

面向检索增强式私有问答大模型的幻觉检测方法

李铂鑫^{1,2}, 鲁 骁¹, 张 霄¹, 王 斌¹

(1. 北京小米移动软件有限公司 小米人工智能实验室, 北京 100085;
2. 中国科学院软件研究所 中文信息处理实验室, 北京 100190)

摘 要: 大模型幻觉现象的存在严重制约了其在实际场景中的落地, 而现有的幻觉检测方法单一, 缺乏面向检索增强生成范式下的幻觉检测工作。针对上述问题, 提出一种面向检索增强式私有问答大模型的幻觉检测方法, 该方法融合了基于不确定性度量指标和基于大模型自动评估两种幻觉检测方法, 综合利用了大模型生成过程中的幻觉特征和大模型自动评估能力。为提升开源大模型自动评估效果, 提出通过指令微调的方式将自动评估能力由闭源大模型蒸馏到开源大模型。此外, 构建了一个面向检索增强生成范式下的幻觉评估集以验证所提方法的有效性。在该数据集上的实验结果表明, 该融合方法取得了相对更高的AUC-ROC值, 相比于基于不确定性度量指标和基于大模型自动评估两种基线方法, 分别显著提升了11.1%和4.3%, 验证了上述两种基线方法具有互补性。此外, 指令微调后的开源大模型的AUC-ROC指标提升了18.6%, 自动评估能力得到显著提升, 验证了指令微调方法对提升开源大模型自动评估能力的有效性。

关键词: 大模型; 幻觉检测; 检索增强生成; 指令微调; 大模型自动评估

DOI: 10.11907/rjdk.241759

扫描二维码阅读全文:



中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2026)001-0039-08

A Hallucination Detection Method for Retrieval-Augmented Private Question-Answering Large Language Models

LI Boxin^{1,2}, LU Xiao¹, ZHANG Xiao¹, WANG Bin¹

(1. Xiaomi AI Lab, Beijing Xiaomi Mobile Software Co., Ltd., Beijing 100085, China;

2. Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The existence of Large Language Models' (LLMs) hallucination phenomenon seriously restricts its implementation in practical scenarios, existing hallucination detection methods were singular, and there was a lack of hallucination detection work under the paradigm of Retrieval-Augmented Generation (RAG). To address these issues, a hallucination detection method was proposed for private Question-Answering (Q&A) LLMs under the paradigm of RAG. This method integrated two types of hallucination detection methods, one based on uncertainty measure indexes and the other on LLMs' automatic evaluation, which effectively harnessed the hallucination features in the LLMs' generation process and the LLMs' automatic evaluation capability. To enhance the automatic assessment of open-source LLMs, a method of instruction fine-tuning was applied to distill the automatic evaluation ability from closed-source LLMs to open-source ones. Furthermore, a hallucination assessment dataset under the RAG's paradigm was constructed to verify the effectiveness of the proposed method. Experimental results on the above dataset indicate that the proposed integrated method achieved the highest AUC-ROC value, significantly increasing by 11.1% and 4.3% compared to the two baseline methods of uncertainty measure index and LLMs' automatic evaluation, respectively, proving the complementary nature of the two baseline methods. Moreover, the instruction fine-tuned open-source LLM exhibits an 18.6% increase in the AUC-ROC value, with a significant improvement in automatic evaluation capability, validating the effectiveness of the instruction fine-tuning method in enhancing the automatic evaluation ability of open-source LLMs.

Key Words: large language model; hallucination detection; retrieval-augmented generation; instruction fine-tuning; automatic evaluation of large language model

收稿日期: 2024-10-30

作者简介: 李铂鑫(1989-), 男, 博士, CCF会员, 北京小米移动软件有限公司小米人工智能实验室高级算法工程师, 研究方向为自然语言处理和深度学习; 鲁骁(1986-), 男, 博士, 北京小米移动软件有限公司小米人工智能实验室高级算法工程师, 研究方向为信息检索、推荐系统和自然语言处理; 张霄(1983-), 女, 硕士, 北京小米移动软件有限公司小米人工智能实验室高级算法工程师, 研究方向为深度学习、大模型和人工智能; 王斌(1972-), 男, 博士, 北京小米移动软件有限公司小米人工智能实验室NLP首席科学家、博士生导师, 研究方向为信息检索和自然语言处理。本文通讯作者: 李铂鑫。

0 引言

大模型幻觉指大模型 (Large Language Models, LLMs) 会生成貌似合理连贯, 但同输入的用户查询意图不一致、同世界知识不一致、与现实或已知数据不符合或无法验证的内容^[1]。幻觉现象的存在会给用户带来误导, 从而使 LLMs 生成结果的可靠性受到挑战, 严重制约了 LLMs 在实际场景中的落地。

本文以智能问答系统为例, 探索 LLMs 应用过程中幻觉现象的识别问题。智能问答系统在各企事业单位有广泛应用, 有望在以 ChatGPT 为代表的 LLMs 的赋能和加持下, 实现更大幅度的升级迭代。然而, 智能问答系统需要根据本地知识库作答, 考虑到内部的制度、政策、文档等内容一般不对外公开 (隐私保护需要), 此类问答系统属于限定领域的问答系统, 也称为私有问答场景。针对私有问答场景, 通用的 LLMs 落地方案是采用检索增强式 (Retrieval-Augmented Generation, RAG) 的开源 LLMs, 其基本思想是让开源 LLMs 根据召回的与用户查询相关的领域知识作答^[2]。鉴于此, 本文对 RAG 范式下开源 LLMs 的幻觉检测问题展开研究。

