

大语言模型的幻觉问题研究综述^{*}



刘泽垣¹, 王鹏江¹, 宋晓斌¹, 张欣², 江奔奔¹

¹(清华大学, 北京 100084)

²(对外经济贸易大学, 北京 100029)

通信作者: 江奔奔, E-mail: bbjiang@tsinghua.edu.cn

摘要: 随着以 Transformer 为代表的预训练模型等深度学习技术的发展, 大语言模型 (LLM) 日益展现出强大的理解力和创造力, 对摘要、对话生成、机器翻译和数据到文本生成等下游任务产生了重要影响, 同时也在图像说明、视觉叙事等多模态领域展现出了广阔的应用前景. 虽然大语言模型具备显著的性能优势, 但深度学习架构使其难以避免内容幻觉问题, 这不仅会削弱系统性能, 还严重影响其可信性和应用广泛性, 由此衍生的法律风险和伦理风险成为掣肘其进一步发展与落地的主要障碍. 聚焦大语言模型的幻觉问题, 首先, 对大语言模型的幻觉问题展开系统概述, 分析其来源及成因; 其次, 系统概述大语言模型幻觉问题的评估方法和缓解方法, 对不同任务的评估和缓解方法类型化并加以深入比较; 最后, 从评估和缓解角度展望应对幻觉问题的未来趋势和应对方案.

关键词: 可信人工智能; 大语言模型; 幻觉; 幻觉评估与缓解

中图法分类号: TP18

中文引用格式: 刘泽垣, 王鹏江, 宋晓斌, 张欣, 江奔奔. 大语言模型的幻觉问题研究综述. 软件学报, 2025, 36(3): 1152-1185. <http://www.jos.org.cn/1000-9825/7242.htm>

英文引用格式: Liu ZY, Wang PJ, Song XB, Zhang X, Jiang BB. Survey on Hallucinations in Large Language Models. Ruan Jian Xue Bao/Journal of Software, 2025, 36(3): 1152-1185 (in Chinese). <http://www.jos.org.cn/1000-9825/7242.htm>

Survey on Hallucinations in Large Language Models

LIU Ze-Yuan¹, WANG Peng-Jiang¹, SONG Xiao-Bin¹, ZHANG Xin², JIANG Ben-Ben¹

¹(Tsinghua University, Beijing 100084, China)

²(University of International Business and Economics, Beijing 100029, China)

Abstract: With the development of deep learning technologies such as pre-trained models, represented by Transformer, large language models (LLMs) have shown excellent comprehension and creativity. They not only have an important impact on downstream tasks such as abstractive summarization, dialogue generation, machine translation, and data-to-text generation but also exhibit promising applications in multimodal fields such as image description and visual narratives. While LLMs have significant advantages in performance, deep learning-based LLMs are susceptible to hallucinations, which may reduce the system performance and even seriously affect the trustworthiness and broad applications of LLMs. The accompanying legal and ethical risks have become the main obstacles to their further development and implementation. Therefore, this survey provides an extensive investigation and technical review of the hallucinations in LLMs. Firstly, the hallucinations in LLMs are systematically summarized, and their origin and causes are analyzed. Secondly, a systematical overview of hallucination evaluation and mitigation is provided, in which the evaluation and mitigation methods are categorized and thoroughly compared for different tasks. Finally, the future challenges and research directions of the hallucinations in LLMs are discussed from the perspectives of evaluation and mitigation.

Key words: trustworthy artificial intelligence; large language model (LLM); hallucination; hallucination evaluation and mitigation

* 基金项目: 国家重点研发计划 (2022YFE0197600)

刘泽垣和王鹏江为共同第一作者.

