面相对于虚假声明过滤领域两种主要基线方法（BCI 和 CCI）的表现。MACI 在另外两种基线方法表现不佳的设定中展现出稳健性能，在大多数群组中持续达到目标覆盖率，同时保持最高的保留率。

与 BCI 的比较。 BCI 仅通过单一阈值保证边际覆盖率，这会因群组之间的难度差异而导致覆盖过度与覆盖不足交替出现。这一现象在全部三个数据集的虚假声明风险群组以及 WikiBio 的浏览量群组结果中尤为明显。此外，BCI 使用文档 D 基于单一极端声明的符合度分数。这种设计舍弃了其余主张中可能有用的事实性信息，并使文档级一致性得分对该单一主张的估计误差过于敏感。由于共形校准会满足目标覆盖率，这种敏感性往往会导致保守阈值，进而过滤掉许多真实声明。相比之下，MACI 使用乘性的累积一致性得分，更直接地反映整个保留集联合具有事实性的可能性。通过汇总多个保留主张的信息，而非依赖单一极端声明，MACI 提供了更细粒度的符合度，从而在保持稳定覆盖率的同时实现更高的保留率。

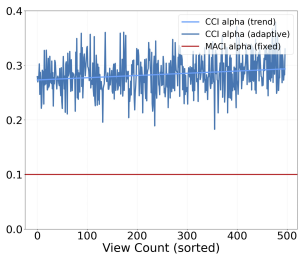
与 CCI 的比较。 CCI 通过采用自适应阈值函数，缓解了 BCI 阈值设定过于保守的问题 g_{CCI}（Cherian 等人（2024）的定理3.1）。它会调整阈值

表 1：三个特征各异的数据集的组条件覆盖率、边际覆盖率和保留率。边际结果位于与数据集名称对应的行中，后续两行则展示了该数据集两种代表性分组标准下的结果。落在 $1-\alpha\pm0.01$ 范围内的覆盖率以绿色圆点标示 ●，而超出该范围（表明覆盖过高或覆盖不足）的数值则以红色箭头标示 ↓↑。与两种保形推断基线方法相比，MACI 在大多数情况下都能达到目标覆盖率，不受群组影响。此外，它在几乎所有群组中的保留率都是最高的。 **覆盖率** 表示覆盖率，**保留率** 表示保留率。未出现覆盖不足且保留率最高的结果以**粗体**标示。所有报告数值均为 30 次重复试验的平均值。CCI 的表现是在固定目标覆盖率（$1-\alpha$）的情况下得出的。

	目标覆盖率：80% ($\alpha=0.2$)						目标覆盖率：90% ($\alpha=0.1$)						目标覆盖率：95% ($\alpha=0.05$)					
	BCI		CCI		MACI		BCI		CCI		MACI		BCI		CCI		MACI	
组	**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**
医学内容																		
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**
解读	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**
行动	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**
虚假声明风险																		
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**
中等	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**
浏览量																		
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**
中等	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**
虚假声明风险																		
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**
中等	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**
问题领域																		
生物/医学	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**
技术/科学	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**
常见	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**
虚假声明风险																		
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**
中等	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**

在样本层面实现更高的保留率。然而，其针对 D 的符合度分数仍仅依赖于单个极端声明。因此，即使 g_{CCI}经过了良好优化，这种基于单一声明的符合度分数仍容易受到该极端声明估计误差的影响，并可能导致阈值过于保守，从而限制保留率的提升。此外， g_{CCI} 采用线性特征空间框架，在应用于某些分组场景时会受到额外限制。每个数据集的分组标准（例如医学内容或虚假声明风险）都是复杂的语义函数，通过提示词和声明解析实现。因此，CCI 提出的简单线性函数和特征难以捕捉这些标准，进而导致覆盖不足或覆盖过度。与 g_{CCI} 不同，我们的分组函数 g 是一个任意可测函数，它将空间划分为有限个组，因此阈值计算不受分组标准复杂性的影响。 g_{CCI} 固有的限制也体现在其保留率上。根据其对分组标准的捕捉效果， g_{CCI} 可能会计算出过于保守的阈值。因此，与直接计算组条件阈值的 MACI 相比，其保留率较低。CCI 提出通过应用反映每个样本特征的逐样本自适应错误率来提高保留率。该方法将α 学习为一个函数，并降低每个样本的 α ，而非仅仅超过最低保留率目标。然而，这种自适应 α 与我们的目标不同。我们的目标是设计过滤规则，以高概率保证筛选集不含虚假声明，从而确保其适用于现实中的高风险领域。调整 α 以提高保留率，会产生难以

在此类场景中部署。图 2 比较了 WikiBio 上 CCI 与自适应 α，以及采用 $\alpha = 0.1$ 的 MACI。上方图表将 CCI 的目标保留率设为 MACI 的平均保留率，并输出逐样本自适应 α，显示 CCI 的 α 值通常高于 MACI 固定的小α。下方表格列出了两种方法的实际覆盖率和保留率。CCI 通过整体提高 α，使保留率几乎达到 MACI 的水平，但由于目标 α 更高，其实际覆盖率低于 MACI。



	CCI（α=自适应）		MACI（α=0.1）	
组	**Cov.**	**Ret.**	**Cov.**	**Ret.**
WikiBio	0.72	0.26	0.90•	**0.28**
浏览量				
Low	0.71	0.24	0.91•	**0.29**
中等	0.73	0.26	0.90•	**0.27**
High	0.74	0.28	0.90•	**0.31**

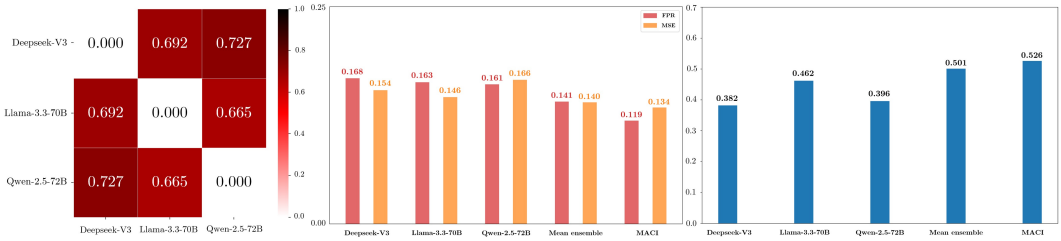
图2：在 WikiBio 数据集上，以浏览量为指标，对 CCI（自适应 α）和 MACI 的性能进行比较。左图的横轴是按浏览量排序的样本索引，纵轴是 α。左图展示了当 CCI（自适应 α）将其目标保留率设为 MACI 的平均保留率时，α 的变化。CCI（自适应 α）以较高的 α 为代价实现更高的保留率，下方表格显示了由此导致的覆盖率下降。

5.2 多大语言模型集成

在第4.1、4.2和4.3节中，我们讨论了事实性评分$\hat{}p$的重要性。因此，我们首先验证所提出的多大语言模型集成和优化方法（第4.3节）在多大程度上提升了$\hat{}p$ 相较于单 LLM 的性能，以及保留率如何相应提高。我们首先发现，不同模型在虚假声明检测方面存在显著分歧。图3 （a）显示了不同大语言模型判定为虚假的主张集合之间较高的杰卡德距离（Jaccard, 1901）。该分析仅针对 MedLFQA 中真实标签为虚假的主张子集进行。较高的距离意味着模型检测虚假声明的模式各不相同，表明通过集成有望显著提升性能。图3（b）显示，从单 LLM 到算术平均集成，再到 MACI，假阳性率和均方误差依次得到改善。图3 （b）还显示，假阳性率的改善始终伴随着均方误差的改善，并证明 MACI 的$\hat{}p$是事实性评分的更优估计量。图3（c）表明，保留率相应的逐步提升符合我们的目标，即通过提高$\hat{}p$的质量来最大化保留率。

5.3 协变量偏移下的 MACI

在已部署系统中，用于拟合 MACI 阈值的校准查询与测试时查询不一定遵循相同的协变量分布。为了研究这种协变量偏移情形，我们在 MedLFQA 上构造了明确的偏移：根据基于 SelfCheck 的事实性评分对响应进行排序，将更多事实性较强的简单查询分配到校准池，并将更多易产生幻觉的困难查询分配到测试池。这样便造成了校准分布与测试分布之间的明显失配，同时保持底层标注流程不变。遵循Tibshirani 等人(2019)提出的密度比校正思路，我们考虑一种变体 MACI-DRE，它估计$r(x) = p_X^{(t)}(x)/p_X^{(s)}(x)$ 测试协变量与校准协变量之间的密度比，所用方法是基于部署时特征（SelfCheck 评分的汇总统计量以及提示/响应长度）的轻量级分类器。随后，我们按照估计出的比率对校准集重新采样，并在该重采样集上运行原始 MACI 流程，不改变 MACI 的任何内部组件。表2 汇总了协变量偏移情形下 MedLFQA 上的结果。MACI 在某些虚假声明风险组中出现覆盖不足或覆盖过度，而 MACI-DRE 则在保持保留率相近的同时，使组级覆盖率更接近目标水平。这表明，MACI 可以与轻量级密度比估计相结合，以减轻实践中的协变量偏移。附录E中对 MACI-DRE 作了详细说明。



(a) Jaccard distance between LLMs　(b) Marginal FPR and MSE　(c) Marginal Retention Ratio

图 3：(a) 展示了不同大语言模型对 MedLFQA 中已知为假的主张所作预测之间的高杰卡德距离，表明虚假声明检测模式具有多样性，从而支持使用集成。(b) 展示了假阳性率从单 LLM 和简单算术平均集成到我们所提出的 MACI 的逐步改善。它还表明，随着假阳性率的改善，均方误差在实践中也会改善；(c) 表明随着假阳性率和均方误差的改善，保留率也会提高。更详细的分析见图7。