1 相关工作

为缓解幻觉带来的不利影响, 针对 LLMs 的幻觉检测工作日益成为学术界和工业界的一项重要研究课题。迄今为止, 研究人员从多个角度进行了探索, 提出了一系列幻觉检测方法, 包括基于不确定性度量指标的方法^[3]、基于自然语言推断 (Natural Language Inference, NLI) 模型的方法^[4]、构建幻觉分类器的方法^[5]、基于 LLMs 自动评估的方法^[6], 以及基于 LLMs 多次采样结果不一致性的方法^[7-8]等。

基于不确定性度量指标的方法基本思想是当 LLMs 对自己生成的模型回复不置信时, 大概率会出现幻觉现象^[3]。因此, 可考虑利用模型回复中各 token 的生成概率, 构建不确定性度量指标, 据此判断幻觉是否发生。根据所使用的 LLMs 是否开源, 又可划分为两种情况: 一是针对开源 LLMs, 可在获取模型回复的同时, 直接获取其中各 token 的生成概率, 该方法也称作灰盒 (grey-box) 方法; 二是针对闭源 LLMs, 无法获取模型回复中各 token 的生成概率, 可利用一个代理语言模型 (proxy Large Language Model, proxy LLM), 近似计算各 token 的生成概率, 也称作黑盒 (black-box) 方法。该类方法的不足是仅考虑了 LLMs 的生成特征, 而忽略了 LLMs 的自动评估能力对幻觉检测的影响。而本文方法综合考虑了上述两种特征, 有助于进一步提升幻觉检测效果。

基于自然语言推断模型的方法将幻觉识别任务转换

为 NLI 任务, 亦即判断检索到的知识库内容与模型回复之间是否存在“蕴含”“矛盾”“中性”的关系, 如果存在“矛盾”关系, 则表明发生了幻觉。该类方法的优点是简单易实现, 并且作为常用组件频繁出现在其他幻觉检测方法中。例如, Factcheck-GPT 中的实验结果表明, 在 MNLI 数据集上微调的模型 RoBERTa-large-mnli, 在做幻觉检测时效果要优于 zero-shot 下的 ChatGPT^[4,9-10]。此类方法存在两方面不足: ①所用自然语言推断数据集为通用语料, 在特定领域的实际场景中表现欠佳; ②没有考虑 LLMs 自动评估能力与生成特征, 从而限制了其幻觉检测效果。而本文方法充分利用了 LLMs 强大的自动评估能力, 无需训练语料, 同时使用了 LLMs 的生成特征, 幻觉识别能力更强。

构建幻觉分类器的方法, 其核心是标注幻觉数据集, 训练幻觉分类器。例如, Lee 等^[5]借鉴了虚假新闻检测中的方法, 即假新闻会表现出统计意义上的语言学特征, 针对 LLMs 的模型回复, 人工构建了 220 个统计语言学特征, 训练一个支持向量机 (Support Vector Machine, SVM) 分类器, 用于幻觉检测。该类方法的瓶颈是大规模高质量的标注数据成本高, 难以获取; 而本文方法无需标注数据, 充分利用了 LLMs 强大的自动评估能力。

利用 LLMs 自动评估的方法通过编写提示词 (prompt) 实现幻觉识别。例如, Guan 等^[6]利用 LLMs 进行自动化的事实检测, 用到的提示词如下所示。

用于 LLMs 自动幻觉评估的提示词示例:

(Task) Answer the following question:

(Input) Facts:

1. {e1}

2. {e2}

...

M. {eM}

Context: {c}

Statement following the context: {s}

(Question) Based on the given facts, is the statement correct? (A) Yes. (B) No. Please answer A or B:

其中, Context(c) 代表检索到的知识库内容, Statement(s) 表示模型回复, 而 {e1}...{eM} 指将 Context(c) 和 Statement(s) 拼接起来作 query, 从外部数据源 (例如谷歌等搜索引擎) 检索到的 M 个证据 (Evidence)。而本文方法除使用 LLMs 的自动评估能力外, 还融合了 LLMs 的生成特征, 幻觉检测效果更好。此外, 本文方法还对 LLMs 的自动评估能力进行了优化, 体现在两个方面: ①探索了细粒度的句子级别的幻觉检测, 有助于降低较长的模型回复对检测效果的干扰; ②利用指令微调的方式将闭源 LLMs 的强大评估能力蒸馏到开源 LLMs, 显著提升了开源 LLMs 的自动评估能力。

基于 LLMs 多次采样结果不一致性的方法假设如下: 如果 LLMs 对生成的答案具有较高的置信度, 那么对同一个用户查询, 采样多次生成的多个答案大概率语义一致。

此类方法的区别体现在两个方面: ① 采样方式不同; ② 多次采样结果不一致性的度量方式不同。SelfCheckGPT^[7]作为该类方法的代表, 通过调节采样方式 (beam search 采样^[11]、top-k 采样^[12]、top-p 采样^[13]等), 为同一个用户查询额外生成多个回复, 计算原始回复与新生成的多个回复之间的一致性, 以此作为原始回复是否发生幻觉的衡量指标。在计算模型结果不一致性方面, 有多种度量方法, 包括 BERTScore^[10]、基于统计语言模型的评分、基于 NLI 模型的评分, 以及基于 LLMs 自动评估的评分等。在 SelfCheckGPT 的基础上进行了拓展, 不同点主要体现在两个方面: ① 在采样方式上, 通过生成多个语义等价查询的方式 (采用复述生成技术, 通过编写提示词, 让大模型自动生成与原始查询语义相同的新查询), 产生多个输出, 并且每个查询只对应一个模型回复; ② 在不一致性度量方式上, 对生成的多个模型回复进行语义聚类, 计算每个类簇样本个数占比的熵并将其作为衡量指标, 熵越大, 不一致性越高。该类方法的一个显著缺点是在幻觉检测过程中需要多次调用 LLMs, 极大地增加了推理时延, 不利于其在实际场景中的落地, 而本文方法仅需额外多调用一次 LLMs。

