# Agent-X 和 Fact-Audit 论文介绍

组会分享

LiuKai

2026.2.04

# 目录

# 本节内容

# Introduction

## 主要工具

本篇文章的创新点除了利用 Agent 协助评价之外，还提出了从语言学的角度去评判是否是 LLM 生成的文章。大致的思路就是先用 LLM 从多个维度独立的去评判是否是机器生成文本，同时还要生成评判意见，最后哦由一个 Meta Agent 来综合这些维度的评判结果，给出最终的结论。这样的检测方法的优点是:

- 无阈值，无需特定的数据集调试阈值
- 可以生成详细的评判意见

其中这篇论文还有一个创新点是，将常用的评判指标将 AUROC 换成 Accuracy。主要理由是: 论文的方法并不会输出一个模棱两可的分数，而是 Yes/No 的二分类结果，同时我们后面也会提到作者利用 Prompt 的校准工作。

## Guidelines

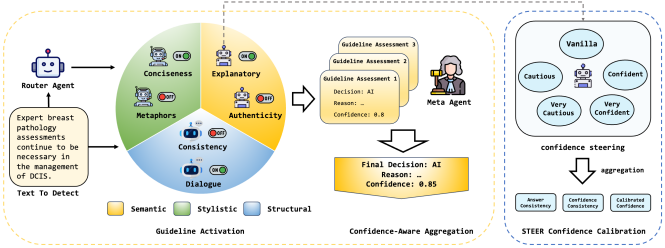
**语义维度**:
人类文本: 做出断言的时候直接，简洁，没有大量的修饰。
机器文本: 一般在做出声明时表现得谨慎、平衡或中立。
**文体维度**:
Precision and Conciseness
人类文本: 简明扼要，直接引入概念。
机器文本: 平衡、解释性强且措辞谨慎
**结构维度**:
人类文本: 句式结构与节奏多变机器文本: 结构统一可预测且高度一致

# Methodology

**Agent 协作机制：** 论文的方法主要是分为三个 Agent 协同工作:

- Router Agent：负责利用 LLM 分析每段输入文本，联合推断其主题领域（例如：医学、法律、文学）和文体属性（例如：正式语体、论证连贯性），然后激活最相关的 Guidelines Base Agent 进行评判，其他不相关的 Base Agent 则保持休眠。
- Base Agent：每个 Base Agent 都会根据 Router Agent 分发的 Guidelines, 对文本进行独立评判，输出二分类结果和评判意见。
- Meta Agent：负责收集所有 Base Agent 的评判结果和意见，并进行综合分析，最终输出文本是否为机器生成的结论。

例如一篇医学论文:Router Agent 会识别出其主题为医学，然后激活 “语义清晰度” 和” 结构精确性” 的 Base Agent 进行评判，最后 Meta Agent 会综合所有激活的 Base Agent 的评判结果，给出最终结论。



# Methodology

关于增加精度，作者运用了 Prompt 校准的技术，避免 LLM 在判断文本中过度自信的问题。在输入给 Base Agent 的 Prompt 中，有五个对称的的词：very cautious, cautious, vanilla, confident, and very confident。据此，Base Agent 要输出三个内容：

- 文本是否为机器生成的二分类结果（Yes/No）
- 对应的置信度等级（从 very cautious 到 very confident 五个等级）
- 评判意见

其中的计算公式是:

$$\kappa_{\text{ans}} = \frac{1}{|\mathcal{P}|} \max_{y \in \{\text{AI},\text{Human}\}} \sum_k \mathbb{I}[f_k(x) = y]$$

$$\mu_c = \frac{1}{|\mathcal{P}|} \sum_k c_k(x), \qquad \sigma_c = \sqrt{\frac{1}{|\mathcal{P}|} \sum_k \big(c_k(x) - \mu_c\big)^2}, \qquad \kappa_{\text{conf}} = \frac{1}{1 + \sigma_c/\mu_c}.$$