表2：MACI 与 MACI-DRE 的比较。MACI-DRE 通过利用重采样校准阈值，减少了出现覆盖不足或覆盖过度情形下的覆盖错误。

	目标覆盖率：80%（$\alpha = 0.2$）				**目标覆盖率：90%**（$\alpha = 0.1$）				**目标覆盖率：95%**（$\alpha = 0.05$）			
	MACI		MACI-DRE		MACI		MACI-DRE		MACI		MACI-DRE	
组	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**
MedLFQA：虚假声明风险												
Low	0.68	0.83	0.76	0.72	0.85	0.57	0.93	0.35	0.94	0.34	0.95	0.29
中等	0.65	0.77	0.84	0.56	0.82	0.55	0.94	0.27	0.87	0.42	0.96	0.19
High	0.77	0.62	0.75	0.67	0.88	0.39	0.89	0.38	0.92	0.27	0.93	0.28

5.4 时间成本

时间效率对于实时过滤至关重要。我们在两个阶段将 MACI 与基线方法进行比较：事实性评分生成阶段，用于生成事实性评分；以及校准阶段，用于优化并计算参数和阈值。我们不计入可忽略不计的过滤时间。表3报告了 WikiBio 数据集上的成本。事实性评分生成阶段通过 OpenRouter 使用 Llama-3.3-70B。结果表明，基于采样的方法由于需要反复生成响应（SelfCheck）或构建知识图谱（FSC-KG），延迟较高。相比之下，MACI 通过采用单遍评分和精简的校准流程，避免了 CCI 复杂的参数搜索，从而实现了最低的总挂钟时间。

表3：时间成本比较。挂钟时间计算方式为：校准时间 +（评分生成时间 × # 测试样本）。数值表示对 500 个 WikiBio 测试样本进行端到端处理所需的时间。请注意，CCI 的评分生成时间与 SelfCheck 相同，因为两者依赖相同的采样过程；基于采样的方法无需单独的校准阶段。

阶段	**SelfCheck**	**FSC-KG**	**CCI**	**MACI**
事实性评分（秒）	3.25 ± 0.43	19.30 ± 2.81	3.25 ± 0.43	**1.20 ± 0.13**
校准（秒）	—	—	10.33 ± 1.18	**3.24 ± 0.65**
挂钟时间（秒）	—	—	1643.91	**598.98**

6 结论

我们通过乘性过滤结构重新表述保形推断，为虚假声明过滤提供了一套具有有限样本、分布无关保证的框架。我们的分析揭示了与预言机事实性评分的偏差如何影响保留率，并由此促使我们采用集成方法来缩小这一差距。基于这些发现，我们开发了 MACI，它采用基于集成的事实性评分和组条件校准，以提供组条件覆盖保证。实验表明，MACI 在实现用户指定覆盖率的同时，大幅提高了事实主张保留率，且运行效率高于现有方法，为在高风险应用中部署大语言模型提供了一种实用方案。

参考文献

Angelopoulos, A. N. and Bates, S. (2022). 保形预测与分布无关不确定性量化简介。*arXiv*预印本 *arXiv:2107.07511*。Audibert, J.-Y. and Tsybakov, A. B. (2007). 插件分类器的快速学习率。《统计年鉴》, 35(2):608–633. Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z., and Ye, J. (2024a). INSIDE：大语言模型的内部状态仍保有幻觉检测能力。发表于第十二届国际学习表征会议（*ICLR*）。Chen, J., Kim, G., Sriram, A., Durrett, G., and Choi, E. (2024b). 利用从开放环境中检索到的证据进行复杂主张验证。发表于*2024*年计算语言学协会北美分会年会：人类语言技术（第*1*卷：长篇论文）论文集，第3569–3587页。Cherian, J., Gibbs, I., and Candes, E. (2024). 通过增强型保形预测方法确保大语言模型的有效性。神经信息处理系统进展, 37:114812–114842. DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., and et al., D. Y. (2025). DeepSeek-V3技术报告。*arXiv*预印本 *arXiv:2412.19437*。Detommaso, G., Bertran, M. A., Fogliato, R., and Roth, A. (2024). 大语言模型置信度评分的多重校准。发表于国际机器学习会议，第10624–10641页。PMLR。Ding, T., Angelopoulos, A., Bates, S., Jordan, M., and Tibshirani, R. J. (2023). 面向多类别的类别条件保形预测。神经信息处理系统进展, 36:64555–64576. Feng, N., Sui, Y., Hou, S., Cresswell, J. C., and Wu, G. (2025). 基于条件保形事实性的检索增强生成响应质量评估。发表于第*48*届国际 *ACM SIGIR* 信息检索研究与发展会议论文集，SIGIR’25，第2832–2836页。ACM。Foygel Barber, R., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). 分布无关条件预测推断的局限。《信息与推断：*IMA*期刊》，10(2):455–482. Gao, C., Gilbert, P. B., and Han, L. (2025). 弥合保形推断中公平性与效率的差距：一种代理辅助群组聚类方法。发表于第*42*届国际机器学习会议论文集，第267卷机器学习研究论文集，第18317–18336页。PMLR。Gibbs, I., Cherian, J. J., and Cand`es, E. J. (2025). 具有条件保证的保形预测。《英国皇家统计学会杂志*B*辑：统计方法论》，87(4):1100–1126. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., and et al., A. M. (2024). Llama 3模型家族。*arXiv*预印本 *arXiv:2407.21783*。Guan, J., Dodge, J., Wadden, D., Huang, M., and Peng, H. (2024). 语言模型会产生幻觉，但可能擅长事实核验。发表于*2024*年计算语言学协会北美分会年会：人类语言技术（第*1*卷：长篇论文）论文集，第1090–1111页。H´ebert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. (2018). 多重校准：面向（计算上可识别的）大众的校准。发表于国际机器学习会议，第1939–1948页。PMLR。Jaccard, P. (1901). 德朗斯河流域及邻近若干地区的阿尔卑斯植物分布。沃州自然科学学会通报, 37:241–72.

Jeong, M.、Hwang, H.、Yoon, C.、Lee, T. 和 Kang, J.（2024）。OLAPH：提升生物医学长篇问答中的事实性。*arXiv* 预印本 *arXiv:2405.12701*。Jung, C.、Noarov, G.、Ramalingam, R. 和 Roth, A.（2023）。批量多重有效保形预测。收录于国际学习表征会议（*ICLR*）。Lee, D. 和 Yu, H.（2025年）。REFIND 参加 SemEval-2025 任务 3：大语言模型中的检索增强事实性幻觉检测。*arXiv* 预印本 *arXiv:2502.13622*。Lei, J.、G'Sell, M.、Rinaldo, A.、Tibshirani, R. J. 和 Wasserman, L.（2018）。回归的无分布预测推断。美国统计学会会刊，113(523):1094–1111。Lewis, P.、Perez, E.、Piktus, A.、Petroni, F.、Karpukhin, V.、Goyal, N.、K¨uttler, H.、Lewis, M.、tau Yih, W.、Rockt¨aschel, T.、Riedel, S. 和 Kiela, D.（2021年）。面向知识密集型自然语言处理任务的检索增强生成。Liu, T. 和 Wu, S.（2025年）。面向长篇文本生成的多群体不确定性量化。收录于第*41*届人工智能中的不确定性会议。Malaviya, C.、Lee, S.、Chen, S.、Sieber, E.、Yatskar, M. 和 Roth, D.（2024）。ExpertQA：专家策划的问题与带归属标注的答案。收录于*2024*年计算语言学协会北美分会会议论文集：人类语言技术（第*1*卷：长文），第3025–3045页。Manakul, P.、Liusie, A. 和 Gales, M.（2023）。SelfCheckGPT：面向生成式大语言模型的零资源黑盒幻觉检测。收录于*2023*年自然语言处理实证方法会议论文集，第9004–9017页。Meng, K.、Bau, D.、Andonian, A. 和 Belinkov, Y.（2022）。在 GPT 中定位并编辑事实关联。神经信息处理系统进展，35:17359–17372。Min, S.、Krishna, K.、Lyu, X.、Lewis, M.、Yih, W.-t.、Koh, P.、Iyyer, M.、Zettlemoyer, L. 和 Hajishirzi, H.（2023）。FACTSCORE：长篇文本生成中事实准确度的细粒度原子评估。收录于*2023*年自然语言处理实证方法会议论文集，第12076–12100页。Mohri, C. 和 Hashimoto, T.（2024）。具有保形事实性保证的语言模型。收录于第*41*届国际机器学习会议（*ICML*）论文集，第36029–36047页。Papadopoulos, H.、Proedrou, K.、Vovk, V. 和 Gammerman, A.（2002）。用于回归的归纳式置信机器。收录于欧洲机器学习会议，第345–356页。Springer。通义千问, :、Yang, A.、Yang, B.、Zhang, B.、Hui, B.、Zheng, B.、Yu, B.、Li, C.、Liu, D.、Huang, F.、Wei, H.、Lin, H.、Yang, J.、Tu, J.、Zhang, J.、Yang, J.、Yang, J.、Zhou, J.、Lin, J.、Dang, K.、Lu, K.、Bao, K.、Yang, K.、Yu, L.、Li, M.、Xue, M.、Zhang, P.、Zhu, Q.、Men, R.、Lin, R.、Li, T.、Tang, T.、Xia, T.、Ren, X.、Ren, X.、Fan, Y.、Su, Y.、Zhang, Y.、Wan, Y.、Liu, Y.、Cui, Z.、Zhang, Z. 和 Qiu, Z.（2025年）。通义千问 2.5 技术报告。*arXiv* 预印本 *arXiv:2412.15115*。Robertson, S. 和 Zaragoza, H.（2009）。概率相关性框架：BM25 及其拓展。信息检索基础与趋势，3(4):333–389。Romano, Y.、Sesia, M. 和 Candes, E.（2020）。具有有效自适应覆盖率的分类。神经信息处理系统进展，33:3581–3591。Sawczyn, A.、Binkowski, J.、Janiak, D.、Gabrys, B. 和 Kajdanowicz, T.（2025年）。FactSelfCheck：面向大语言模型的事实级黑盒幻觉检测。*arXiv* 预印本 *arXiv:2503.17229*。Tian, K.、Mitchell, E.、Zhou, A.、Sharma, A.、Rafailov, R.、Yao, H.、Finn, C. 和 Manning, C.（2023）。只需询问校准：从经人类反馈微调的语言模型中获取校准后的置信度分数的策略。收录于*2023*年自然语言处理实证方法会议论文集，第5433–5442页。