综上所述可以看出, 以有的幻觉检测工作缺乏将不同幻觉检测方法相融合的探索, 且大多针对自由问答场景, 缺少 RAG 范式下的幻觉检测。针对上述不足, 本文提出面向检索增强式私有问答大模型的幻觉检测方法, 同时融合了基于不确定性度量指标和基于 LLMs 自动评估两种幻觉检测方法, 有助于增强幻觉检测效果。

2 本文方法

本文提出了一种面向检索增强式私有问答大模型的幻觉检测方法, 其工作流程如图 1 所示。

首先, 对于输入的用户查询“夫妻可以内推吗?”, 调用召回模块, 用于生成 top-k 个与用户查询相关的文档, 其中召回模块有多种实现方式, 本文采用当前比较流行的智源研究院通用向量模型 (BAAI General Embedding, BGE); 其次, 构建提示词, 输入给相应的 (原生) LLMs 和 (微调后的) 自评估 LLMs, 用于获取判断幻觉是否发生的不确定性得分 $score_1$ 和自评估得分 $score_2$; 最后, 将两者进行线性加权, 即可得到最终幻觉得分 $score_{final}$ 。

该方法同时融合了具有互补性的两种幻觉检测方法, 一是基于不确定性度量指标的方法, 二是基于 LLMs 自动评估的方法。背后动机如下: ① 基于不确定性度量指标的方法, 利用 LLMs 生成过程中各 token 的生成概率构建不确定性度量指标, 是衡量幻觉现象是否发生的一个关键特征^[3]; ② 大模型具有较强的反思能力^[14-17]和自动评估能力^[6, 18-19], 可以自动评价生成答案的质量, 类似人类做题后的自我检查。不难看出, 大模型生成与大模型反思是两个相对独立的过程, 将这两个过程对应的幻觉检测

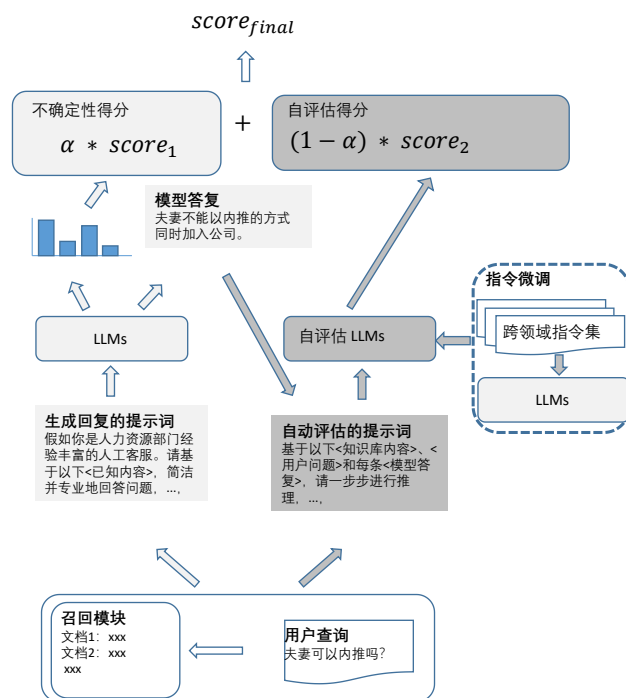


Fig. 1 The workflow of the method in this paper

图 1 本文方法工作流程

指标融合起来, 有助于增强幻觉检测效果。此外, 考虑到开源 LLMs 的指令理解能力与闭源 LLMs 相比存在较大差距, 本文提出通过指令微调的方式将自动评估能力由闭源 LLMs 蒸馏到开源 LLMs, 以进一步提升开源 LLMs 的幻觉识别能力。本文将详细阐述基于不确定性度量指标的方法和基于 LLMs 自动评估的方法, 并介绍两者的融合方式。

2.1 基于不确定性度量指标的方法

该类方法的基本思想已在相关工作部分提及, 根据所使用的 LLMs 是否开源, 可将其划分为灰盒方法和黑盒方法。其中, 灰盒方法直接利用开源 LLMs 模型回复中各 token 的生成概率, 而黑盒方法则针对闭源 LLMs 设计, 利用一个代理语言模型, 近似计算各 token 的生成概率。此类方法的不确定性度量指标计算方式又可分为两类, 一是基于对数概率, 二是基于熵。共有 4 种形式, 如下所示:

(1) 基于对数概率的不确定性度量指标, 如式 (1)、式 (2) 所示。

$$Avg(-\log p) = -\frac{1}{J} \sum_j \log(p_{ij}) \quad (1)$$

$$Max(-\log p) = \max_j (-\log(p_{ij})) \quad (2)$$

(2) 基于熵的不确定性度量指标, 如式 (3) 一式 (5) 所示。

$$Avg(H) = \frac{1}{J} \sum_j H_{ij} \quad (3)$$

$$Max(H) = \max_j (H_{ij}) \quad (4)$$

$$H_{ij} = -\sum_{\tilde{w} \in p_{ij}(\tilde{w})} p_{ij}(\tilde{w}) \log p_{ij}(\tilde{w}) \quad (5)$$

