Agent-X 和 Fact-Audit 论文介绍

组会分享

LiuKai

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

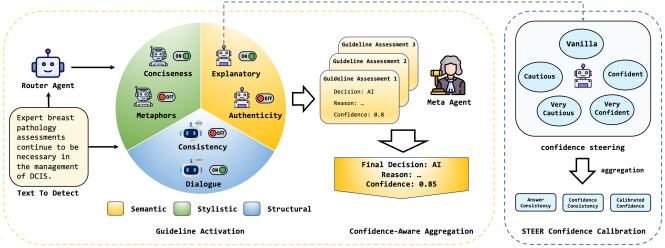
语义维度:
人类文本: 做出断言的时候直接，简洁，没有大量的修饰。
机器文本: 一般在做出声明时表现得谨慎、平衡或中立。
文体维度:
Precision and Conciseness
人类文本: 简明扼要，直接引入概念。
机器文本: 平衡、解释性强且措辞谨慎
结构维度:
人类文本: 句式结构与节奏多变机器文本: 结构统一可预测且高度一致

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:

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

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

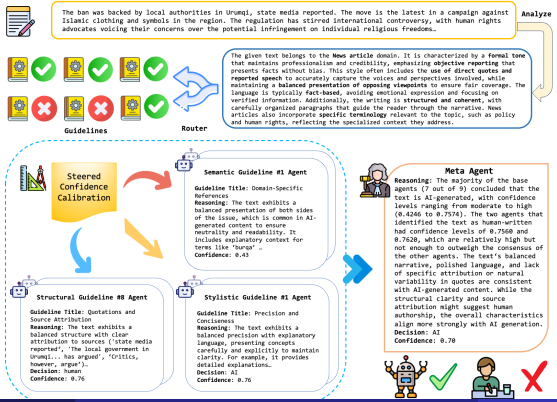
$$C_{\text{cal}}(x) = \mu_c \cdot \kappa_{\text{ans}} \cdot \kappa_{\text{conf}}.$$

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

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

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

本节内容

参考文献 I