收稿时间: 2024-01-23; 修改时间: 2024-05-03; 采用时间: 2024-06-20; jos 在线出版时间: 2024-12-10

CNKI 网络首发时间: 2024-12-11

1 研究背景介绍

1.1 语言模型发展概述

大语言模型(以下简称大模型)是指基于大规模语料库进行预训练的超大型深度学习模型,在解决文案写作、知识库回答、文本分类、代码生成、文本生成等下游自然语言处理任务中表现出强大的能力^[1-4]。大语言模型多基于Transformer架构开展预训练,利用自注意力以更好地捕捉词汇、语法和语义等语言知识,在处理长期依赖性方面比其他神经网络架构具有更好的表现^[5],同时可充分利用硬件的并行性,达到训练比以往规模更大、功能更强的模型训练效果。大模型训练所需数据集的规模可达千亿级别,相比于普通预训练模型更加复杂和庞大,而正是因为模型规模的增大,打破了原有的模型性能定律,产生了涌现能力。例如,参数规模更加庞大的GPT-3模型可以通过上下文学习解决少样本问题,而较小规模的GPT-2模型则在此方面表现不佳^[6]。虽然这些模型可生成表达流畅、语法规范的文本,但却易于输出与输入文本、真实世界知识相矛盾的事实错误或逻辑错误,从而产生幻觉问题(hallucination)。

1.2 大模型幻觉问题

幻觉问题在大语言模型中广泛存在,已经成为自然语言生成面临的最大挑战之一。大模型可能自信地输出错误或者不存在的答案,这种潜在的幻觉问题将极大地限制其在实际场景中的应用,衍生出法律和伦理风险。例如,对文本摘要生成的研究表明目前最先进模型生成的摘要中约有30%存在偏离事实的幻觉问题,严重影响了模型的可靠性和可用性^[7]。此外,幻觉问题广泛存在于几乎所有的自然语言生成下游任务中,如文本摘要、对话生成、机器翻译和数据到文本生成任务等^[8-11]。

1.2.1 大模型幻觉问题的提出

大模型的广泛应用极大地推动了以自然语言生成为代表的下游任务部署。自然语言生成任务可分为开放式和非开放式。开放式语言生成任务是指输入不完整且输出语义不包含在输入中的任务类型。大模型需要利用知识图谱或语料库中的知识来创建输入中不包含的新内容,而由于幻觉问题的存在,大模型所生成的内容可能不符合真实世界的知识。例如,在新闻写作生成任务中,模型可能生成并未在真实世界发生或存在的内容^[12]。

相比之下,在非开放式语言生成任务中,大模型需要根据输入生成文本,同时为输出文本提供完整甚至额外信息。在此类任务的实际应用中可能存在模型生成的内容与输入信息的事实不一致的情况。例如,文本摘要生成任务中,模型生成的摘要内容与输入文档不一致,且与源文档存在出入^[13];机器翻译任务中,模型生成的译文与原文内容不一致^[14]。因此,无论是开放式语言生成任务,还是非开放式语言生成任务,幻觉问题均普遍存在。

1.2.2 大模型幻觉问题的伦理影响和法律风险

ChatGPT等大模型的纷至推出成为全球科技竞争焦点。大模型成为引领新一轮科技革命和产业革新的战略性技术,具备成为赋能千行百业的通用基础设施的潜能。而随着大模型的不断发展,其幻觉问题引发的伦理影响和法律风险也日益严峻。一方面,迅速迭代的模型愈发具有创造力,导致用户过度依赖。这种依赖增加了用户对于模型能力的信任从而对内容幻觉缺乏警惕;另一方面,人类认知存在摩西幻觉(Moses illusion),当事实中的部分内容被错误但相似的信息替代时,人类往往难以识别^[15]。伴随着大模型商业化落地的进程,其引发的法律与伦理风险已经成为掣肘其广泛部署的关键。在一些高风险场景下,一旦大模型产生幻觉,可能造成难以救济的风险。例如,若司法审判环节采用大模型生成裁判文书,一旦出现幻觉,可能直接威胁裁决公正和当事人的基本权利。在医疗领域,如果基于患者信息生成的治疗方案出现幻觉,将直接威胁患者的健康和生命安全。同样,即使在机器翻译应用场景中,如果生成的药品说明书存在幻觉,也有可能产生导致严重危害患者生命的风险。因此,厘清大模型幻觉问题的成因与类型,剖析评估方法和缓解方法具有理论和实践层面的双重意义。