其中, H 为 Heat function 的首字母, 表示信息熵。

2.2 基于 LLMs 自动评估的方法

基于 LLMs 自动评估的方法通过编写提示词的方式实现幻觉识别。在已有工作中, 此类方法一般调用的是 ChatGPT、GPT4 等闭源 LLMs, 并不适用本文聚焦的私有问答场景。鉴于此, 本文采用开源 LLMs 作为自动评估器, 根据评估时的文本粒度和是否使用微调后的 LLMs, 共有以下 3 种实现方式:

(1) 段落级别评估。让 LLMs 自动评估检索到的知识库内容是否支持模型回复的生成, 用到的提示词如下所示。

LLMs prompting with passage 方法中的提示词示例:

前提: {context}

假设: {model_ans}

前提支持假设吗? 请回答"是"或"否"。

回答:

其中, context 指知识库内容, model_ans 表示模型回复, 根据 LLMs 的判断结果, 得出模型回复发生幻觉的概率。用 $L = \{L_1, L_2, L_3\}$ 表示 LLMs 的判断结果集合, $S = \{S_1, S_2, S_3\}$ 表示可能的幻觉概率集合, 其中 S_i 代表 L_i 对应的幻觉概率。存在如下映射关系: $\{L_1 = '是', S_1 = 0.0; L_2 = '否', S_2 = 1.0; L_3 = 'N/A', S_3 = 0.5\}$, 其中 "N/A" 表示 LLMs 输出异常时的情形。于是, 当 LLMs 的判断结果为 1 时, 可以计算出如式(6)所示的幻觉概率。

$$s = \sum_{i=1}^3 \mathbb{I}\{l = L_i\} S_i \quad (6)$$

其中, $\mathbb{I}\{l = L_i\}$ 为指示函数: 当 l 取值为 L_i 时, 其值等于 1; 否则, 其值等于 0。

(2) 句子级别评估。利用自然语言处理工具 spacy (<https://spacy.io/>), 将模型回复拆分成多个句子, 让 LLMs 自动判别知识库内容对每个句子的支持程度, 每个支持程度各对应一个幻觉得分, 最后取所有句子支持程度的均值作为最终幻觉得分, 用到的提示词如下所示。

LLMs prompting with sents 方法中的提示词示例:

基于以下<知识库内容>、<用户问题>和每条<模型答复>, 请一步步进行推理:

第一步: 详细对比每条<模型答复>与<知识库内容>, 识别出与<知识库内容>不一致的地方;

请注意, 这里所说的不一致, 指的是每条<模型答复>中出现而<知识库内容>中没有的内容、以及每条<模型答复>与<知识库内容>相违背的内容;

第二步: 根据第一步识别的"与知识库内容不一致的内容", 判断每条<模型答复>忠于<知识库内容>的程度, 并从下述列表中, 选择输出一个合适的程度: ["非常高", "高", "中等", "低", "非常低"]。

请输出每条<模型答复>的判断结果, 输出格式为 json, json 中的字段依次为: "忠诚度"。整体输出格式为 json 列表, 整体

json 的字段为: "整体结果"。

<知识库内容> {context}

{model_response_splits}

<用户问题> {user_query}

<输出结果> {json 格式}

其中, "model_response_splits" 表示拆分为多个句子的模型回复, 格式如下: "<模型回复 1>句子 1\n\n<模型回复 2>句子 2\n\n..."; "user_query" 代表用户查询, 即输入的用户问题。用 $L = \{L_1, L_2, L_3, L_4, L_5, L_6\}$ 表示所有的支持程度集合, $S = \{S_1, S_2, S_3, S_4, S_5, S_6\}$ 表示幻觉得分集合; 其中, S_i 代表 L_i 对应的幻觉得分, $L_6 = '其他'$ 表示 LLMs 输出异常时的情形。定义如下映射关系: $\{L_1 = '非常高', S_1 = 1.0; L_2 = '高', S_2 = 0.8; L_3 = '中等', S_3 = 0.5; L_4 = '低', S_4 = 0.3; L_5 = '非常低', S_5 = 0.1; L_6 = '其他', S_6 = 0.0\}$ 。则对于包含 k 个句子的模型回复, 可以得到一个支持程度列表 $l = \{l_1, \dots, l_k\}$, 其中 $l_j \in L$ 表示第 j 个句子对应的支持程度, 于是最终幻觉得分如式(7)所示。

$$s = \sum_{j=1}^k \sum_{i=1}^6 \mathbb{I}\{l_j = L_i\} S_i \quad (7)$$

这里, $\mathbb{I}\{l_j = L_i\}$ 为式(6)所示的指示函数; 需要注意的是, 作为超参数, 每个支持程度对应的幻觉得分 S_i 可通过网格搜索的方式在验证集上自动寻得局部最优值。

(3) 指令微调版的句子级别评估。在前期实验中, 发现开源 LLMs 的自动评估效果并不理想, 与闭源 LLMs (例如 ChatGPT、GPT4 等) 相比存在较大差距。一种有效的解决方案是通过指令微调的方式, 将自动评估能力由闭源 LLMs 蒸馏到开源 LLMs^[20-22]。于是, 本文构建了一个用于提升开源 LLMs 自动评估能力的指令集, 该指令集共有 490 条指令, 数据源取自公司内部店员自主问答项目, 包含手机和生态链商品的卖点、产品参数、售后维修等维度的数据。针对用户查询, 利用召回模型获取相关的知识库内容, 用于构建提示词, 其指令格式见上文 LLMs prompting with sents 方法中的提示词。