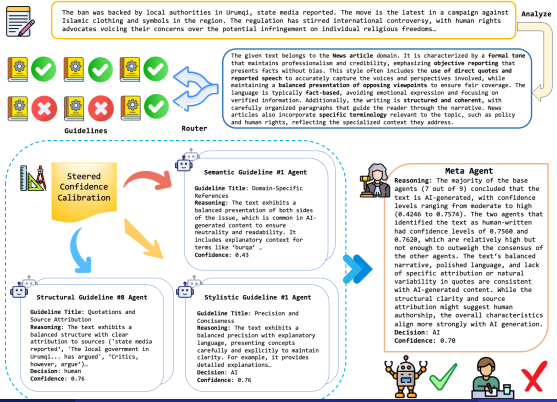
$$C_{\text{cal}}(x) = \mu_c \cdot \kappa_{\text{ans}} \cdot \kappa_{\text{conf}}.$$

$$k^* = \arg\min_k \left| c_k(x) - C_{\text{cal}}(x) \right|, \qquad f_{\text{final}}(x) = f_{k^*}(x).$$

$$c_{cal} = \text{平均分}(\mu_c) \times \text{答案是否一致}(\kappa_{ans}) \times \text{置信度的一致}(\kappa_{conf})$$

# Methodologies

而在 Meta Agent 中，如果一个智能体非常自信（且经过校准验证），而另一个智能体犹豫不决，Meta Agent 会更听从自信那个智能体的意见，还会“阅读”基础智能体写的理由。如果理由写得逻辑严密、证据确凿，该智能体的意见会被优先考虑。如果被激活的专家意见不一致，会结合语境评估并调和，分析为什么会有分歧，并试图达成一个基于共识的结论。最后生成评价意见，与 Base Agent 类似，Meta Agent 也会通过 Steering Conf 进行校准

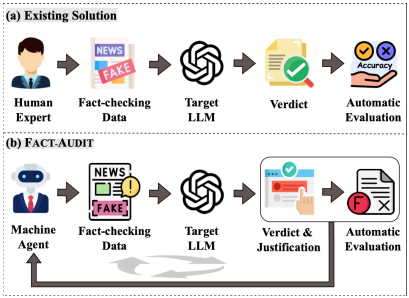


# 本节内容

# Introduction

这篇论文主要是提出了一个生成用于测试 LLM 检测事实能力的数据集，这里面使用了多个 LLM 进行协作。我们先介绍它创建的数据集与其他静态数据集的区别和优劣。首先是在流程上，原来的是由人类专家给出数据，然后给 LLM 进行检测，直接评判计算准确率。在 Fact-Audit 中，使用 Agent 生成数据，这样就节省了人类成本，然后还有对 LLM 生成的检测信息进行评价，最后把结果反馈给 Agent 进行迭代优化。原有检测方法的缺点是:

- 静态数据集，如果预先训练过就会不真实
- 需要人工专家
- 只关注结果而忽略过程



# Methodology

在 Methodology 上，作者解释原有的取样方法的弊端，原有的采样方法属于蒙特卡洛采样方法，在数学上十分低效，数学上的低效 ($O(1/\sqrt{N})$) 蒙特卡洛采样的收敛速度很慢。意味着你需要抽取海量的测试题（$N$ 要非常大），才能准确评估出一个模型到底行不行。如果题目太少，评估结果就不可靠。同时还有个最大的缺点: 大多数常见的、简单的知识。模型在训练阶段已经见得多了，随机抽样很容易抽到这些，导致评分虚高。因此作者调整了采样和评分，首先定义标准期望:

$$\mathbb{E}_{p(x)}[\mathcal{F}_\alpha(x)] = \int p(x)\mathcal{F}_\alpha(x)\mathrm{d}x$$

受重要性采样 (Importance Sampling) 的启发，引入提议分布 $q(x)$ 来提高效率，该过程调整为:

$$\mathbb{E}_{p(x)}[\mathcal{F}_\alpha(x)] = \int q(x)\mathcal{F}_\alpha(x)\frac{p(x)}{q(x)}\mathrm{d}x = \mathbb{E}_{q(x)}\left[\mathcal{F}_\alpha(x)\frac{p(x)}{q(x)}\right]$$