1.3 本文结构

本文第2节对大模型幻觉问题进行概述,分析幻觉问题的来源及成因,并从不同角度给出了幻觉问题的定义。第3节和第4节分别综述大模型幻觉的评估方法和缓解方法,并对其展开分析对比。最后,第5节在以上调研的基础上展望幻觉问题的应对方案和研究方向。

2 问题分析

2.1 大模型幻觉的定义

早在预训练阶段,“幻觉”一词已经在自然语言处理中被广泛采用,通常是指模型生成的内容对提供的源内容无意义或不忠实^[16].现阶段,考虑到多样化的生成内容往往包含不属于源内容的事实知识,大模型幻觉通常是指模型生成的文本不忠实于信息源或者与现实世界的事实不符.例如,在摘要生成任务中,大模型生成的摘要可能无法对应源文本中的事实正确信息;在对话生成任务中,对话产生的输出与对话历史或外部事实相矛盾;在机器翻译任务中,则是指大模型产生完全脱离原始材料的错误翻译.

对于不同的下游任务,大模型对幻觉问题的容忍度存在显著差异.例如,数据文本生成或文章撰写任务对幻觉的容忍度相对较低,因为这些任务可能涉及高风险决策或学术研究.而在对话生成任务中,只要避免事实性错误,对幻觉的容忍度则相对较高.由于大模型强大的泛化能力和适应性,更多的新型任务如推荐系统、创意设计、智能检索等幻觉风险也随之浮现.因此,大模型的幻觉定义需要以类型化方式从多个角度归纳.

2.2 大模型幻觉的分类

2.2.1 内在幻觉与外在幻觉

根据大模型与输入信息源的关系,可以将幻觉问题分类为内在幻觉和外在幻觉^[16].

内在幻觉是指大模型生成与输入信息源内容相矛盾的输出.例如,在摘要生成任务中,生成的摘要与文章源内容相矛盾.再如,在机器翻译任务中,翻译输出的文本与源文本提供的内容不同且相违背.

外在幻觉是指无法从输入信息源中验证模型所生成的输出.此类输出既不能被源内容支撑,也不能被源内容反驳.以摘要生成任务为例,如果生成的输出在来信息源中没有提及,则用户既不能从源内容找到生成的输出的证据,也不能证明输出是错误的.因此,外在幻觉并不总是错误的,其可能来自事实正确的外部信息,并产生一定的创造性,从而有助于广告营销、内容创作类下游任务的完成.

2.2.2 封闭域幻觉与开放域幻觉

基于推理的作用范围和上下文限制,推理模式可分类为封闭域推理和开放域推理.封闭域推理指的是模型在特定的上下文或领域中进行推理,而开放域推理则是指模型在更广泛的领域或情境中进行推理.基于大模型在特定任务或领域中推理的表现,可以将幻觉分为封闭域幻觉和开放域幻觉.

OpenAI 团队正是以封闭域和开放域两个角度来评估 GPT-4 模型的潜在幻觉问题的^[17].具体而言,封闭域幻觉是指模型仅提供了给定的上下文信息源,但在内容生成过程中臆造了超出上下文的信息.例如,在摘要生成任务中,大模型生成的摘要包含了源文章不存在的信息.开放域幻觉则是指模型在更广泛的领域或任务中生成的输出存在虚假的情况.此种情况下,模型可能在没有参考任何特定输入上下文的情况下自发提供有关议题的错误信息.例如,在对话生成任务中,大模型为研究者提供了不存在的文献、为律师提供了不存在的案例,此类幻觉也引发了社会和各国监管机构对于生成式人工智能的普遍担忧^[18].

2.2.3 输入矛盾幻觉、上下文矛盾幻觉以及事实矛盾幻觉

基于大模型在处理信息时可能出现的矛盾类型,可以将幻觉分类为输入矛盾幻觉、上下文矛盾幻觉与事实矛盾幻觉^[19].输入矛盾幻觉由大模型偏离用户输入导致.其中用户输入包括两个组成部分:任务指令和任务输入.大模型响应与任务指令之间的矛盾通常反映为模型误解了用户意图.当模型生成的内容和任务输入之间出现矛盾,出现类似内在幻觉的情形,则体现为输入矛盾幻觉.