其中, 输出结果调用 GPT-4-32k (<https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4>) 接口生成。基于上述指令集, 利用 LoRA 微调方式对开源 LLMs 进行指令微调^[23], 指令微调过程中采用了 deepspeed 框架 (<https://www.deepspeed.ai>) 和两块 Nvidia A800 80G 的 GPU, 对应的超参数如表 1 所示。

Table 1 Hyperparameters for instruction fine-tuning

表 1 指令微调超参数

超参数	lora_r	lr	epoch	max_length
数值	64	2e-5	10	2048

2.3 融合方法

本文将基于不确定性度量指标的方法和基于 LLMs 自动评估的方法相融合, 具体通过线性加权的方式予以实现, 旨在进一步提升幻觉检测效果, 融合方式如式(8)所示。

$$score_{final} = \alpha * score_1 + (1 - \alpha) * score_2 \tag{8}$$

其中, α 为平衡因子, 介于 $[0, 1]$ 区间; $score_{final}$ 表示融合方法最终得分; $score_1$ 代表式(1)–式(4)所示不确定性度量指标中的一种, $score_2$ 表示式(6)或式(7)所示的 LLMs 自动评估得分, 并且都已经作了最小–最大值标准化处理。最小–最大值标准化处理的目的是使每种方法赋予每个样本的得分都处于 $[0, 1]$ 区间, 便于不同方法下的结果相融合, 具体实现如式(9)所示。

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}} \tag{9}$$

其中, $x = (x^{(1)}, \dots, x^{(i)}, \dots, x^{(m)})^T$ 表示某种幻觉检测方法关于 m 个样本的初始打分, 这里 x 表示 m 维的列向量, x_{min} 代表向量 x 的最小值, x_{max} 表示向量 x 的最大值。可以看出, 经过最小–最大值标准化处理后, 第 i 个样本的得分 $x_{norm}^{(i)} \in [0, 1]$ 。

3 实验与结果分析

3.1 数据集与评估指标

为了验证所提方法的有效性, 本文构建了一个面向人力资源 (Human Resource, HR) 领域问答场景的幻觉评估集。该评估集采自 HR 问答场景, 涉及考勤休假、离职/入职、五险一金、落户、薪酬、招聘、培训等多个业务, 所有样本的答案都由 LLMs 生成, 大模型答案 (模型回复) 是否正确, 由 HR 一线客服人员进行评估。本文将其中的错例都判为幻觉, 其余判为非幻觉, 构建的幻觉评估集名称为 HR–RAG–Hallu, 其样本统计信息如表 2 所示。

Table 2 Sample statistics of hallucination evaluation set HR–RAG–Hallu

表 2 幻觉评估集 HR–RAG–Hallu 的样本统计

类别	数量	平均答案长度	每个答案平均句子个数
总数	183	103.2	3.2
幻觉总数	78	103.9	3.1
非幻觉	105	102.7	3.3

可以看到, 幻觉和非幻觉样例的平均答案长度和每个答案的平均句子个数基本一致。在表 2 构建的幻觉评估集上, 提示词格式采用 RAG 范式下的标准模版, 由指令、已知内容 (知识库内容) 和问题 (用户查询) 3 部分组成, 具体格式如下所示。

RAG 范式下 HR 问答场景中的提示词示例:
假如你是人力资源部门经验丰富的人工客服。请基于以下<已知内容>, 简洁并专业地回答问题, 并严格遵守<规则>。
规则:
不允许在答案中添加编造成分。不允许在答案中作过多的解释。不允许在答案中提建议。如果无法从中得到答案, 请说“根据已知信息无法回答该问题”或“没有提供足够的

相关信息”。
另外, 答案请使用中文。
<已知内容:

{context}>
问题: {user_query}

此外, 本文还对 HR 问答场景中 RAG 范式下的幻觉现象进行了分类, 大致分为“完全漂移”和“部分漂移”两大类, “部分漂移”又细分为 5 个小类, 详情如表 3 所示。

Table 3 A comprehensive overview of hallucination phenomenon classifications in RAG paradigm within HR question answering scenarios
表 3 HR 问答场景中 RAG 范式下的幻觉现象分类情况一览

幻觉类别 (一级)	幻觉类别 (二级)
完全漂移: 模型基于自身学到的知识作答, 不过完全超出了<知识库内容>的范围, 即<模型回复>与<知识库内容>完全不相关	完全漂移
张冠李戴式: 指<模型回复>答非所问、错答的现象。示例:<用户查询>=“武汉怎么办理社保卡”, <模型回复>=FAQ 中“武汉社保卡开通金融账户”对应的答案	
忽视关键信息: 指的是<知识库内容>中有特殊说明, 而模型在回复时, 忽略了该重要信息。示例:<用户查询>=“南京婚假”, 标准答案=“南京婚假是 15 个连续自然日”, <模型回复>=“南京婚假为连续 10 个自然日。”, 忽略了 FAQ=“婚假期限”答案中的“当地另有规定的除外”	
部分漂移:<模型回复>与<知识库内容>部分相关, 但存在少量事实性错误或关键信息缺失、没有完全遵循指令等问题	部分漂移
缺少关键信息: 指<模型回复>中缺少了<知识库内容>中的重要信息, 例如超链接、联系人、联系方式等	
反问式: 指模型并没有直接回复<用户查询>, 而是追问了一句	
非标准拒绝话术: 指模型没有按照指令中的标准拒绝话术进行拒绝	

本文评估指标采用 AUC–ROC (Area Under the ROC Curve), 即 ROC 曲线下面积, 旨在衡量正样本得分高于负样本得分的概率。与正确率指标相比, AUC–ROC 值不依赖于样本类别分布且不受阈值的影响, 可以更准确地评价分类模型的优劣, 其计算方式如式 (10) 所示。

$$AUC - ROC = \frac{\sum_{i \in \{\text{正类样本}\}} rank_i - \frac{M(M+1)}{2}}{M \times N} \tag{10}$$