简单来说，就是引入 $q(x)$ , 来让难题有更高的采样概率，同时还调整了不同题目的分数，如果题目很难，我们可以看到，得分就会低，这样才能综合评估能力。

# Fact-Audit Algorithm

## Algorithm 1 Fact-Audit

初始化事实核查测试场景 $\Theta_0$，并设置记忆库 $\mathcal{M} = \varnothing$。
对于 $i := 0$ 到 $n$ 重复：
  将候选集 $\mathcal{X}$ 置空。
  **Stage 1: Prototype Emulation**
  当 $|\mathcal{X}| < k$ 时：
    Appraiser: $\theta_i \sim P(\Theta_i)$。
    Inquirer: $x \sim q(x|\theta_i)$。
    若 $x$ 通过质量检测，则并入 $\mathcal{X}$。
  **Stage 2: Fact Verification with Justification**
  $\mathcal{M} := \mathcal{F}_\alpha(\mathcal{X}) \dfrac{p(x)}{q(x|\Theta_i)}$。
  对于 $j := 0$ 到 $m$ 重复：
    Prober: $x \sim \rho(\mathcal{M})$。
    $\mathcal{M} := \mathcal{M} \cup \left\{ \mathcal{F}_\alpha(x) \dfrac{p(x)}{q(x|\theta_i)} \right\}$。
  **Stage 3: Adaptive Updating**
  更新 $\Theta_{i+1} \sim \pi(\Theta_{i+1}|\Theta_i, \mathcal{M})$。

## 备注

可以看到，Appraiser 根据当前参数 $\Theta_i$ 采样一个具体的测试方向 $\theta_i$ 然后 Inquirer 根据方向 $\theta_i$ 生成具体的问题 $x$，Quality Inspector 检查 $x$ 是否符合逻辑、是否有事实错误。只有合格的 $x$ 才会加入集合 $\mathbb{X}$。
$\mathcal{F}_\alpha(\mathbb{X})$ 是让被测模型回答这些问题。分式 $\frac{p(\mathbb{X})}{q(\mathbb{X}|\Theta_i)}$ 是重要性权重。
最后 Adaptive Updating.
$\Theta_{i+1} \sim \pi(\Theta_{i+1}|\Theta_i, \mathcal{M})$ 根据这一轮的测试结果，更新下一轮的生成参数 $\Theta$。

# Methodology

在出题具体实现上，文章实现了三个分工明确的智能体。

- Appraiser：主要负责出考纲，考纲分为三个经典类别，复杂主张 (Complex Claims)：需要多步推理的。假新闻 (Fake News)：故意误导的信息。社会谣言 (Social Rumors)：社交媒体上流传的传闻。
- Inquirer：根据给定的大纲出考题，考察方式也分三种：claim 模式只依靠 LLM 自身知识回答；evidence 模式引入维基百科证据检验推理；wisdom-of-crowds 模式提供模拟社交媒体评论考察群体智慧利用能力。
- Quality Inspector：负责“审题和校对”，确保题目质量与多样性。

Evaluator 则对 LLM 的回答进行评分，$\mathcal{F}_\alpha(x)$ 输出包含评分 $s$ 与评语 $c$。Prober 读取记忆库 $\mathcal{M}$，定位薄弱题目并重新设计试题。

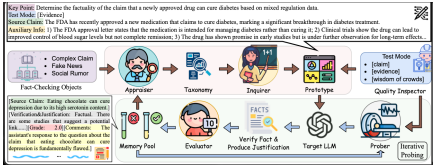



图: Fact-Audit 流水线

# Methodology

对于上述返回出来的结果，Appraiser 会针对记忆池 $\mathcal{M}$ 中低评分的题目进行分析，然后着重出这一方面的题目，也就是抓着 LLM 的弱点来出题，从而提高采样效率。而作者在 Metric 上设置了三个指标

- IMR: 代表低分事实核查响应占问题总数的比例。
- JFR: 目标 LLM 进行了正确的判决预测但提供的理由较差的案例百分比。
- Grade: 分数越高，说明模型的回答越准确、理由越充分。

# Limitations

这篇文章的局限性主要是它只负责找问题而不负责解决问题，它可以非常精准的发现 LLM 在事实核查方面的问题，但是并没有提供解决方案。未来可以考虑让框架不仅能审计模型，还能生成高质量的训练数据。这样，开发者就可以利用这些数据来微调模型，真正实现模型性能的提升，形成“评估-改进”的闭环。

# 本节内容

# 参考文献 I