计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2025/19(05)-1295-07
doi: 10.3778/j.issn.1673-9418.2408080

大语言模型幻觉现象的分类识别与优化研究

何　静[1],沈　阳[2+],谢润锋[3]
1. 北京航空航天大学 人文与社会科学高等研究院,北京 100191
2. 清华大学 新闻与传播学院,北京 100084
3. 北京工业大学 信息学部,北京 100124
+ 通信作者 E-mail: 287773664@qq.com

摘　要:随着大语言模型在自然语言理解和生成任务上的广泛应用,其在医疗、法律和科研等高精度领域的表现被愈发关注。然而,幻觉现象作为大语言模型普遍存在的问题,极大制约了其在这些领域的实际应用。当前,针对大语言模型幻觉现象的评估和优化尚存在显著不足:缺乏高质量的高精度领域幻觉评估数据集;现有幻觉评估方法大多依赖单一模型,未能充分利用多模型间的差异性优势;不同模型在幻觉类型和幻觉率上表现存在较大差异,尚未有有效方法来降低高幻觉率模型的幻觉现象。该研究采用数据集构建-群体智能选举-幻觉分类与量化-先验知识优化的系统流程,全面评估和优化了大语言模型在医疗问答领域的幻觉现象。根据公开数据集Huatuo,结合GPT4生成问题答案和人工标注的形式构建了医疗问答领域大模型幻觉评估数据集;使用GPT4o、GPT4、ChatGLM4、Baichuan-13B和Claude 3.5等先进的大语言模型对数据集中的问题生成答案。通过一种基于群体智能的方法,选举出一个LeaderAI,它将各模型的回答与参考答案进行比较,从而确定各模型的幻觉率。进一步将幻觉分为事实性幻觉和忠实性幻觉两类。研究结果表明,在LeaderAI的指导下,被评估的大模型的幻觉率显著下降,特别是忠实性幻觉率明显降低。

关键词:大语言模型;幻觉识别;幻觉分类;模型优化
文献标志码:A　**中图分类号:**G206;TP18

Research on Categorical Recognition and Optimization of Hallucination Phenomenon in Large Language Models

HE Jing[1], SHEN Yang[2+], XIE Runfeng[3]
1. Institute for Advanced Studies in Humanities and Social Sciences, Beihang University, Beijing 100191, China
2. School of Journalism and Communication, Tsinghua University, Beijing 100084, China
3. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

Abstract: With the widespread application of big language models in natural language understanding and generation tasks, their performance in high-precision fields such as healthcare, law, and scientific research has received increasing attention. However, the phenomenon of hallucinations, as a common problem in large language models, greatly restricts their practical application in these fields. At present, there are significant shortcomings in the evaluation and optimization of hallucination phenomena in large language models. Firstly, there is a lack of high-quality and high-precision domain hallucination evaluation datasets. Secondly, most of the existing hallucination assessment methods rely on a single model, which fails to take full advantage of the differences between multiple models. Finally, there are significant differences in the performance of different models in terms of hallucination types and rates, and there is currently no effective method to

基金项目:国家自然科学基金(62406016);青少年心理健康与危机智能干预安徽省哲学社会科学重点实验室项目(SYS2023A07)。
This work was supported by the National Natural Science Foundation of China (62406016), and the Project of Youth Mental Health and Crisis Intelligent Intervention Anhui Provincial Key Laboratory of Philosophy and Social Sciences (SYS2023A07).
收稿日期:2024-08-22　**修回日期:**2025-01-03

reduce the hallucination phenomenon in high hallucination rate models. This paper adopts a systematic process of dataset construction, swarm intelligence election, hallucination classification and quantification, and prior knowledge optimization to comprehensively evaluate and optimize the hallucination phenomenon of large language models in the field of medical question answering. Firstly, based on the publicly available dataset Huatuo, a large model illusion evaluation dataset in the medical question answering field is constructed by combining GPT generated question answers and manual annotation. Secondly, advanced big language models such as GPT4o, GPT4, ChatGLM4, Baichuan-13B, and Claude 3.5 are used to generate answers to questions in the dataset. By using a swarm intelligence based method, a LeaderAI is elected, which compares the answers of each model with reference answers to determine the illusion rate of each model. Finally, hallucinations are further divided into two categories: factual hallucinations and fidelity hallucinations. The research results indicate that under the guidance of LeaderAI, the illusion rate of the evaluated large models significantly decreases, especially the fidelity illusion rate.