其中, M 表示正类样本数量, N 代表负类样本数量, 本文正类代表幻觉, 负类代表非幻觉; $rank_i$ 表示将所有样本按幻觉得分由低到高排序后, 样本 i 对应的序号, 此处得分越高, 序号越大。

3.2 对比方法

本文使用的基准大模型来自于 Baichuan 系列 (<https://github.com/baichuan-inc/Baichuan2>), Chat 版模型使用的是 Baichuan2–13B–Chat, 代理语言模型使用的是 Baichuan2–

13B-Base。采用的对比方法如下:

(1) Baichuan2-13B-Chat's probabilities (LLM, grey-box)^[3]。基于不确定性的度量方法,此处直接使用 LLMs 的生成概率,该方法的不确定性计算方式如式(1)一式(4)所示。

(2) Baichuan2-13B-Base's probabilities (Proxy LLM, black-box)^[3]。基于不确定性的度量方法,此处使用代理语言模型的生成概率,该方法的不确定性计算方式如式(1)一式(4)所示。

(3)基于自然语言推断模型的方法^[4]。简称“NLI”,将幻觉识别(尤其是 RAG 模式下的)转换为自然语言推断任务,即判断检索到的知识库内容(前提)与模型回复(假设)之间是否存在“蕴含”“矛盾”“中性”的关系。如果存在“矛盾”关系,则表明发生了“幻觉”。需要注意的是,考虑到模型回复一般是由若干个句子构成的段落,本文基于 NLI 的方法探索了两种实现方式:①一次性推断,将整个模型回复当作前提,判断知识库内容与模型回复之间发生“矛盾”的概率,对应方法名为“NLI with passage”;②多次推断,将模型回复拆分成多个句子,逐一判断知识库内容与模型回复中每个句子之间发生“矛盾”的概率,再通过取平均的方式计算最终结果,对应方法名为“NLI with sents”。

(4) LLMs prompting^[6]。对应 2.2 节中的 3 种实现方式:①段落级别评估,类似于 NLI with passage 方法,不同之处是无需标注数据,对应的方法名为“LLMs prompting with passage”;②句子级别评估,对应的方法名为“LLMs prompting with sents”;③句子级别评估,但使用了指令微调后 Chat 版的大模型,对应的方法名为“LLMs prompting with sents (sft)”。

(5)融合方法。利用式(8)将基于不确定性度量指标和基于 LLMs 自动评估两种方式进行线性加权,对应的方法名为“Baichuan2-13B-Chat's probabilities with Avg(H)+LLMs prompting with sents (sft)”。

3.3 实验结果与分析

表 4 展示了不同的大模型幻觉检测方法在评估集 HR-RAG-Hallu 上的实验结果,其中,最优值加粗表示、次优值加下划线,出于篇幅方面的考虑,融合方法仅展示 top-5 的结果。下文分别分析单模型效果和融合方法的效果。

(1)单模型效果分析。针对单一模型,方法 LLMs prompting with sents(sft)效果最好,AUC-ROC 值为 78.4%。基于不确定性度量指标的方法,无论是灰盒方法还是黑盒方法,通过 Avg(H)方式计算不确定性的效果最好,并且两者结果很接近,分别为 71.6%和 71.5%,都显著高于其他 3 种方式。此时,在 RAG 范式下的私有问答场景中,灰盒方法直接使用了 Chat 版 LLMs 的生成概率,在响应时间方面占据优势:省去了黑盒方法调用代理语言模型获取生成概率的时间。此外,基于对数概率的不确定性度量指标的 3

Table 4 Experimental results on hallucination evaluation set HR-RAG-Hallu

表 4 幻觉评估集 HR-RAG-Hallu 上的实验结果

方法类型	方法名称	AUC-ROC/%
Baichuan2-13B-Chat's probabilities (LLM, grey-box)	Max(-logp)	37.5
	Avg(-logp)	28.4
	Max(H)	59.1
	Avg(H)	71.6
Baichuan2-13B-Base's probabilities (Proxy LLM, black-box)	Max(-logp)	54.5
	Avg(-logp)	38.4
	Max(H)	64.6
	Avg(H)	71.5
NLI	w/passage	75.6
	w/sents	71.1
LLMs prompting (Baichuan2-13B-Chat)	w/passage	50.8
	w/sents	59.8
	w/sents (sft)	78.4
	Baichuan2-13B-Chat's probabilities with Avg(H)+LLMs prompting with sents (sft)	82.7
融合方法	Baichuan2-13B-Base's probabilities with Avg(H)+LLMs prompting with sents (sft)	82.6
	NLI with passage + LLMs prompting with sents (sft)	80.9
	Baichuan2-13B-Chat's probabilities with Max(H)+LLMs prompting with sents (sft)	79.5
	NLI with sents + LLMs prompting with sents (sft)	79.0

种方法: Baichuan2-13B-Chat's probabilities with Max(-logp)、Baichuan2-13B-Chat's probabilities with Avg(-logp)和 Baichuan2-13B-Base's probabilities with Avg(-logp),其 AUC-ROC 值均小于 0.5,效果甚至不及随机猜测。