Tibshirani, R. J.、Foygel Barber, R.、Candes, E. 和 Ramdas, A.（2019）。协变量偏移下的保形预测。神经信息处理系统进展，32。Vovk, V.（2012）。归纳式保形预测器的条件有效性。收录于亚洲机器学习会议，第475–490页。PMLR。Vovk, V.、Gammerman, A. 和 Shafer, G.（2005）。随机世界中的算法学习。Springer。Wang, X.、Wei, J.、Schuurmans, D.、Le, Q.、Chi, E.、Narang, S.、Chowdhery, A. 和 Zhou, D.（2023）。自一致性提升语言模型中的思维链推理。收录于国际学习表征会议（*ICLR*）。Wang, Y.、Wang, M.、Manzoor, M. A.、Liu, F.、Georgiev, G. N.、Das, R. J. 和 Nakov, P.（2024）。大语言模型的事实性：一项综述。收录于*2024*年自然语言处理实证方法会议论文集，第19519–19529页。Zhang, L.、Song, D.、Wu, Z.、Tian, Y.、Zhou, C.、Xu, J.、Yang, Z. 和 Zhang, S.（2025年）。通过深层内部表征分析检测大语言模型中的幻觉。收录于第三十四届国际人工智能联合会议论文集，*IJCAI-25*，第8357–8365页。国际人工智能联合会议组织。

附录

附录概述。 附录 A 包含论文中省略的主要理论结果的证明。附录B提供了更多方法细节，包括对集成目标以及我们框架中所用经验量的精确定义。附录 C回顾保形推断及其在虚假声明过滤中的应用的背景知识。附录 D报告数值实验的实现细节、数据集和评估指标，并给出补充结果。附录E包含更多实验结果。附录 F 报告了我们在研究中对大语言模型的使用。

A 主要结果的证明

引理 1。 对于每个 $i \in \{1, \ldots, n\}$、每个阈值 $\tau \in [0,1]$ 以及每个辅助随机化$U_i \sim \mathrm{Unif}(0,1)$，我们有

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

证明。 固定 $i \in \{1, \ldots, n\}$、阈值 $\tau \in [0,1]$ 和随机化变量 $U_i \sim$**Unif(0, 1)**。根据 E_i 的定义，

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

（⇒）假设 $E_i \le \tau$。那么，根据下确界的定义，存在 $\tau^* \le \tau$ 使得

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

由于保留集 $F(\hat{p},\ t,\ U_i;\ P_i,\ C_i)$ 关于 τ 单调不增，因此我们有

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

（⇐）反过来，假设

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

那么 τ 属于集合

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

因此，根据下确界的定义，我们得到 $E_i \le \tau$。结合两个方向的证明，即可得出所需的等价关系。也就是说，

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

□

A.1 定理 1 的证明

该证明遵循保形预测中边际覆盖率的标准论证，并在我们的设置中重新陈述如下。

下界。 根据算法 1 和引理 1，所有保留主张均具备事实性的事件可写为

$$\{F_{n,\alpha}(P_{n+1}, C_{n+1}) \subseteq A_{n+1}\} \iff \{E_{n+1} \le \hat{Q}_{1-\alpha}\}.$$

由于样本 $(P_i,\ C_i,\ Y_i)$ 是可交换的，且随机化 U_i 是独立同分布的，因此符合度分数 $\{E_1, \ldots, E_{n+1}\}$ 本身也是可交换的。

$$\mathbb{P}\big(E_{n+1} \le \hat{Q}_{1-\alpha}\big) \ \ge\ 1 - \alpha.$$

利用上述等价关系，我们得到

$$\mathbb{P}\big(F_{n,\alpha}(P_{n+1}, C_{n+1}) \subseteq A_{n+1}\big) \ \ge\ 1 - \alpha,$$

这证明了边际覆盖下界。

上界。 假设一致性分数 $\{E_i\}_{i=1}^{n+1}$ 以概率 1 互不相同，从而排除并列的可能性。记其次序统计量为 $E_{(1)},\dots,E_{(n+1)}$；在此条件下，它们几乎必然构成严格递增序列。令 $k=\lceil (1-\alpha)(n+1)\rceil$。根据定义，

$$\{F_{n,\alpha}(P_{n+1},C_{n+1})\subseteq A_{n+1}\} \iff \{E_{n+1}\le E_{(k)}\}.$$

由于一致性分数可交换且互不相同， E_{n+1} 的秩在 $\{1,\dots,n+1\}$ 上均匀分布。因此，

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

最后，由于

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

我们得出

$$\mathbb{P}\big(F_{n,\alpha}(P_{n+1},C_{n+1})\subseteq A_{n+1}\big) \le 1-\alpha+\frac{1}{n+1},$$

这就证明了上界。

A.2 定理的证明2

固定 $k\in\{1,\dots,K\}$ 其中 $\mathbb{P}(g(P_{n+1},C_{n+1})=k)>0$。根据算法 1 和引理 1，

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

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

给定 $g(P_{n+1},C_{n+1})=k$。那么符合度分数 $\{E^{(k)}_i:\ i\in\mathcal{I}_k\}\cup\{E^{(k)}_{n+1}\}$ 是可交换的。令 $m=|\mathcal{I}_k|$ 并设 $r=\lceil (1-\alpha)(m+1)\rceil$。若 $E^{(k)}{}_{(1)}\le\cdots\le E^{(k)}{}_{(m+1)}$ 是次序统计量，则

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

根据可交换性， $E^{(k)}_{n+1}$ 占据 $m+1$ 个排名中的任意一个的可能性相同。如果截断位置出现并列，则事件 $\{E^{(k)}_{n+1}\le E^{(k)}_{r}{}_{()}\}$ 只会变得更有可能。因此，

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

证明至此完成。

引理2 (预言机事实性评分下的均匀性)。 若$\hat{}p=p^*$，则在给定(P_i,C_i) 的条件下，符合度分数 E_i 在 $[0,1]$上服从均匀分布。

证明。 回顾第4.1节， $F_{\text{预言机}\tau}(P_i,\ C_i)$ 表示阈值为 τ时的保留主张集。符合度分数定义为使保留集完全由事实性主张构成的最小阈值：

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

根据随机化预言机过滤器的构造，保留规则经过校准，满足

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

此等式对任意 $\tau\in[0,1]$ 均成立。因此，

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

等价地， E_i 的条件分布函数为

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

因此，在给定（P_i，C_i）的条件下，我们有 $E_i^*\sim$Unif(0，1)。 □

A.3 定理 3

为简化记号，我们省略对（$P,\ c$）的依赖，并在整个证明中记作 $\hat{p} := \hat{p}(P,\ c)$ 和 $p^* := p^*(P,\ c)$。我们还要强调，下面的论证聚焦于一个采用固定阈值 τ 的简化阈值规则。本文不分析 MACI 中使用的完整数据依赖型乘法过滤过程，而是考察一个基本决策规则 $h_p = \mathbb{1}\{p \geq \tau\}$，以阐明 p 中的估计误差如何影响保留率。

由（2）可知，保留率可写为

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

在阈值情形下，若 $F_\tau(p;\ P,\ C) = \{c\colon\ p \geq \tau\}$，则可简化为 $R(p,\ \tau) = \mathbb{E}[h_p]$ 且 $h_p := \mathbb{1}\{p \geq \tau\}$。因此，保留率差距为

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

其中我们使用了不等式 $|\mathbb{E}[Z]| \leq \mathbb{E}[|Z|]$。

由于 $h_{\hat{p}},\ h_{p^*} \in \{0,1\}$，它们的绝对差恰好在两个阈值决策不一致时等于 1。因此

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

固定 $\epsilon > 0$。若 $h_{\hat{p}} \neq h_{p^*}$，则一个分数高于 τ，另一个低于它。这只可能在两种情况下发生：

1. p^* 位于距 τ 不超过 ϵ 的范围内，即 $|p^* - \tau| \leq \epsilon$。2. p^* 距 τ 比 ϵ 更远，但 $\hat{p}$ 越过阈值，从而迫使 $|\hat{p} - p^*| > \epsilon$。

因此，

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

取概率并应用并集界可得

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

对于（$\mathbb{I}$），由马尔可夫不等式，

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

对于（$\mathbb{III}$），假设（边界条件）保证

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

因此，对于每一个 $\epsilon > 0$，

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

令 $V := \mathbb{E}[(\hat{p} - p^*)^2]$。不等式

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

对任意 $\epsilon > 0$ 成立。因此，我们可以对 ϵ 最小化右端。通过设定 $\epsilon = V^{1/(\beta+2)}$（至多相差常数因子）来平衡这两项贡献，可得

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

其中 $\mathfrak{C}'$ 仅取决于 $(\mathfrak{C}, \beta)$。至此，证明完成。

算法 1自适应保形推断（ACI）

1: **输入：** 大小为 n_{cal} 的校准数据集 $\mathcal{D}_{\text{cal}} = \{(P_i, C_i, Y_i)\}_{i=1}^{n_{\text{cal}}}$、一个新实例 (P_{n+1}, C_{n+1})、一个黑盒分类器 $\hat{}p$，以及一个误差水平 $\alpha \in (0,1)$。