Key words: large language model; hallucination recognition; hallucination classification; model optimization

2024年1月，OpenAI的Sora像一颗炸弹，再一次引爆全球。早在2023年，以大语言模型（large language model，LLM）为代表的人工智能生成技术（artificial intelligence generated content，AIGC）的涌现，已具备出色的学习能力和泛化能力，能够依据所处的错综复杂的环境自主生成并顺利执行各项任务，并可以基于喂养数据进行推理、学习、创造、交流等多种智能活动，但不可避免具备幻觉现象。正如Zhang等人[1]提到的，这些大模型中存在一个普遍的问题，大模型会生成看似合理，但其实偏离了用户意图、偏离之前所生成的上下文或偏离了事实知识的回答，即大模型回答中存在幻觉。这一现象不仅影响了模型输出的可靠性和安全性，更制约了其在对准确性和一致性要求极高领域（如医疗、法律和科学研究等）中的应用。在此背景下，评估和减缓大模型幻觉至关重要。

群体智能（crowd intelligence）方法借鉴了人类社会中集体决策和智慧群体的概念，旨在通过集成多专家的答案来提高整体准确性。在这种方法中，每个专家的答案都会作为群体成员的“投票”，并通过一个专家（leader）进行最终的答案决策。这种方法不仅可以有效降低单个专家在回答问题时产生的幻觉率，还可以利用不同专家的优势互补，从而提高整体答案的可靠性。

本文在研究中引入群体智能方法，通过选举一个LeaderAI，将多模型生成的答案与参考答案进行对比，综合评估各个模型的幻觉率及其分类幻觉率的情况。本文将幻觉分为两类：事实性幻觉，指生成内容与已知事实不符或存在事实捏造；忠实性幻觉，指生成内容未忠实于指令或上下文，包括指令不一致、上下文不一致和逻辑不一致。其具体区别如表1所示。LeaderAI在此过程中发挥了关键作用，不仅能准确识别模型输出中的幻觉现象，还能指导其他模型优化生成过程。实验结果表明，LeaderAI方法能显著降低大语言模型的幻觉率，尤其是对忠实性幻觉的优化效果更为明显。

表1 大模型幻觉分类

Table 1 Hallucination classification of LLMs

提问：糖尿病患者可以通过吃蜂蜜代替糖吗？		
幻觉类型	事实性幻觉	忠实性幻觉
回答内容	是的，蜂蜜是天然的，可以帮助糖尿病患者稳定血糖水平	蜂蜜富含维生素和矿物质，对提高免疫力很有帮助，因此是一种健康的食品
问题分析	错误：蜂蜜虽然是天然食品，但仍然含有大量果糖和葡萄糖，会升高血糖水平，不适合糖尿病患者代替糖使用	偏题：回答内容虽无事实错误，但与提问“糖尿病患者是否可以用蜂蜜代替糖”无关，未忠实于用户意图

本文贡献可以总结为以下三点：第一，基于公开的Huatuo数据集，通过GPT4o结合人工标注的方式构建了高质量的医疗问答领域大模型幻觉评估数据集。第二，根据所构建的幻觉评估数据集设计Prompt测试当前先进LLM的幻觉率及事实性和忠实性幻觉率，包括GPT4o、GPT4、ChatGLM4、Baichuan-13B和Claude 3.5。第三，在实验中验证了LeaderAI方法的有效性。通过引入LeaderAI模型生成的先验知识，降低了高幻觉率模型的幻觉率，提高了整体模型在医疗问答领域的可靠性。

1 相关工作

1.1 国内主要研究进展

AI幻觉成因层面，围绕大语言模型展开系统化定义。胡泳[2]分析ChatGPT幻觉源于其“推理”原理，与数据集、数据压缩与提示（prompt）直接相关，平衡创造力和准确性成为AIGC未来发展的关键。莫祖英等人[3]采用数据测试实验方法，认为AIGC虚假信息主要包括事实性虚假和幻觉性虚假两种类型，产生的根源与大规模语言模型、预训练数据集和人工标注三个要素有关，为