直接基于 LLMs 自动评估方法的 AUC-ROC 值分别为 50.8% (LLMs prompting with passage) 和 59.8% (LLMs prompting with sents),仅比随机猜测略好一点,反映出开源 LLMs 在自动评估方面的局限性。一个重要原因是,开源 LLMs 的指令理解能力存在局限性,对于其较复杂的指令,会出现“输出的评估结果与模型答复数量不一致”“输出结果是非法的 json 格式而无法正确解析”等问题。不过,在仅使用少量指令进行有监督微调后,句子级别评估效果得到显著提升,AUC-ROC 值增加了 18.6%,达到 78.4%,验证了指令微调方法对提升开源 LLMs 自动评估能力方面的有效性。

(2)融合模型效果分析。基于不确定性度量指标与基于 LLMs 自动评估(微调后)相融合的方法取得了更佳效果,AUC-ROC 值为 82.7,相较于用于融合的两种基线方法 Baichuan2-13B-Chat's probabilities with Avg(H)和 LLMs prompting with sents(sft)分别提升了 11.1%和 4.3%,提升幅度显著,充分验证了本文想法,即基于不确定性度量指标的幻觉检测方法和基于 LLMs 自动评估的幻觉检测方法具有互补性。其次是融合方法 Baichuan2-13B-Base's proba-

bilities with Avg (H) + LLMs prompting with sents (sft) , AUC-ROC 值为 82.6,缺点是会额外调用代理语言模型,从而导致响应时间增加,具体而言,代理语言模型平均时延增加了 3.9%,即由 2.05 s 上升到了 2.13 s。

图 2 显示了式 (8) 中平衡因子 α 对实验效果的影响,其中 $score_1$ 和 $score_2$ 分别对应基线方法 LLMs prompting with sents (sft) 和 Baichuan2-13B-Chat’s probabilities with Avg(H) 的 AUC-ROC 得分。当 α 分别取 1.0 和 0.0 时, AUC-ROC 得分相对较低,分别为 78.4 和 71.6,即单独使用基线方法 LLMs prompting with sents(sft) 和 Baichuan2-13B-Chat’s probabilities with Avg(H) 时的得分。当 α 取值介于 0.3~0.9 时, AUC-ROC 值稳定地居于 81.6~82.7 之间,一方面验证了上述两种方法进行融合的有效性,另一方面也反映出本文融合方法的鲁棒性,即在较大程度上对 α 不敏感。需要注意的是,当 α 取值较小时(小于 0.2), 融合方法虽然高于基线方法 Baichuan2-13B-Chat’s probabilities with Avg(H), 但略低于另一个基线方法 LLMs prompting with sents (sft)。一个重要原因是基线方法 Baichuan2-13B-Chat’s probabilities with Avg(H) 的得分偏低, 当 α 取值较小时(小于 0.2), 融合方法的最终得分中 Baichuan2-13B-Chat’s probabilities with Avg(H) 方法的得分所占比重较高,从而影响了融合效果。

表 5 展示了幻觉评估集 HR-RAG-Hallu 上的两个样例,最后一列对应本文融合方法给出的幻觉得分。

可以看出,输入给 LLMs 的召回文档并未包含这两个

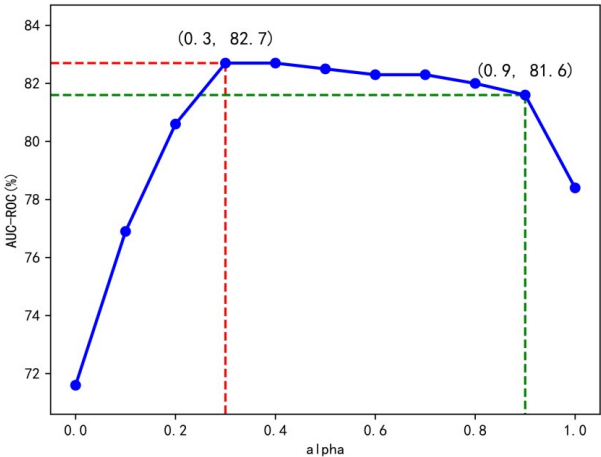


Fig. 2 The influence of the balance factor α in the fusion method
图 2 融合方法中平衡因子 α 的影响

用户查询的答案,根据 RAG 范式下 HR 问答场景中的提示词, LLMs 本应回复“根据已知信息无法回答该问题”,但 LLMs 却没有遵从指令,而是都作出了回答。其中,第一个用户查询的标准答案是“可以内推,但不享受奖金”,而 LLMs 却错误地回复“夫妻不能以内推的方式同时加入公司”,给用户造成了误导;针对第二个用户查询, LLMs 的回复超出了召回文档的范围,属于幻觉分类中的“完全漂移”,细看其实是一些无用的套话,会降低用户体验。而本文所提融合方法给上述两个样例打出较高的幻觉得分,可通过卡阈值的方式对其进行过滤,从而降低幻觉风险。