2: **输出：** 满足边际覆盖率的筛选集 $F_{n,\ \alpha}(P_{n+1},\ C_{n+1})$。

— 校准阶段 —3: **对于** $i = 1,\ ..,\ n_{\text{cal}}$ **执行**4: 对 $U_i \sim$ 进行均匀采样（0，1）。5: 令 $A_i = \{\, c_{i,j} \in C_i : y_{i,\ j} = 1 \,\}$ 为事实性声明集合。6: 计算符合度分数

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

7: **结束循环**

8: 计算经验分位数

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

— 过滤阶段 —

9：采样 $U_{n+1} \sim \text{Unif}(0,1)$。

10：构建共形滤波器

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

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

B 方法细节

B.1 多大语言模型集成目标的细节

本附录给出了支撑我们多大语言模型集成（MACI）的代理目标、经验量和优化过程的形式化定义，以补充第4.3节中的描述。

回顾（3）可知

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

由于无法同时优化真正例率（TPR）和假阳性率（FPR），我们设定容差 $\delta \in (0,\ 1)$，使得 $\text{TPR}(p,\ \tau) \ge 1-\delta$。令 $\tau_{p,\ \delta}$ 表示真实声明中事实性评分的 δ 分位数。总体目标随后为

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

其中，FPR 是阈值 $\tau_{p,\ \delta}$ 下的假正率。

由于分布 $\mathbf{P}$ 未知，我们使用留出集 $\mathcal{D}_{\text{opt}} = \{(P_\ell,\ C_\ell,\ Y_\ell)\}_{\ell=1}^{n_{\text{opt}}}$近似该目标。令 $N_1 = \sum_{\ell=1}^{n_{\text{opt}}} |\{c_{\ell,j} \in C_\ell : y_{\ell,\ j} = 1\}|$ 表示真实声明的总数。真实声明中的经验 δ 分位数为

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

对于文档（P_ℓ, C_ℓ, Y_ℓ），经验假正率定义为

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

其中 $a \vee b = \max(a,b)$。经验优化问题则为

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

算法 2 多大语言模型自适应保形推断（MACI）

1: 输入： 数据 $\mathcal{D}_{\text{opt}} = \{(P_i, C_i, Y_i)\}_{i=1}^{n_{\text{opt}}}$ 和 $\mathcal{D}_{\text{cal}} = \{(P_i, C_i, Y_i)\}_{i=1}^{n_{\text{cal}}}$，大小分别为 n_{opt} 和 n_{cal}，一个新实例 (P_{n+1}, C_{n+1})，一组基础分类器 $\{\hat{p}_m\}_{m=1}^{M}$，一个分组函数 g，一个误差水平 $\alpha \in (0,1)$，以及一个真正率容差 $\delta \in (0,1)$。

2: 输出： 满足组条件覆盖率的过滤子集 $\hat{F}^{(k_{\text{test}})n,\ \alpha}(P_{n+1},\ C_{n+1})$。

— 优化与校准阶段 —

3: 对于 每个组 $k \in \{1, \ldots, K\}$ **执行**

4: Define th$^{\text{e opt}}$ imization indices $\mathcal{I}_{\text{opt},k} = \{\, i \in \mathcal{D}_{\text{opt}} :\ g(P_i, C_i) = k \,\}$。

5: 对于任意候选权重 w，计算经验阈值

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

其中 $N_{1,\ k} = \sum_{i \in \mathcal{I}_{\text{opt}',\ k}} |\{c_{i,j} \in C_i : y_{i,\ j} = 1\}|$ 是组 k 中真实声明的数量。

6：通过求解计算最优集成权重 w_k^*

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

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

7: **结束循环**

8：**对于** 每个组 $k \in \{1,\ \ldots,\ K\}$

执行9：定义校准索引 $\mathcal{I}_{\text{cal},\ k} = \{\, i \in \mathcal{D}_{\text{cal}} :\ g(P_i,\ C_i) = k \,\}$。10：定义组条件集成分类器 $\hat{p}_k^*(c) = \hat{p}_{\text{ens}}(c; w_k^*)$。11：**对于**每个$i \in \mathcal{I}_{\text{cal},\ k}$ **执行**12：从 Unif(0，1) 中采样 $U_i \sim$。13：令 $A_i = \{\, c_{i,j} \in C_i :\ y_{i,\ j} = 1 \,\}$ 为事实性声明集合。14：计算一致性分数

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

15：**结束循环**

16：计算组条件经验分位数

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

17: **结束循环**

— 过滤阶段 —

18：确定新实例所属的组：$k_{\text{test}} = g(P_{n+1},\ C_{n+1})$。19：获取对应的最优权重 $w_{k_{\text{test}}}^*$ 和阈值 $\hat{Q}^{(k_{\text{test}})\ 1-\alpha}$。20：定义组条件集成分类器 $\hat{p}_{k_{\text{test}}}^*(c) = \hat{p}_{\text{ens}}(c; w_{k_{\text{test}}}^*)$。21：从 Unif(0，1) 中采样 $U_{n+1} \sim$。22：构建自适应共形滤波器：

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

23：**返回** $\hat{F}^{(k_{\text{test}})n,\ \alpha}(P_{n+1},\ C_{n+1})$。

在黑盒大语言模型中，无法直接针对 p^* 进行微调。作为替代，令 $\{p_m\}_{m=1}^{M}$ 表示基础事实性分数，$w = (w_1、\ldots\ldots、w_M)$表示权重之和为一的非负权重向量。集

成预测器为

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

并且权重通过以下方式进行优化

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

C 背景

C.1 保形推断

保形推断（CI）（Papadopoulos 等人，2002；Vovk 等人，2005；Lei 等人，2018；Angelopoulos 与 Bates，2022）是一种统计框架，可为任意机器学习模型提供与分布无关的不确定性量化。在数据可交换这一唯一假设下（独立同分布数据满足该条件），CI 会为新的测试点X_{n+1} 生成预测集 $C(X_{n+1})$，使其以用户指定的至少 $1-\alpha$的概率包含真实标签 Y_{n+1}。这通过使用留出校准数据集 D_{calib}的校准流程来实现。其核心机制是定义非一致性评分函数 $S(\cdot,\cdot)$，用于衡量数据点(X_i, Y_i)与模型预测的不符程度。例如，对于评分函数为 $\hat{}p$ 的概率分类器，一种常见分数为 $S(X_i, Y_i) = 1 - \hat{}p(Y_i \mid X_i)$，其中更高的分数表明真实标签被赋予了更低的概率。这些分数针对 D_{calib}中的每个样本计算，阈值$\hat{}\tau$ 则取分数排序列表中第 $\lceil(|D_{\text{calib}}|+1)(1-\alpha)\rceil$位的值来确定。对于新的测试点 X_{n+1}，预测集的构造方式是纳入所有非一致性分数不超过该阈值的可能标签 $y \in \mathcal{Y}$，即$C(X_{n+1}) = \{y \in \mathcal{Y} \mid S(X_{n+1}, y) \leq \hat{}\tau\}$。该构造提供了强有力的有限样本边际覆盖保证 $\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1-\alpha$，为构建可靠的机器学习系统奠定了稳健基础。

C.2 使用保形推断进行虚假声明过滤

Mohri 和 Hashimoto（2024）将置信区间框架改编用于从大语言模型（LLM）的输出中过滤虚假声明，并提出一种基础方法，我们称之为基础保形推断（BCI）。该过程从一组 n 提示 $\{P_i\}_{i=1}^n$开始。对于每个提示 P_i，大语言模型会生成一个回答 R_i，随后将其划分为一组独立主张 $C_i = \{c_{i,1}, \ldots, c_{i,N_i}\}$。每个主张$c_{ij}$ 都对应一个真实二元标签$y_{ij} \in \{0,1\}$，其中$y_{ij} = 1$ 表示真实主张，$y_{ij} = 0$ 表示虚假声明。因此，每个数据点都是一个元组 $D_i = (P_i, C_i, Y_i)$，并假设数据集 $\{D_i\}_{i=1}^n$ 独立同分布地取自某个未知的联合分布 P。评分函数 p会为每项主张赋予一个置信度分数 $p(c_{i,j})$。该评分函数可以通过多种方式构建，例如直接查询大语言模型（Tian 等人，2023；Guan 等人，2024），或捕捉频率（Wang 等人，2023；Manakul 等人，2023）。他们的正式目标是输出一个主张筛选集 $F_{n\alpha}(P_i, C_i) \subseteq C_i$，其中不包含任何虚假声明，且错误率不超过用户指定的水平，即

$$\mathbb{P}\big(F_{n,\alpha}(P_{n+1}, C_{n+1}) \not\subseteq A_{n+1}\big) \leq \alpha,$$

他们将筛选集定义为所有分数超过校准后的全局阈值$\hat{}\tau$的主张，即 $F_{\hat{\tau}}(P_i, C_i) := \{\, c_{i,j} \in C_i : p(P_i, c_{i,j}) \geq \hat{}\tau \,\}$. 阈值 $\hat{}\tau$ 由保形程序确定。具体而言，他们将每个样本 (C_i, Y_i) 的非一致性分数定义为确保所有保留主张均为真的最低可能置信度分数阈值 τ：

$$S(C_i, Y_i) := \inf\{\tau \in [0,1] : F_\tau(P_i, C_i) \subseteq A_i\}$$

针对校准集中的所有样本计算该非一致性分数 D_{calib}。随后，将全局阈值$\hat{}\tau$ 设为这些非一致性分数的$(1-\alpha)$ 分位数，具体说明见第C.1节。他们表明，如果数据样本可交换，该方法便能满足所需的概率保证。