AIGC虚假信息的进一步研究提供了理论基础。张欣[4]提出基于网络文本语料库的训练可能嵌入算法偏见，人类反馈强化学习可能加剧AI幻觉生成与传播风险。

AI幻觉解决层面，结合具体行业提出实践性解决方案。陈建兵等人[5]从AI治理角度出发，认为解决AI幻觉问题需采取综合性的举措，其中包括对数据进行精细化处理、优化算法设计、构建有效的监控与反馈体系，以确保模型所生成的答案能够符合人类的评判标准和预期，保障其整体的安全性与稳定性。王禄生[6]具体探讨法律AI产生的知识完满幻觉、知识权威幻觉与知识生成幻觉具体幻觉问题，并认为未来需要在语料源头端、训练过程端、结果生成端强化法律数据供给、法律指令微调与法律知识验证，确保技术扩散的可及性与均等化，克服“知识幻觉”以实现法律人工智能的进一步迭代。漆晨航[7]提出针对AI虚假信息，应建立虚假信息协同治理机制，通过战略规划统一规范概念，并传导至法律制度与信息执法内容实践中。胡泳[8]提出多维度安排实现问责分配、重视互联网平台的把关人作用、保障新闻媒体的正常信息供应等途径和方法解决幻觉等虚假信息问题。

1.2　国外主要研究进展

AI幻觉成因层面，技术内生原因与人脑机制类比探索。Wendland[9]从技术评估角度探讨该主题的研究项目，围绕AI自我意识发展研究，提出过程中可能产生幻觉问题，认为AI幻觉的成因与技术上的黑箱原理密切相关。Loeb[10]基于丘脑皮质连接及其与皮质注意力相关的假定功能，探索人类产生幻觉的主要研究进程和成因，并类比提出人工神经网络为主的受生物启发的AI模型，对产生AI幻觉问题的成因可能性与影响展开分析。Zhang等人[11]围绕LLM幻觉分类，提出输入冲突、上下文冲突和事实冲突三类，并分别总结出语料库、数据标注等原因，并针对性提出解决意见。Ye等人[12]围绕多种下游任务中广泛观察到的代表性幻觉进行五种分类，即机器翻译、问答、对话系统、摘要系统、基于大型语言模型的知识图谱和视觉问答。Bawden等人[13]认为诸如大小写区别等扰动会严重影响LLM的分析判断，传统应对方式的泛化性不足，造成幻觉问题加深。Umapathi等人[14]提出没有准确、可靠和可访问来源的记忆信息会导致不同类型的幻觉产生，外部知识介入是解决问题的关键。Bang等人[15]提出用于使用公开可用的数据集定量评估LLM的框架，由于无法访问外部知识库，LLM会从参数记忆中产生更多的外在幻觉。Lin等人[16]提出衡量LLM生成答案真实性的问题框架，研究得出最大的模型通常最不真实，建议使用除模仿网络文本之外的训练目标进行微调，提高真实性。

AI幻觉解决层面，提出了数据优化、人工审核等方式。Dziri等人[17]深度挖掘AI幻觉问题在数据的收集与应用方面重点困境，以数据为中心的解决方案构建全新的、更具可解释性与稳定性的对话大模型。Devanny等人[18]围绕LLM分析AIGC工具的应用价值，认为虽然存在相对显著的幻觉问题，但是可审计的分析评估流程与可信赖的人工审核可以极大降低幻觉风险的危害。Pan等人[19]重点分析LLM本身被提示或引导去修复其自身输出中的问题，即事后自我修正的技术方案解决LLM幻觉问题。Weller等人[20]结合新闻学信息传播理论，提出“根据-提示”模型，指导LLM根据观察到的文本为基础作出回应，并构建QUIP（quantization with incoherence processing and lattice codebooks）分数评估指标量化文本基础。Gaur等人[21]围绕LLM数字应用相关幻觉问题，创建并使用SVAMP数据集的符号版本鼓励符号推理与数字答案对齐，为LLM配备提供简洁和可验证推理的能力，并使其更具可解释性。Dale等人[22]提出一种方法来评估信息源对生成内容的贡献百分比，通过低的源贡献来识别和降低幻觉问题。

目前已有研究致力于大模型幻觉的检测、评估及纠正[23-24]，但相关工作仍然存在以下几个问题：

第一，当前幻觉研究通常基于通用领域，很少涉及垂直领域特别是医疗问答领域，因此很难知道当前大模型的医疗知识水平及其掌握程度。

第二，单一智能方法应对幻觉现象具有局限性，这类方法往往依赖于单个模型对幻觉现象的识别和纠正，缺乏多角度的综合判断能力[20]。

第三，不同LLM生成的回复中，幻觉的程度和类型各不相同，缺乏对不同类型幻觉的分类分析。是否可以在量化并分类模型的幻觉程度之后，找到一种方法来提升高幻觉率模型的表现呢？