Table 5 Case analysis of hallucination detection
表 5 幻觉检测案例分析

序号	用户查询	大模型答案	召回文档	幻觉得分
1	夫妻可以内推吗	夫妻不能以内推的方式同时加入公司	文档 1->问:内推具体流程 答:您好,您通过内推网站推荐候选人后,负责相关岗位的 HR 会转给业务部门进行简历筛选,简历筛选过后 HR 会跟该候选人联系安排面试 文档 2->问:怎么算内推成功 答:认定方式: (1)当候选人被多名内推人或职位推荐时,认定其最终入职职位的最近一次推荐记录为有效推荐 (2)当候选人被多名内推人或职位推荐,均未成功入职所推荐职位,一年内由招聘人员调配至精准职位并成功入职时,认定最近一次推荐记录为有效推荐 (3)当候选人未成功入职内推人所推荐职位,且未被其他内推人推荐或通过其他渠道被推荐,一年内入职公司其他职位时,仍认定内推人的推荐为有效推荐	0.75
2	您好,我上班途中发生了交通事故,趾骨骨折,请问可以申请什么保险?	您好,根据您的已知信息,您需要了解的是关于上班途中发生的交通事故是否可以申请保险以及可以申请哪些保险。首先,我们需要明确您的公司是否为员工提供了商业补充保险。其次,我们需要了解公司的商业补充保险是否涵盖了意外伤害和医疗费用的保障。如果您公司的商业补充保险确实包含了这些保障,那么您可以在符合保险合同条款的情况下申请理赔。具体的理赔流程和所需材料,您可以参考您公司的商业补充保险的相关文件和指导。希望这些信息对您有所帮助	文档 1->问:理赔申请单 答:您好,理赔申请单下载路径:<company app>-在线文档-人事类-补充商业保险 文档 2->问:金额大于 1 万元,怎么申请商保理赔 答:您好,不同的就医场景所需材料不同,可查询补充商保相关-业务指南 中 2.1.2 查找具体材料。将所有材料提交给保险公司后,按照其指示进行后续流程。理赔申请单模板请查询: <url> 文档 3->问:试用期可以报销商保吗 答:您好,试用期期间可以正常进行商保理赔。具体流程可以参考<company app>-在线文档-人事类-补充商业保险	0.88

4 结语

本文研究了检索增强范式下开源大模型的幻觉检测问题,提出了一种面向检索增强式私有问答大模型的幻觉检测方法,该方法综合运用了大模型的生成特征和自动评估能力,在构建的基于 HR 智能问答领域的幻觉评估集 HR-RAG-Hallu 上,相较于两种用于融合的基线方法, AUC-ROC 指标分别显著提升了 11.1% 和 4.3%。

但同时,本文研究尚存在一些不足之处,例如未考虑领域内的知识对幻觉检测的帮助、对大模型的自身特征挖掘得不够充分等。未来工作中,可探索结合领域内知识图谱的细粒度幻觉检测,并研究大模型内部状态向量对幻觉检测的影响。

参考文献:

- [1] ZHANG Y, LI Y, CUI L, et al. Siren's song in the AI ocean: a survey on hallucination in large language models [DB/OL]. <https://arxiv.org/abs/2309.01219>.
- [2] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks [C]// Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020: 9459-9474.
- [3] ZHANG T, QIU L, GUO Q, et al. Enhancing uncertainty-based hallucination detection with stronger focus [C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 915-932.
- [4] WANG Y, REDDY R G, MUJAHID Z M, et al. Factcheck-gpt: end-to-end fine-grained document-level fact-checking and correction of llm output [DB/OL]. <https://arxiv.org/abs/2311.09000>.
- [5] LEE B W, AROCKIARAJ B F, JIN H. Linguistic properties of truthful response [C]// Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), 2023: 135-140.
- [6] GUAN J, DODGE J, WADDEN D, et al. Language models hallucinate, but may excel at fact verification [C]// Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024: 1090-1111.
- [7] MANAKUL P, LIUSIE A, GALES M. Selfcheckgpt: zero-resource black-box hallucination detection for generative large language models [C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 9004-9017.
- [8] ZHAO Y, YAN L, SUN W, et al. Knowing what llms do not know: a simple yet effective self-detection method [C]// Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024: 7044-7056.
- [9] WILLIAMS A, NANGIA N, BOWMAN S. A broad-coverage challenge corpus for sentence understanding through inference [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 1112-1122.
- [10] LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized bert pre-training approach [DB/OL]. <https://arxiv.org/abs/1907.11692>.
- [11] HOKAMP C, LIU Q. Lexically constrained decoding for sequence generation using grid beam search [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1535-1546.
- [12] FAN A, LEWIS M, DAUPHIN Y. Hierarchical neural story generation [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 889-898.
- [13] HOLTZMAN A, BUYS J, DU L, et al. The curious case of neural text de-generation [DB/OL]. <https://arxiv.org/abs/1904.09751>, 2020.
- [14] ASAI A, WU Z, WANG Y, et al. Self-rag: learning to retrieve, generate, and critique through self-reflection [DB/OL]. <https://arxiv.org/abs/2310.11511>, 2024.
- [15] MADAAN A, TANDON N, GUPTA P, et al. Self-refine: iterative refinement with self-feedback [C]// Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023: 46534-46594.
- [16] SHINN N, CASSANO F, GOPINATH A, et al. Reflexion: language agents with verbal reinforcement learning [C]// Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023: 8634-8652.
- [17] HUANG J, GU S, HOU L, et al. Large language models can self-improve [C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 1051-1068.
- [18] WANG P, LI L, CHEN L, et al. Large language models are not fair evaluators [C]// Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024: 9440-9450.
- [19] ZHENG L, CHIANG W L, SHENG Y, et al. Judging llm-as-a-judge with mt-bench and chatbot arena [C]// Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023: 46595-46623.
- [20] WANG Y, KORDI Y, MISHRA S, et al. Self-instruct: aligning language models with self-generated instructions [C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 13484-13508.
- [21] JIANG Y, CHAN C, CHEN M, et al. Lion: adversarial distillation of proprietary large language models [C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023: 3134-3154.
- [22] XU C, SUN Q, ZHENG K, et al. Wizardlm: empowering large language models to follow complex instructions [DB/OL]. <https://arxiv.org/abs/2304.12244>, 2024.
- [23] HU E J, WALLIS P, ALLEN-ZHU Z, et al. Lora: low-rank adaptation of large language models [DB/OL]. <https://arxiv.org/abs/2106.09685>, 2022.

(责任编辑:孙娟)