上下文矛盾幻觉是大模型在产生冗长或多回合响应时可能出现的自相矛盾现象.其主要体现为模型在生成的文本中出现前后不一致,或在对话过程中出现自相矛盾.这种幻觉主要由大模型在长期记忆方面的限制,或识别相关上下文方面的不足导致.

事实矛盾幻觉则是指当大模型生成与既定世界知识相矛盾的信息或文本时所产生的幻觉.事实冲突幻觉的来源可能是多种多样的,并可能在大模型生命周期的不同阶段引入.

2.3 大模型幻觉的成因

本节将分别从数据层、模型层与应用层分析大模型的幻觉成因,如图1所示。

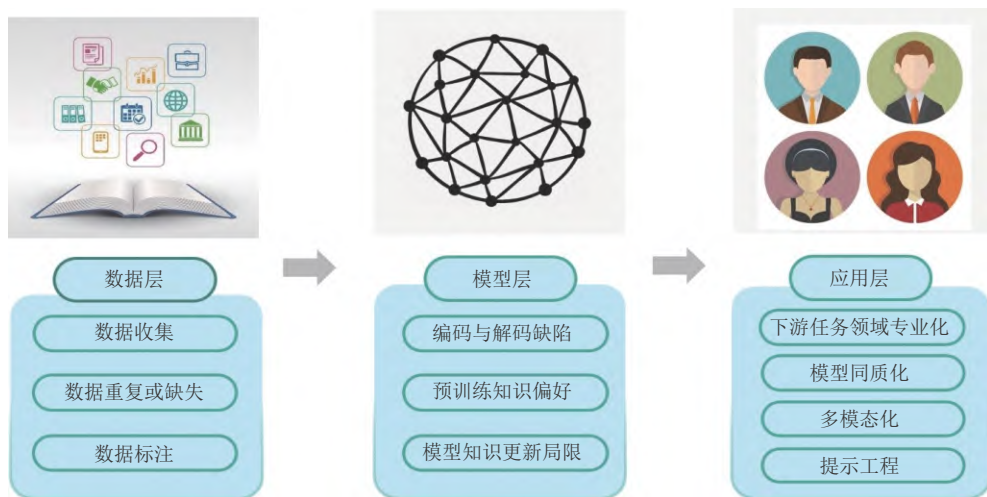


图1 大模型幻觉成因

2.3.1 数据层

大模型训练需要海量多源数据。这些训练数据多来自于网络,数据质量参差不齐。当训练数据采自于充斥着虚假信息和毒害信息的社交媒体和公共网页时,模型难以避免地会吸收并重现这些不真实的信息,造成训练数据源与模型生成内容之间的分歧,从而潜在性地促使模型生成不忠实于所提供数据源的文本,导致幻觉的产生。

2.3.1.1 数据收集

为了节约数据成本、提高数据生成效率,大模型训练数据并非全部经过人工校验,而是配合采用启发式收集方法。该方法旨在适当精度范围内找到快速可行的解决方案,但可能会损失部分准确性。在收集大规模数据集时,部分模型会启发式地选择超出提供的信息源以外的真实信息作为信息源和目标配对,从而导致目标引用包含信息源不支持的信息。例如,有研究^[20]指出,在 RotoWire 数据集的摘要生成任务中,编写的摘要中有大约 40% 的内容无法直接映射到任何输入的表格记录,这导致了模型在学习时产生不符合事实的幻觉。

2.3.1.2 数据重复或缺失

来自训练语料库的重复示例会导致模型偏向于生成重复的、来自重复示例中记忆短语的输出。数据缺失则可能导致模型产生偏离于事实的预测,这种预测会被下游任务继承,从而产生幻觉问题。目前,Google 团队已经发现在某些大模型数据集中存在许多近似重复的样本和长重复的子串,导致模型在无提示输出时直接复制重复文本^[21]。因此,这些数据集中存在的重复和缺失极易产生不遵循事实的幻觉问题。