为解决上述提到的三个问题，本文首先基于公开医疗问答数据集，利用GPT4o生成医疗问题的回复，结合人工标注构造了高质量的医疗领域大模型幻觉评估数据集。其次基于所构建的数据集，定义了“幻觉率”作为模型生成错误内容的比例，并将幻觉分为事实性幻觉和忠实性幻觉两类。然后设计prompt测试并量化各个大模型在医疗领域问答中的幻觉程度，对两类幻觉的概率进行分析。实验结果表明，不同模型在事实性幻觉和忠实性幻觉上的表现存在显著差异。最后采用LeaderAI模型生成先验知识，辅助LLM进行回答。实验结果表明，该方法能够降低高幻觉率模型的幻觉率，提高大模型在医疗问答领域的可靠性。

2 方法与实验

2.1 医疗问答领域幻觉评估数据集构建

为评估现有大模型的幻觉率，本文基于医疗问答领域的Huatuo数据集（https://github.com/FreedomIntelligence/Huatuo-26M）展开研究。医疗问答领域因其高度专业性和复杂性，对大模型的事实性判断能力、逻辑一致性和用户意图理解能力提出了严苛要求。本文利用正则匹配从Huatuo数据集中抽取出2 000条肝胆病和内分泌病相关数据作为种子数据构造为幻觉评估数据集，采用GPT4o、Copilot和文心一言标注生成参考答案。由于该幻觉评估数据集中参考答案由大模型生成，数据集仍然可能存在幻觉，为进一步降低数据集中存在的幻觉，对于大模型标注出的存在问题的数据，通过人工查询相关资料纠正这些数据中存在的幻觉表述。采用如图1所示的方式，通过GPT4o、GPT4、ChatGLM4、Baichuan-13B和Claude 3.5生成种子数据中问题的回答，其中虚线框内容为LMM针对该问题生成的回答。

根据问题，用一句话且简要地给出正确答案。
问题：脂肪肝的手术治疗有些什么？
正确答案：脂肪肝的手术治疗包括肝移植和减肥手术。

图1 生成幻觉数据集的prompt格式

Fig.1 Prompt format for generating datasets of hallucination

2.2 LeaderAI选举

构建完医疗幻觉测试集之后，选举一个LeaderAI对于有效的决策至关重要。在选举过程中，使用GPT4o模型作为中立评价工具，进行公平客观的选举。具体而言，使用如图2所示的prompt，明确要求模型作为中立的评价者，基于参考答案对候选模型的回答进行逐一对比分析。提示词中明确限定了模型的角色为中立评价者，而不是候选者，同时评分是基于参考答案的对比进行的。这种设计使得评价过程聚焦于对答案正确性的判断，与评价工具本身的生成能力无关。LLM作为通用性工具，具备对比分析和判断答案是否正确的能力，因此，使用GPT4o或其他模型在此场景下都可以完成公平评估。通过综合分析所有候选模型的表现，最终确定出在医疗领域中最具有潜力的LeaderAI，为后续的幻觉率测试提供基础。经过选举后，GPT4o被选为LeaderAI。

你是一名医疗领域专家，请作为中立的评价者，根据以下问题五名助理的回答，对比参考答案进行分析评估，选举出最佳的助理。
问题n：……
参考答案：……
助理1回答：……
助理2回答：……
助理3回答：……
助理4回答：……
助理5回答：……
回复：助理1

图2 LeaderAI选举的prompt格式

Fig.2 Prompt format for LeaderAI election

2.3 幻觉率测试

为评估当前大模型对于医疗问答领域的幻觉情况，本文基于性能表现、多样性和可用性选取了当前先进的大语言模型作为候选模型。这些模型包括国际领先的通用性模型GPT4o、GPT4和Claude 3.5，以及本地化开发的中文模型ChatGLM4和Baichuan-13B。采用如图3所示的prompt格式，通过选举出的LeaderAI，对于每个待测模型生成的回答进行判断，若回答被判定为回答错误则表示该待测模型产生幻觉。根据幻觉分类原则，判断该幻觉属于事实性幻觉或忠实性幻觉，同时计算了两类幻觉在所有幻觉中的占比，以区分并量化模型在不同幻觉上的表现。本文计算以下指标：