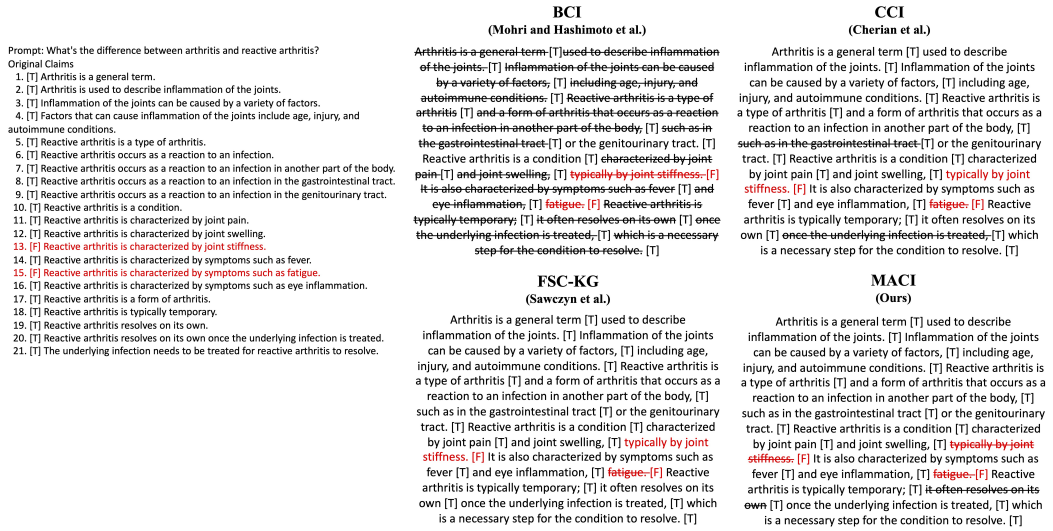


图 4：MedLFQA 中独立拆解的主张示例，以及四种方法筛除这些主张中的虚假声明后的汇总结果。BCI 得出的结果较为保守，而 CCI 和 FSC-KG 的保留率较高，但未能筛除所有虚假声明；相比之下，MACI 成功筛除了所有虚假声明。

D 实验详情

转换到对数空间。 我们的实现将所有乘法计算都在对数空间中进行。给定主张级别的事实性评分 $\hat{p}(P,\ c)\ \in\ [0,1]$ 和一个较小的 $\epsilon\ >\ 0$，我们应用基于补数的变换

$$\ell(P,c) := -\log\bigl(1-\hat{p}(P,c)+\epsilon\bigr),$$

令 π 表示由 $\hat{p}$ 诱导的排序（按升序排列，因此首先考虑低置信度主张），并定义补数乘积

$$\Pi_k(P,C) := \prod_{j=1}^{k}\bigl(1-\hat{p}(P,c_{\pi(j)})+\epsilon\bigr).$$

于是我们得到精确的对数乘积恒等式

$$-\log\Pi_k(P,C)\ =\ \sum_{j=1}^{k}\ell(P,c_{\pi(j)}).$$

因此，我们筛选器中使用的预算规则 $\sum_{j=1}^{k}\ell(P,\ c_{\pi(j)})\ \leq\ \tau$等价于 $\Pi_k(P,\ C)\ \geq\ e^{-\tau}$。由于映射 $x\ \mapsto\ -\log\ x$ 在 $(0,1]$, 上严格单调，因此这种对数空间实现保留了诱导出的选择集合和保形分位数，同时通过避免乘积下溢提高了数值稳定性。

D.1 数据集

MedLFQA。 对于医学问答任务，Cherian 等人(2024) 使用 MedLFQA 基准中的提示构建了一个实验数据集（Jeong 等人，2024）。为生成数据，他们首先提示 GPT-3.5-Turbo 生成新的响应，再由 GPT-4o 将这些响应解析为原子声明。对于至关重要的真实标注步骤，他们采用了一种自动验证程序。针对每条生成的主张，他们提示 GPT-3.5-Turbo 验证该主张是否得到原始 MedLFQA 基准所提供的参考答案支持，实际上是将参考答案视为真实来源文本。我们从该数据集中随机抽取了 2,000 个样本，共包含 33,833 条主张，用于实验。

WikiBio。 Cherian 等人(2024) 遵循 FACTSCORE（Min 等人，2023）数据集的原则，构建了一个新的大规模基准，用于评估大语言模型输出的事实性。为生成数据，他们提示 GPT-3.5-Turbo 为从维基百科抽取的 8,516 个姓名撰写简短传记。为了避免人工标注的高昂成本，他们随后采用 FACTSCORE 程序的一个变体进行事实核查。针对每条生成的主张，他们使用 BM25 算法（Robertson 和 Zaragoza，2009）从维基百科检索真实依据段落，随后提示 GPT-3.5-Turbo 验证该主张是否得到检索文本的支持。为方便起见，本文将这一完整数据集称为 WikiBio。我们随机抽取了 2,000 个样本，共包含 53,804 条主张，用于实验。

ExpertQA。 Malaviya 等人（2024）构建了 ExpertQA 数据集，这是一个用于评估大语言模型输出的事实性和归因的大规模基准。为生成数据，他们首先邀请了来自 32 个领域的 484 位合格专家，根据各自的职业经历提出具有挑战性、旨在寻求信息的问题。为确保标注质量，他们随后采用专家参与式评估流程，由同一批专家验证六个具有代表性的大语言模型所生成响应中的句子级主张。利用该数据集中丰富的人工标注信息，我们为每项主张构建了二元真实依据。在我们的研究所用数据集中，ExpertQA 最具挑战性，标注也最为严谨，因为其内容由领域专家直接验证。我们随机抽取了 2,000 个样本，共包含 11,538 项主张，用于实验。

在这 2,000 个样本中，1,500 个用于校准阶段，500 个用于过滤（测试）阶段。图 4 展示了我们所用数据样本的实际格式。

D.2 分组标准

我们采用复杂的分组标准，这些标准很可能在现实中出现，但同时需要提示与回复解析以及数值。我们创建了一项适用于全部三个数据集的分组标准，以及三项反映各数据集特征的分类标准。标准如下：

常见：虚假声明风险。 这是通过分析提示词与回复文本的特征计算得出的综合风险指数。风险分数随回复长度增加、列表或数字出现频率升高，以及包含“总是”“从不”或“治愈”等绝对或确定性表达而上升。相反，当出现引用来源或证据的表达（如“根据”或“研究表明”）时，风险分数下降。该指数仅基于文本特征估计包含虚假信息的潜在风险。

MedLFQA：医学内容。 医疗相关问题根据用户提示词的“意图”分为三个群组：

信息寻求（信息）：询问特定疾病或药物事实信息的情形，使用诸如“是什么”“症状”或“治疗”等关键词。

Inte解读寻求（解读）：请求解读某种具体症状
or c意味着什么，使用“这意味着什么”或“我需要担心吗”等措辞。
行动寻求（行动）：询问有关行动或治疗的具体指导，使用以下措辞
例如“我应该，”“我可以服用，”或“该如何做”。

WikiBio：浏览量。根据 WikiBio 数据集中每个人的维基百科页面累计浏览量对组进行划分。这一指标用于反映公众对该人物的兴趣或认知度。

ExpertQA：问题领域。根据官方元数据中指定的学术领域，将 ExpertQA 数据集中的问题（提示）划分为三个高层级领域：
生物学/医学（生物/医学）：生命科学和健康相关领域，例如医疗保健、医学、生物学、化学和心理学。

技术/科学（技术/科学）：工程和物理相关领域，例如工程与技术、物理学和天文学。

常见：不属于上述两类的所有其他学术领域。

D.3 事实核查提示

下面的提示词是我们输入给 M 大语言模型的提示词示例，用于获取每个主张的文字化事实性评分。

事实核查提示

指令。 你是一位能够访问庞大知识库的专业事实核查员和逻辑学家。你的任务是评估从模型对给定提示词的回答中提取的特定主张的事实准确性。请遵循以下步骤，以得出精确的评估：

1. **语境分析**：在原始提示词的语境中理解该主张。2. **知识核验**：运用你的一般世界知识以及所提供的信息，对照既定事实、适用时的科学或医学共识，以及基本逻辑一致性来核查该主张。3. **概率估计**：估计该主张在事实上为真的概率。**评分标准（真实性概率）。** 在 $[0.0, 1.0]$之间分配一个连续分数，以表示你的置信度：

- **1.0（确定为真）**：该主张在公理上为真，或已由非常有力的证据验证。
- **0.8–0.9（高置信度）**：专家普遍认可该主张为真；可能缺少一些细节背景，但核心内容正确。
- **0.4–0.6（模棱两可/不确定）**：该主张存在争议、仅得到部分支持，或缺乏足够的信息来核实。
- **0.1–0.3（低置信度）**：该主张与现有证据相矛盾，或包含重大错误。
- **0.0（错误）**：该主张已被证明确实错误、在逻辑上不可能，或完全是捏造的。**输入。原始提示词：**{prompt_text}**待评估的主张：**{claim_text}**输出格式。**请返回一个包含数组 "evaluations" 的单个 JSON 对象：{ "evaluations": [{ "claim_id": 1, "reasoning": "此处填写逐步推理……", "score": 0.85 }]}

D.4 选择大语言模型

尽管我们的方法论与模型无关，适用于任何大语言模型，但我们选择使用 Llama-3.3-70B-Instruct（Grattafiori 等人，2024）、Qwen-2.5-72B-Instruct（Qwen 等人，2025年）和 DeepSeek-V3（DeepSeek-AI 等人，2025年）。我们选择这三种白盒模型，是因为它们具有较高的透明度和可复现性。它们公开可用且服务稳定，使我们能够公开分享并控制所有设置，例如解码参数、日志和校准流程，从而让我们的工作易于复现。这种开放性也减少了我们对版本更新所带来的策略或过滤器隐性变化的依赖；在可复现性和审计至关重要的领域，这是一项关键优势。

D.5 基于采样的方法

SelfCheck。 Manakul 等人（2023）提出了一种黑盒、零资源方法：针对同一查询，从大语言模型中采样多个响应，并量化原始响应与样本之间的内容一致性，以检测幻觉。具体而言，该方法针对单个提示词生成多个随机响应，并通过汇总句子/段落级别的相互一致性来计算响应可靠性，