（1）幻觉率：用于衡量模型整体输出的可靠性，定义为模型产生幻觉的回答数量占总回答数量的比例。

$$R=\frac{N_{\mathrm{h}}}{N}\times 100\% \tag{1}$$

其中，N_{h}为模型生成的幻觉回答数量，N为模型生成的总回答数量。

（2）事实性幻觉：用于评估模型在事实一致性上的表现，定义为模型生成的事实性幻觉回答数量占产生幻觉回答数量的比例。

你是一名医疗专家。根据问题以及参考，判断回答是否正确。若正确返回“正确”。若错误表示出现幻觉。请根据幻觉分类原则返回“事实性幻觉”或“忠实性幻觉”。
幻觉分类原则如下。1. 事实性幻觉：回答包含可以基于现实世界信息的事实，但存在矛盾的情况；回答包含无法根据已确立的现实世界知识进行验证的事实。2. 忠实性幻觉：回答偏离了用户的指令；回答与用户提供的上下文信息不一致的情况；回答存在内部逻辑矛盾。
问题：脂肪肝的手术治疗有些什么？
参考答案：肝移植术；肝肝移植；肝活体肝移植；肝移植；肝成人肝移植；肝脏移植；胃减容手术；脂肪肝移植；自体原位肝移植；腹腔镜胃减容术；肝切除术
回答：脂肪肝的手术治疗包括肝移植和减肥手术。
返回：正确

图3 测试大模型幻觉率的prompt格式

Fig.3 Prompt format for testing hallucination rate of LLMs

$$R_{\text{factuality}}=\frac{N_{\text{factuality}}}{N_{\text{h}}}\times 100\% \tag{2}$$

其中，$N_{\text{factuality}}$ 为模型生成的事实性幻觉回答数量，N_{h} 为模型生成的幻觉回答数量。

(3)忠实性幻觉：用于评估模型回答是否忠实于指令和上下文，定义为模型生成的忠实性幻觉回答数量占产生幻觉回答数量的比例。

$$R_{\text{faithfulness}}=\frac{N_{\text{faithfulness}}}{N_{\text{h}}}\times 100\% \tag{3}$$

其中，$N_{\text{faithfulness}}$ 为模型生成的忠实性幻觉回答数量，N_{h} 为模型生成的幻觉回答数量。

(4)幻觉率降幅

$$R_{\text{reduce}}=\frac{N_{\text{h}}-N_{\text{h}}'}{N_{\text{h}}}\times 100\% \tag{4}$$

其中，N_{h} 表示原幻觉数量(原幻觉总数量、原真实性幻觉数量或原忠实性幻觉数量)；N_{h}' 表示引入LeaderAI后的幻觉数量(引入LeaderAI后的幻觉总数量、真实性幻觉数量或忠实性幻觉数量)。

本文对5个待测模型进行测试，其中幻觉率的测试结果如表2所示。结果表明，GPT4o的幻觉率相对最低，为46.70%。说明它在生成内容时具有相对较高的准确性和可靠性，同时也符合LeaderAI的选举结果。Baichuan-13B的幻觉率最高，达到了68.30%，这意味着它更容易产生偏差或不准确信息。GPT4、Claude 3.5和ChatGLM4幻觉率基本持平，在这5个模型中属于中等水平。

表2　不同大模型幻觉率

Table 2　Hallucination rates of different LLMs

大模型	幻觉	幻觉率/%
GPT4o	934	46.70
GPT4	1 221	61.05
Claude 3.5	1 248	62.40
ChatGLM4	1 285	64.25
Baichuan-13B	1 366	68.30

如表3所示，整体来看GPT4o、GPT4和ChatGLM4的表现相对均衡，在忠实性方面相较于其他大模型表现得更好。具体而言，GPT4o不仅在整体幻觉率方面较低，在忠实性幻觉方面的表现也更加突出，这使得它在对输入信息的忠实性和生成内容的准确性之间取得了较好的平衡。Claude 3.5和Baichuan-13B的事实性幻觉和忠实性幻觉比例相对接近，表明这两类幻觉的发生概率相当。实验结果表明，不同模型在事实性幻觉和忠实性幻觉上的表现存在显著差异。那么，如何降低幻觉率，提高模型对输入信息的忠实性和内容的准确性，成为模型优化输出的核心目标之一。

表3　不同大模型分类幻觉率

Table 3　Categorized hallucination rates of different LLMs

大模型	事实性幻觉	忠实性幻觉	事实性幻觉率/%	忠实性幻觉率/%
GPT4o	597	337	63.92	36.08
GPT4	747	474	61.18	38.82
Claude 3.5	599	649	48.00	52.00
ChatGLM4	808	477	62.88	37.12
Baichuan-13B	685	681	50.15	49.85

2.4　低幻觉率模型指导高幻觉率模型

从上述实验结果可以分析得到，幻觉率高的LLM缺乏足够的领域知识，难以对较专业的问题进行有效回答，同时不同的大模型对于医学领域问答在事实性幻觉和忠实性幻觉上的表现存在显著差异。为了降低大模型的幻觉率，本文假设强大的大模型具备更丰富的知识、更低的幻觉率，从而希望能够通过其优化较弱大模型的表现。这一方法在一些特定场景下具有实际应用价值。例如，在计算资源受限的场景中，利用低幻觉率模型生成的高质量先验知识指导多个轻量级模型，可有效降低成本并提升回答的可信性。此外，在需要领域适配或多模型协作的任务中，这一方法也展现出显著优势。通过引入更强大模型的先验知识，可以缓解较弱模型在复杂任务中的幻觉问题，提高其整体性能和适应性。本文已知GPT4o作为选举出的LeaderAI，并具有最低的幻觉率。为利用LeaderAI中丰富的领域知识，采用图4所示的prompt让LeaderAI生成先验知识。最后使用图5所示的prompt，基于之前所构建的幻觉评估数据集对各模型再次进行测试，测试结果如表4所示。

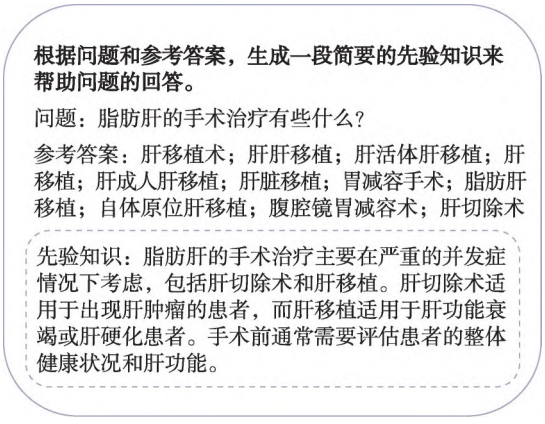

图4　采用LeaderAI生成先验知识的prompt格式

Fig.4　Using LeaderAI to generate prompt format for prior knowledge generation

表4为引入LeaderAI先验知识后各个大模型的幻觉情况。可以发现所有的大模型在LeaderAI先验知识的辅助下，除GPT4o自身以外，其他模型幻觉率都有着明显的下降。其中Baichuan-13B的幻觉率下降幅度达

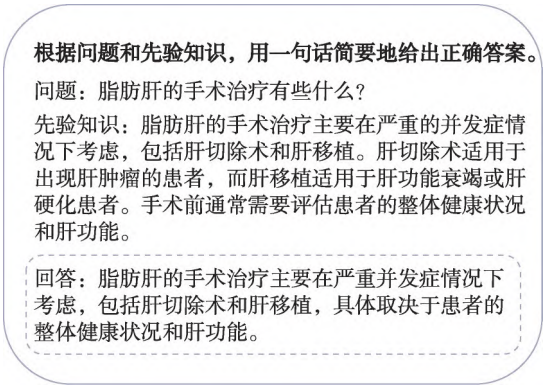

图5 结合先验知识测试大模型幻觉率的prompt格式

Fig.5 Combining prior knowledge to retest prompt format for LLMs hallucination rate

表4 引入LeaderAI先验知识辅助后各大模型幻觉率

Table 4 Introducing LeaderAI prior knowledge to assist in hallucination rate of each LLM

大模型	幻觉	幻觉率/%	降幅/%
GPT4o	902	45.10	3.43
GPT4	926	46.30	24.16
Claude 3.5	994	49.70	20.35
ChatGLM4	997	49.85	22.41
Baichuan-13B	1 008	50.40	26.21

到了26.21%。各个模型在LeaderAI先验知识的帮助下提升了对事实和背景信息的理解，使其在医疗问答领域的表现更加稳健。同时可以发现，引入LeaderAI后，各个模型的幻觉率都趋近于LeaderAI的幻觉率，表明在prompt中引入先验知识对模型的影响较大。