其中使用语义相似度、基于自然语言推断的矛盾信号以及问答一致性等指标。在我们的实验中，应用 SelfCheck 流程时，我们使用 Llama-3.3-70B-Instruct 模型针对同一提示词生成多个样本。

FactSelfCheck（FSC）。Sawczyn 等人（2025年）提出了一种检测事实层面幻觉的方法：从回答和多个样本中提取事实单元以构建事实图，然后汇总所有样本中每项事实的支持信号与矛盾信号。该过程包括从初始回答和多个样本中提取事实（例如，实体-关系-实体三元组），计算这些事实的共识程度，并将其汇总为事实、句子或段落层面的分数，最后进行基于阈值的过滤。在我们的实验中，我们也使用 Llama-3.3-70B-Instruct 生成 FSC 过程所需的多个样本。

D.6 评估指标

To 为评估我们提出的方法，我们考察两个关键方面：预言机的质量 -
app近似事实性评分函数的能力，以及最终过滤过程的性能。

覆盖率。覆盖率是验证我们的保形推断程序理论保证的主要指标。如果筛选集 $F(C_i)$ 中不包含任何幻觉性声明，则认为样本 D_i “受到覆盖”。经验覆盖率是测试集中成功受到覆盖的样本比例。对于给定的错误率 α，有效的保形程序预期会产生接近或超过 $1-\alpha$ 的经验覆盖率。

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

保留率。 覆盖率衡量过滤器的安全性，而保留率衡量其效用。保留率是指过滤后剩余主张占主张总数的平均比例，表明原始文本内容保留了多少：

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

E 附加结果

与采用统一事实性分数的 BCI、CCI 和 MACI 进行比较。 定理 3 表明，事实性评分的质量会直接影响 MACI 的保留率。因此，通过基于多大语言模型的加权优化集成提升事实性评分的质量，将直接提高保留率，使其高于基于保形推断的基线方法 MACI。因此，有必要验证 MACI 中排除多大语言模型集成部分、以模拟 Oracle 事实性的自适应保形评分结构，对保留率的贡献有多大。为此，我们将 MACI 的事实性评分固定为 BCI 和 CCI 所采用的频率评分。也就是说，我们在排除多大语言模型集成组件的情况下，对比 MACI 与 BCI、CCI 的覆盖率和保留率。表 4 显示，即便采用相同的统一事实性分数，MACI 仍取得最高保留率。这表明，MACI 受 Oracle 启发的自适应保形推断结构本身优于基线方法。

与 MultiValid 保形推断的比较。 多重有效保形预测领域的研究提出了一个稳健的框架，可同时保证多个子组和阈值区间的覆盖率。具体而言，Jung 等人(2023）提出的批量多重有效保形预测定义了多重有效覆盖率的概念，即在单一阈值函数下，同时满足多个群组的组条件覆盖率以及不同阈值水平的阈值级条件覆盖率。形式上，设 $s(X,Y)$ 表示一个符合度分数， $\Gamma_\tau(X)$ 表示由阈值函数 $\tau(\cdot)$所诱导的预测集，并设 $\mathcal{G}, \mathcal{I}$ 表示由群组和分数区间构成的族，则多重有效覆盖率要求对所有 $G \in \mathcal{G}$ 和 $I \in \mathcal{I}$，

$$\mathbb{P}\big(Y \in \Gamma_\tau(X) \,\big|\, X \in G,\ s(X,Y) \in I\big) \ \geq\ 1-\alpha-\varepsilon,$$

表 4：与采用统一事实性分数的共形基线进行比较。覆盖率在 $1-\alpha \pm 0.01$ 范围内的值以绿色圆点标记 •，超出该范围的值以红色箭头标记 ↓↑。

	目标覆盖率：80% ($\alpha = 0.2$)						目标覆盖率：90% ($\alpha = 0.1$)						目标覆盖率：95% ($\alpha = 0.05$)					
	BCI		CCI		MACI		BCI		CCI		MACI		BCI		CCI		MACI	
组	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**
MedLFQA：虚假声明风险																		
Low	0.84↑	0.07	0.83↑	0.68	0.82↑	**0.70**	0.94↑	0.03	0.91•	0.41	0.91•	**0.49**	0.97↑	0.01	0.95•	0.28	0.96•	**0.32**
中等	0.83↑	0.06	0.81•	0.66	0.81•	**0.71**	0.89•	0.03	0.90•	0.39	0.91•	**0.46**	0.94•	0.01	0.95•	0.25	0.96•	**0.30**
High	0.73↓	0.06	0.78↓	0.43	0.81•	**0.58**	0.88↓	0.01	0.89•	0.22	0.90•	**0.36**	0.94•	0.01	0.94•	0.12	0.95•	**0.24**

在小容差 $\varepsilon \geq 0$ 下成立。这比标准组条件覆盖率要求更强，因为后者仅以组成员身份为条件。

虽然这种方法提供了强有效性保证，但由于需要同时满足多个群组和阈值水平，它可能导致预测集过于保守且规模较大。MACI 面向一种更具体的应用场景，重点是对大语言模型生成的响应进行虚假声明过滤，旨在为少数有意义的群组（例如根据虚假声明风险划分的群组）维持组条件覆盖率，同时实现适用于实际系统的保留率。具体而言，如果 $\mathcal{G}_{\text{risk}}$ 表示一个特定于任务的分区（例如按虚假声明风险划分为低、中等、高），MACI 将确保

$$\mathbb{P}\big(Y \in C(X) \,\big|\, Z \in g\big) \;\geq\; 1-\alpha, \qquad \forall g \in \mathcal{G}_{\text{risk}},$$

而不再额外以分数区间为条件。因此，MACI 专注于在虚假声明过滤这一特定任务中平衡覆盖率与效率，而非针对不同群组和阈值区间提供通用的多重有效保证。

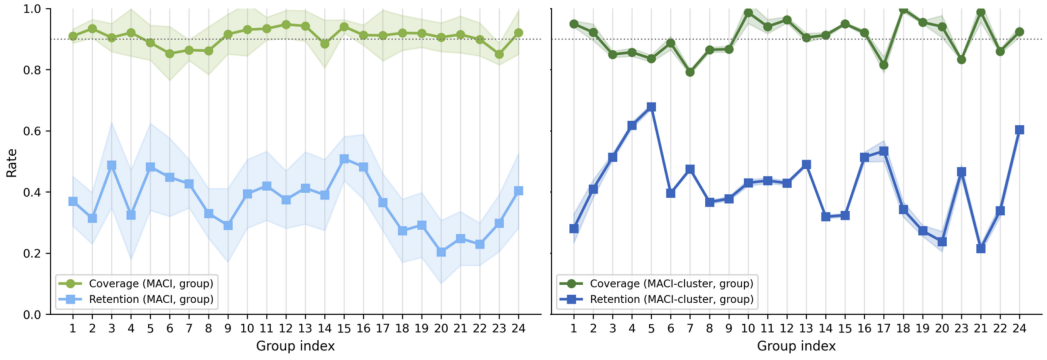
为定量评估这一差异，我们实现了 MultiValid 保形推断（MVCI），这是对 Jung 等人 (2023) 的 BatchMVP 的改进版本，专为虚假声明过滤场景设计，并将其与 MACI 进行比较。我们在 WikiBio 数据集上开展了实验，两种方法均采用相同的组定义（例如，基于虚假声明风险划分群组）、相同的校准/测试集划分以及相同的目标错误率 α。

表 5 展示了两种方法在组条件覆盖率和保留率方面的比较结果。MVCI 在每个群组中都达到目标覆盖率，但由于采用保守阈值，以同时保证不同群组和阈值水平下的覆盖率，其保留率显著较低。相比之下，在相同的组定义和显著性水平下，MACI 在维持相近组条件覆盖率的同时，实现了远高于 MVCI 的保留率。这体现了两者不同的设计理念：MVCI 优先保证强多重有效覆盖率，而 MACI 则着重在虚假声明过滤中平衡组条件保证与实际保留率。

表 5：与 MultiValid 保形推断 (MVCI) 的比较。覆盖率在 $1-\alpha \pm 0.01$ 范围内的值以绿色圆点标记 •，而超出此范围的值以红色箭头标记 ↓↑。

	目标覆盖率：80% ($\alpha = 0.2$)				目标覆盖率：90% ($\alpha = 0.1$)				目标覆盖率：95% ($\alpha = 0.05$)			
	MVCI		MACI		MVCI		MACI		MVCI		MACI	
组	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**
WikiBio：虚假声明风险												
Low	0.80•	0.11	0.82↑	**0.40**	0.89•	0.03	0.90•	**0.23**	0.95•	0.02	0.94•	**0.17**
中等	0.81•	0.08	0.81•	**0.42**	0.91•	0.02	0.90•	**0.25**	0.96•	0.01	0.95•	**0.12**
High	0.81•	0.03	0.81•	**0.45**	0.90•	0.01	0.90•	**0.28**	0.95•	0.01	0.96•	**0.09**

与基于采样的方法进行比较。虽然我们的主要关注点是基于 CI 的基线方法，但与近期非 CI 方法进行比较也很有实际意义。我们将 MACI 的组条件覆盖率、边际覆盖率和保留率与适用于黑盒大语言模型且不依赖检索的基于采样的方法进行比较。第D.5节简要介绍了这些基线方法。与基于 CI 的方法不同，基于采样的方法不提供统计保证；相反，它们会计算事实性评分 $p \in [0,1]$，从而能够通过



(a) MACI (b) MACI-Cluster

图 5：使用 MACI 和 MACI-cluster 方法时，群组数量较多情况下的覆盖率和保留率。上图显示了划分为 24 个群组后的结果。与 MACI-cluster 相比，MACI 更能维持各组的平均覆盖率，但方差也更大。使用组聚类方法可以通过牺牲严格的组条件覆盖率，获得实用的结果。