表5和表6分别为引入LeaderAI先验知识后各个大模型的事实性幻觉和忠实性情况。从表中可以看出，引入LeaderAI后，各模型的忠实性幻觉占比均显著下降，其中Claude 3.5的忠实性幻觉占比下降了52.70%。也就是说，在prompt中引入LeaderAI的先验知识能够使模型的回答更加贴合用户的指令，与提供的上下文信息一致，且内部逻辑更加合理。这意味着模型在理解和响应用户需求方面有所提升，回答的质量得到改善。然而，从表5分析可以发现，各模型的事实性幻觉占比存

表5 引入LeaderAI先验知识辅助后各大模型事实性幻觉率

Table 5 Introducing LeaderAI prior knowledge to assist in factuality hallucination rate of each LLM

大模型	事实性幻觉	事实性幻觉率/%	降幅/%
GPT4o	608	67.41	−1.84
GPT4	627	67.71	16.06
Claude 3.5	687	69.11	−14.69
ChatGLM4	691	69.31	14.48
Baichuan-13B	708	70.24	−3.36

表6 引入LeaderAI先验知识辅助后各大模型忠实性幻觉率

Table 6 Introducing LeaderAI prior knowledge to assist in faithfulness hallucination rate of each LLM

大模型	忠实性幻觉	忠实性幻觉率/%	降幅/%
GPT4o	294	32.59	12.76
GPT4	299	32.29	36.92
Claude 3.5	307	30.89	52.70
ChatGLM4	306	30.69	35.85
Baichuan-13B	300	29.76	55.95

在一定波动，特别是Claude 3.5模型，部分忠实性幻觉转化为事实性幻觉。具体表现为Claude 3.5在理解用户意图和生成忠实内容方面有所提升，但仍然避免不了事实性幻觉的出现。

3 结论

为探索大模型对于医疗问答领域中存在的幻觉，本文首先基于公开数据集Huatuo并采用GPT4o结合人工标注的方式构建幻觉评估数据集。使用GPT4o、GPT4、ChatGLM4、Baichuan-13B和Claude 3.5大语言模型对数据集中的问题生成答案。通过群体智能方法选举出一个LeaderAI，它将各模型的回答与参考答案进行比较，计算各模型的幻觉率，并根据设计的幻觉原则prompt将幻觉分为事实性幻觉和忠实性幻觉两类。实验结果表明，在LeaderAI的指导下，各模型的幻觉率显著降低，尤其是忠实性幻觉率大幅减少。

参考文献：

[1] ZHANG S, PAN L M, ZHAO J Z, et al. The knowledge alignment problem: bridging human and external knowledge for large language models[EB/OL]. (2023-05-23) [2024-08-22]. https://arxiv.org/pdf/2305.13669.pdf.

[2] 胡泳. 当机器人产生幻觉，它告诉我们关于人类思维的什么?[J]. 文化艺术研究, 2023, 16(3): 15-26.
HU Y. When robots hallucinate: what does it tell us about human thinking?[J]. Studies of Culture and Art, 2023, 16(3): 15-26.

[3] 莫祖英, 盘大清, 刘欢, 等. 信息质量视角下AIGC虚假信息问题及根源分析[J]. 图书情报知识, 2023, 40(4): 32-40.
MO Z Y, PAN D Q, LIU H, et al. Analysis on AIGC false information problem and root cause from the perspective of information quality[J]. Documentation, Information & Knowledge, 2023, 40(4): 32-40.

[4] 张欣. 面向产业链的治理: 人工智能生成内容的技术机理与治理逻辑[J]. 行政法学研究, 2023(6): 43-60.
ZHANG X. Industry chain-oriented governance: technological mechanisms and governance logic in the management of artificial intelligence generated content[J]. Administrative Law Review, 2023(6): 43-60.

[5] 陈建兵, 王明. 负责任的人工智能: 技术伦理危机下AIGC的治理基点[J]. 西安交通大学学报(社会科学版), 2024, 44

(1): 111-120.
CHEN J B, WANG M. Responsible artificial intelligence: governance fundamentals for AIGC in the ethical crisis of technology[J]. Journal of Xi'an Jiaotong University (Social Sciences), 2024, 44(1): 111-120.
[6] 王禄生. ChatGPT类技术: 法律人工智能的改进者还是颠覆者?[J]. 政法论坛, 2023, 41(4): 49-62.
WANG L S. ChatGPT-like technology: improver or disruptor of legal AI?[J]. Tribune of Political Science and Law, 2023, 41(4): 49-62.
[7] 漆晨航. 生成式人工智能的虚假信息风险特征及其治理路径[J]. 情报理论与实践, 2024, 47(3): 112-120.
QI C H. Research on the risks of disinformation from generative artificial intelligence and its governance paths[J]. Information Studies (Theory & Application), 2024, 47(3): 112-120.
[8] 胡泳. 人工智能驱动的虚假信息: 现在与未来[J]. 南京社会科学, 2024(1): 96-109.
HU Y. AI-driven disinformation: present and future[J]. Nanjing Journal of Social Sciences, 2024(1): 96-109.
[9] WENDLAND K. Demystifying artificial consciousness-about attributions, black swans, and suffering machines[J]. Journal of AI Humanities, 2021, 9: 137-166.
[10] LOEB G E. Remembrance of things perceived: adding thalamocortical function to artificial neural networks[J]. Frontiers in Integrative Neuroscience, 2023, 17: 1108271.
[11] ZHANG Y, LI Y F, CUI L Y, et al. Siren's song in the AI ocean: a survey on hallucination in large language models [EB/OL]. (2023-09-03)[2024-08-22]. https://arxiv.org/pdf/2309.01219.pdf.
[12] YE H B, LIU T, ZHANG A J, et al. Cognitive mirage: a review of hallucinations in large language models[EB/OL]. [2024-08-22]. https://arxiv.org/abs/2309.06794.
[13] BAWDEN R, YVON F. Investigating the translation performance of a large multilingual language model: the case of bloom[C]//Proceedings of the 24th Annual Conference of the European Association for Machine Translation. Stroudsburg: ACL, 2023: 157-170.
[14] PAL A, UMAPATHI L K, SANKARASUBBU M. Med-HALT: medical domain hallucination test for large language models [C]//Proceedings of the 27th Conference on Computational Natural Language Learning. Stroudsburg: ACL, 2023: 314-330.
[15] BANG Y J, CAHYAWIJAYA S, LEE N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity[C]//Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 675-718.
[16] LIN S, HILTON J, EVANS O. TruthfulQA: measuring how models mimic human falsehoods[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 3214-3252.
[17] DZIRI N, KAMALLOO E, MILTON S, et al. FaithDial: a faithful benchmark for information-seeking dialogue[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 1473-1490.
[18] DEVANNY J, DYLAN H, GROSSFELD E. Generative AI and intelligence assessment[J]. The RUSI Journal, 2023, 168(7): 16-25.
[19] PAN L M, SAXON M, XU W D, et al. Automatically correcting large language models: surveying the landscape of diverse self-correction strategies[EB/OL]. (2023-08-06) [2024-08-22]. https://arxiv.org/pdf/2308.03188.pdf.
[20] WELLER O, MARONE M, WEIR N, et al. "According to ..." prompting language models improves quoting from pre-training data[C]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 2288-2301.
[21] GAUR V, SAUNSHI N. Reasoning in large language models through symbolic math word problems[C]//Findings of the Association for Computational Linguistics: ACL 2023. Stroudsburg: ACL, 2023: 5889-5903.
[22] DALE D, VOITA E, BARRAULT L, et al. Detecting and mitigating hallucinations in machine translation: model internal workings alone do well, sentence similarity even better [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 36-50.
[23] 岳颀, 张晨康. 多模态场景下AIGC的应用综述[J]. 计算机科学与探索, 2025, 19(1): 79-96.
YUE Q, ZHANG C K. Survey on applications of AIGC in multimodal scenarios[J]. Journal of Frontiers of Computer Science and Technology, 2025, 19(1): 79-96.
[24] 张钦彤, 王昱超, 王鹤羲, 等. 大语言模型微调技术的研究综述[J]. 计算机工程与应用, 2024, 60(17): 17-33.
ZHANG Q T, WANG Y C, WANG H X, et al. Comprehensive review of large language model fine-tuning[J]. Computer Engineering and Applications, 2024, 60(17): 17-33.

何静(1989—),女,四川遂宁人,博士,讲师,主要研究方向为人工智能、大数据等。
HE Jing, born in 1989, Ph.D., lecturer. Her research interests include artificial intelligence, big data, etc.

沈阳(1974—),男,江西赣州人,博士,教授,主要研究方向为人工智能、大数据等。
SHEN Yang, born in 1974, Ph.D., professor. His research interests include artificial intelligence, big data, etc.

谢润锋(1999—),男,福建漳州人,硕士研究生,主要研究方向为自然语言处理。
XIE Runfeng, born in 1999, M.S. candidate. His research interest is natural language processing.