阈值（例如 0.5）。表6 显示，基于采样的方法具有较高的保留率，但覆盖率较低。这凸显了它们难以满足严格要求的局限性，即筛选集不含任何虚假声明。此外，它们的目标覆盖率并非由用户指定，因此无法预测。这些问题凸显了在需要用户指定较高 $1-\alpha$的高风险场景中使用 MACI 的必要性。

表 6：与基于采样的方法（一种具有代表性的黑盒、非检索虚假声明过滤方法）进行比较。基于采样的方法通常表现出极低或不稳定的覆盖率，以及较高的保留率。这表明它们不适用于「所有主张都必须是事实性内容」这一严格目标（**覆盖率**的定义）。相比之下，MACI（α = 0.1）能够可靠地保证用户所需的覆盖率。

	SelfCheck		FSC-text		FSC-KG		MACI	
组	**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	
医学内容								
Info	0.49	0.98	0.59	0.85	0.58	0.87	**0.90**	0.48
解读	0.54	0.98	0.59	0.90	0.63	0.93	**0.90**	0.47
行动	0.64	0.95	0.74	0.79	0.73	0.83	**0.90**	0.53
虚假声明风险								
Low	0.64	0.98	0.70	0.86	0.71	0.89	**0.89**	0.52
中等	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	
组	**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	
浏览量								
Low	0.11	0.95	0.42	0.69	0.46	0.64	**0.91**	0.21
中等 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
虚假声明风险								
Low	0.19	0.98	0.41	0.78	0.43	0.72	**0.90**	0.23
中等 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

应用组聚类方法。MACI 采用孟德尔式组条件符合方案：对于每个预定义的组 $g \in \mathcal{G}$，我们根据属于该组的校准样本分别校准一个阈值。令 R_g 表示组g内的非一致性分数分布，并令 $\hat{}$ $q_g^{1-\alpha}$表示 g中校准分数的经验$(1-\alpha)$分位数。将该组特定的分位数用作阈值，可在校准样本与测试样本之间满足群组条件可交换性的前提下，提供精确的有限样本组条件保证

$$\mathbb{P}(Y \in C_g(X) \,|\, Z = g) \;\geq\; 1-\alpha, \qquad \forall g \in \mathcal{G},$$

然而，当特定组的校准样本量 n_g 较小时，经验分位数 $\hat{}q_g^{1-\alpha}$可能具有较高方差，进而导致阈值不稳定，并使各组的覆盖率和保留率变得更加不稳定。

Gao 等人 （2025年）提出通过聚类分数分布相似的群组，并在聚类层级进行共形校准，以缓解这一问题。形式化地，令 $h:\ \mathcal{G} \to \mathcal{K}$ 将每个群组 g 映射到一个簇 $k = h\,(g)$，并令 R_k 表示将分配给簇 k的所有群组的校准分数合并后得到的分数分布。随后，在聚类空间中应用 Mondrian 共形推断，得到阈值 r_k^{α}，供所有满足 $g \;\; h\,(g) \;\; = k$共享。

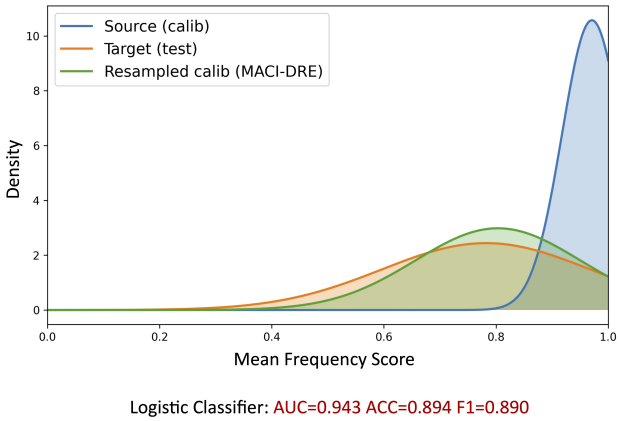


图 6：MedLFQA 数据集中的协变量偏移和重采样分布。蓝色和红色曲线分别显示基于协变量偏移变量平均频率分数的校准分布和测试分布。绿色曲线显示使用二元分类器重采样得到的分布。

这会增加每个阈值的有效校准样本量并降低方差。与此同时，组条件保证有所放宽：如果每个群组 g 的分数分布与其所属簇 $k = h(g)$ 的分数分布足够接近，即

$$\sup_{t\in\mathbb{R}} \left|F_g(t) - F_k(t)\right| \;\le\; \varepsilon_g,$$

where F_g 和 F_k 分别是 R_g 和 R_k的累积分布函数，那么组条件覆盖率会从
精确的 $1-\alpha$ 水平降至
m 的

$$\mathbb{P}\{Y \in C_k(X) \,|\, Z = g\} \;\ge\; 1 - \alpha - \varepsilon_g,$$

对所有 $g \in \mathcal{G}$而言，如 Gao 等人(2025年) 所示。换言之，组聚类以牺牲严格的有限样本组条件覆盖率为代价，换来了更低的方差和更稳定的阈值。

我们通过结合 Gao 等人的组聚类机制与 MACI 的事实性评分和阈值设定方案，构建了一个变体 MACI-Cluster。首先，我们使用虚假声明风险分数定义细粒度组，在符合度分数分布空间中进行组聚类，然后学习簇级阈值，并由每个簇内的组共享这些阈值。随后，我们从经验组条件覆盖率和保留率两方面，对比 MACI-Cluster 与原始 MACI 在许多组上的表现。具体而言，MACI-Cluster 使用通过对向量化的符合度分数直方图进行 k-均值得到的 $K = 8$ 个簇，并为每个簇应用一个保形阈值。

图 5 展示了在 24 种组划分下，MACI 和 MACI-Cluster 在各组上的覆盖率和保留率。在 100 次重复运行中，MACI 的组级覆盖率平均接近用户指定的水平，但方差很大，导致实际覆盖率不稳定。MACI-Cluster 的组级覆盖率并未恰好收敛到目标水平，这是由于簇与评估组之间存在分布差异；不过，其方差较小，因此当组数较多且每组样本较少时，覆盖率和保留率都显著更加稳定。因此，当组数增加且样本量较小时，组聚类方法是一种能够提供稳定结果的有效方法。

协变量偏移下的 MACI。用于拟合保形阈值的校准数据与已部署系统在测试时遇到的查询，不一定遵循相同的协变量分布。这种情形通常形式化为协变量偏移：校准数据 $\{(X_i, Y_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} \mathrm{P}_{XY}^{(s)}$，且测试点 $(X_{n+1}, Y_{n+1}) \sim \mathrm{P}_{XY}^{(t)}$ 满足 $\mathrm{P}_{Y|X}^{(s)} = \mathrm{P}_{Y|X}^{(t)}$，而它们的边际协变量分布 $\mathrm{P}_X^{(s)}$ 和 $\mathrm{P}_X^{(t)}$ 可能不同。令 $p_X^{(s)}$ 和 $p_X^{(t)}$ 分别表示 $\mathrm{P}_X^{(s)}$ 和 $\mathrm{P}_X^{(t)}$ 的密度（相对于一个共同的支配测度）。Tibshirani 等人（2019）对此类

失配通过使用密度比对校准样本进行重加权来处理

$$r(x) \;=\; \frac{p_X^{(t)}(x)}{p_X^{(s)}(x)}.$$

令 $S_i = S(X_i, Y_i)$ 表示源校准集上的非一致性分数。对于目标误覆盖水平 α，加权经验分位数定义如下

$$\widehat{q}_\alpha \;=\; \inf\Big\{q \;:\; \frac{\sum_{i=1}^n r(X_i)\,\mathbb{1}\{S_i > q\}}{\sum_{i=1}^n r(X_i)} \;\le\; \alpha\Big\},$$

它将常见的未加权经验尾概率替换为密度比加权版本，使其针对$\mathbf{P}_t$，而非$\mathbf{P}_s$。

我们将这一密度比校正原则应用于 MACI 流程，并将由此得到的变体称为 MACI-DRE。我们不修改 MACI 内部的分位数计算，而是先使用估计值 $\widehat{r}(x)$，通过重要性重采样构造密度比加权的校准集，然后在该重采样集上运行原始 MACI 算法。具体而言，给定来自 P_s 的校准样本$\{(X_i, Y_i)\}_{i=1}^n$ 和估计比率 $\widehat{r}(X_i)$，我们抽取索引

$$I_1, \ldots, I_n \;\sim\; \text{Multinomial}(n;\, \pi_1, \ldots, \pi_n), \qquad \pi_i \;=\; \frac{\widehat{r}(X_i)}{\sum_{j=1}^n \widehat{r}(X_j)},$$

并构造重采样校准集 $\{(X_{I_k}, Y_{I_k})\}_{k=1}^n$。对于任意有界可测函数 f,

$$\mathbb{E}\Big[\frac{1}{n}\sum_{k=1}^n f(X_{I_k}, Y_{I_k})\Big] \;=\; \sum_{i=1}^n \pi_i f(X_i, Y_i) \;\approx\; \mathbb{E}_{(X,Y)\sim P_t}\big[f(X,Y)\big] \quad \text{if} \quad \widehat{r}(x) \approx \frac{p_t(x)}{p_s(x)}.$$

因此，当密度比估计准确时，在重采样校准集上运行 MACI，与直接从目标分布 P_t 中抽取校准集并运行 MACI 在期望意义下是等价的，同时保持 MACI 的所有内部组件（集成权重学习、子组阈值估计）不变。

针对 MedLFQA 中的协变量偏移情形，我们在校准分布与测试分布之间构造了显式的协变量偏移。对于每个数据样本 $\{(P, R, C)\}$，我们通过 SelfCheck（Manakul 等人，2023）方法获取主张级事实性评分，并定义一个标量特征

$$s(x) \;=\; \frac{1}{m(x)} \sum_{j=1}^{m(x)} \hat{p}_{\text{SelfCheck}}(x, j),$$

其中，$m(x)$ 表示回答中的原子声明数量。我们按照 $s(x)$ 对所有样本进行排序，并将上部区域（较大的 $s(x)$，平均而言事实性更高的回答）指定为源测试池P_s，将下部区域（较小的 $s(x)$，更易出现幻觉的回答）指定为目标校准池 P_t。这样，在保持底层标注流程不变的同时，就 $s(x)$ 的分布而言产生了明显的边缘协变量偏移。

密度比是根据部署时可观测且无需真实标签的特征估计的。对于每个样本 x，我们构建一个特征向量

$$\phi(x) \;=\; \big(\bar{s}(x), \text{std}(s(x)), \text{len(prompt)}, \text{len(response)}, 1\big),$$

其中，$\bar{s}(x)$ 和 $\text{std}(s(x))$ 分别表示声明级 SelfCheck 评分的均值和标准差，其余分量则编码简单的提示词与回答长度统计。随后，我们在 $\phi(x)$ 上训练一个二元分类器，以区分源样本和目标样本，其中 P_s 的标签为 0，P_t 的标签为 1。使用平衡先验的逻辑回归，估计得到的密度比如下

$$\widehat{r}(x) \;=\; \frac{\hat{p}(Y = 1 \mid \phi(x))}{1 - \hat{p}(Y = 1 \mid \phi(x))},$$

随后将其用于上述重要性重采样步骤。

与一致性分数变体的比较。 MACI 的乘法分数源于预言机形式化。在理想预言机为每个声明分配联合事实性评分的情况下，将多个验证器的分数相乘，可以对该联合分数进行简单近似。为确保数值稳定性，我们采用以下对数乘积形式：

$$S_{\text{mult}}(c) = \sum_{i=1}^{N_i} \log p_i(c).$$

共形校准取决于一致性分数的排序，而非其绝对值，因此乘积形式与对数乘积形式之间存在单调变换，且二者提供相同的覆盖保证。

我们将该乘法分数与另外两种聚合规则进行比较。第一种是对数和形式，

$$S_{\text{log-sum}}(c) = \log\Big(\frac{1}{N_i}\sum_{i=1}^{N_i} p_i(c)\Big).$$

第二种是幂平均形式，

$$S_{\text{PM}}(c;\lambda) = \Big(\frac{1}{N_i}\sum_{i=1}^{N_i} p_i(c)^{\lambda}\Big)^{1/\lambda},$$

通过改变指数 λ，我们可以探索不同的聚合行为。

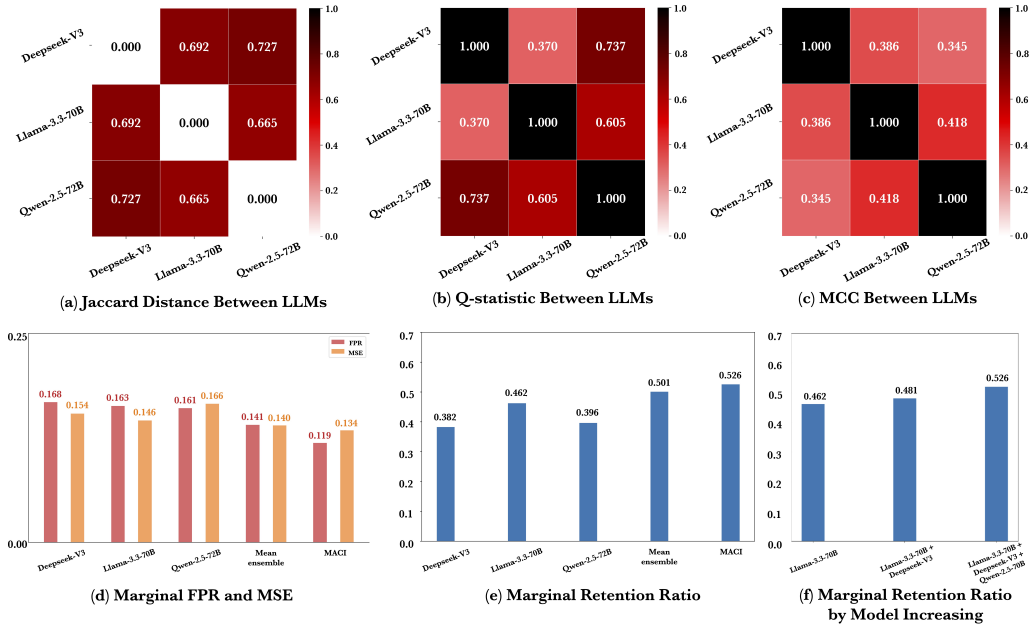
表 7 报告了这两种变体与 MACI 的覆盖率和保留率。总体而言，三种方法的覆盖率都接近目标覆盖率，但幂平均在各种设置下的保留率始终最低。对数和形式在低风险组和中等风险组中的保留率与 MACI 相当，但在高风险组以及 α较小时，其保留率明显下降。MACI 的对数乘积分数有明确的预言机公式依据，与共形校准框架兼容，并且在我们的实验中整体表现最佳，因此我们在正文中将其作为默认选择。基于求和的（对数和）聚合仍是一种有前景的替代方案，未来可通过更精细的设计进一步探索。

表7：在MedLFQA上，跨虚假声明风险组对一致性分数变体（幂平均（$\lambda = 2$）、对数和以及MACI的对数积）的比较。

	目标覆盖率：80%（$\alpha = 0.2$）						**目标覆盖率：90%**（$\alpha = 0.1$）						**目标覆盖率：95%**（$\alpha = 0.05$）					
	幂平均		对数和		MACI		幂平均		对数和		MACI		幂平均		对数和		MACI	
组	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**
MedLFQA：虚假声明风险																		
Low	0.81	0.67	0.81	0.75	0.79	**0.78**	0.91	0.42	0.90	**0.52**	0.89	**0.52**	0.96	0.29	0.95	0.33	0.95	**0.37**
中等	0.80	0.66	0.82	**0.70**	0.79	**0.70**	0.91	0.45	0.90	**0.46**	0.91	**0.46**	0.95	**0.31**	0.95	**0.31**	0.95	**0.31**
High	0.81	0.54	0.80	0.60	0.80	**0.64**	0.91	0.35	0.91	0.36	0.89	**0.41**	0.96	0.23	0.95	0.24	0.95	**0.26**

联合概率建模。 我们的主要实验设置将每个模型回答分解为相互独立的原子声明，然后仅使用{提示词 c_i} 对为每条声明估计事实性评分。这一设计遵循预言机公式，并支持在声明层面进行保形分析，同时隐含地采用了声明之间相互独立的假设。然而，在现实的大语言模型输出中，声明之间可能存在相关性、冗余或逻辑依赖关系，因此这一假设无法与底层生成过程完全吻合。

在我们的框架中，事实性检查交由用作验证器的 M 大语言模型负责，而不是在生成模型的单次解码过程中解决。从这一视角来看，一个重要的考量是：与明确呈现完整声明集的联合建模方法相比，逐声明建模策略是否仍然有效。为此，我们在逐声明设置之外，补充进行了一项联合建模实验：验证器一次接收整个声明集，并输出条件分数 $\{p(c_i \mid C)\}_{i=1}^{N_i}$。具体而言，我们将 {提示词 $\{c_1, \ldots, c_{N_i}\}$}作为输入，以便验证器原则上能够利用各项主张之间的相互作用来分配事实性评分。



(a) Jaccard Distance Between LLMs (b) Q-statistic Between LLMs (c) MCC Between LLMs

(d) Marginal FPR and MSE (e) Marginal Retention Ratio (f) Marginal Retention Ratio by Model Increasing

图 7：（a）–（c）展示了多样性诊断指标，包括杰卡德距离、Q统计量和马修斯相关系数；（d）–（e）展示了 MACI 的性能指标；（f）则说明了保留率随集成规模变化的趋势。（a）、（b）和（c）表明，大语言模型针对虚假声明的预测存在高度分歧且相关性较低，说明其检测模式各不相同，充分体现了采用集成方法的必要性。（d）展示了从单 LLM 和算术平均集成到我们提出的 MACI，其假阳性率和均方误差逐步改善；（e）则显示，随着这些错误率的改善，保留率也随之提高。（f）表明，随着集成中的模型数量增加（$k = 1$, 2、3），并选取最优组合，保留率持续提升，验证了扩展集成规模的效率。

表 8显示，逐声明设置与联合设置得出的覆盖率和保留率特征相近。这表明，即使验证器能够获取完整声明集，也能有效地将每项主张视为一个相对独立的决策单元；从验证器的角度来看，逐声明建模在实践中是可行的选择。换言之，我们在基于预言机的分析中采用的独立性假设，既是便于理论分析的简化，也是一种近似；与更偏向联合建模的策略相比，这种近似不会显著降低性能。

表 8：MACI 与 MACI-Joint 的比较。MACI-Joint 根据完整的声明集对主张进行联合评分，而 MACI 则独立地对每项主张评分。

	目标覆盖率：80%（$\alpha = 0.2$）				**目标覆盖率：90%**（$\alpha = 0.1$）				**目标覆盖率：95%**（$\alpha = 0.05$）			
	MACI		MACI-Joint		MACI		MACI-Joint		MACI		MACI-Joint	
组	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**	**Cov.**	**Ret.**
MedLFQA：虚假声明风险												
Low	0.79	**0.78**	0.80	0.75	0.89	**0.52**	0.91	**0.52**	0.95	**0.37**	0.95	0.36
中等	0.79	0.70	0.80	**0.74**	0.91	0.46	0.90	**0.47**	0.95	0.31	0.95	**0.34**
High	0.80	**0.64**	0.80	0.60	0.89	**0.41**	0.90	0.38	0.95	**0.26**	0.95	0.25

F 关于使用大语言模型的声明

我们仅使用大语言模型进行少量编辑和脚本自动化；核心思想、实验和分析均由作者完成。