2.3.1.3 数据标注

标注数据集为模型提供了宝贵的训练资源,常被应用于指令微调和对齐等关键步骤。但在大模型的发展背景下,数据标注方法正逐渐从纯人工方式向机器辅助、甚至全机器标注转变。这种变革无疑提高了数据生成的效率,但也带来了刻板印象、文本编写错误、个体间回答风格的不一致性,以及跨国和跨地区的文化差异等问题,从而使模型在解读或生成内容时出现偏离事实或预期的幻觉^[22]。

2.3.2 模型层

模型的结构设计也是大模型幻觉的关键成因之一。Parikh 等人^[23]的研究表明,即便采用了高质量的训练数据,模型因其固有结构仍可能出现幻觉。这种由模型结构引发的幻觉,可能与编码器、解码器、模型预训练阶段的知识偏好、曝光偏差及模型知识更新局限有关。

2.3.2.1 编码与解码缺陷

在大模型架构中,编码器的作用是将输入数据解析为有深度的内部表现形式,解码器的作用是从编码器获取编码输入并生成最终目标序列.如果编码器在这种表征上的学习不够精确,可能会导致对训练数据的不同部分进行错误的编码关联,从而产生与原输入不一致的误导输出,引发幻觉问题.关联错误也与不完全的编码有关,若解码器关注了输入数据源的错误编码部分,会导致生成错误的结果^[24].这种错误的关联将导致生成的内容中混淆了两个相似实体之间的事实,产生偏离于事实的幻觉.考虑到模型任务对长上下文的不断增长的依赖,编码器处理多模态数据的准确性和效率,以及解码器的策略设计成为了决定模型是否出现幻觉的关键因素之一.

2.3.2.2 预训练知识偏好

在正式部署前,大模型会历经预训练阶段,其中与训练数据集相关的知识会被编码为模型的参数并持续被模型记忆.Longpre 等人^[25]的研究揭示,在生成过程中,大模型更倾向于依赖记忆的参数知识,而非实时的输入信息.这表明,模型在生成输出时更多地使用了先验参数知识,而非实时输入,这可能成为引发幻觉输出的又一来源.

2.3.2.3 模型知识更新局限

大模型无法及时更新迭代是其产生事实错误幻觉的关键来源之一.大模型训练的权重更新方式导致其具有灾难性遗忘局限.当使用新的数据集训练已有的大模型时,该模型将会失去对原数据集识别的能力,而且模型的参数越多、网络结构层数越深,模型的表达能力越强,就越容易出现灾难性遗忘,进而增加了产生幻觉的可能性.

2.3.3 应用层

随着大模型在各种应用场景中的广泛使用,一些关键的应用层趋势同样可能成为产生幻觉的诱因.下游任务的领域专业化可能导致模型在特定领域中失去泛化能力.模型同质化趋势使得不同任务使用相似的预训练模型,可能导致对某些输入过于敏感.模型处理多种数据类型的数据多模态化趋势,增加了模型的复杂性和产生幻觉的可能性.而提示工程虽提高了模型响应的灵活性,但也可能使模型过度依赖提示.

2.3.3.1 下游任务领域专业化

在为特定下游任务部署大模型时,模型往往展现出领域专业化特征.以摘要生成为例,大模型会被训练为从输入文本中抽取核心内容,并简洁地重组为摘要.为确保摘要的准确性和专业性,模型可能需要专门针对特定领域进行优化,这可能会限制模型处理其他广泛任务的能力.这种专业化可能导致模型在实际应用中与训练数据的分布存在偏差,从而引发幻觉问题^[26].

2.3.3.2 模型同质化

目前,主流大模型多基于 Transformer 架构的序列建模方法训练部署完成.在该基础架构下,通过大规模的数据训练和微调产生了更为先进的模型,迸发出强大的涌现能力;同时,采用相同的基础架构也导致了模型的同质化.同质化和涌现能力以一种相互作用的方式显著影响大模型性能^[26].一方面,涌现为模型带来了创新能力,并为解决更多的下游任务增加了可能性;另一方面,同质化虽然可以在广泛的应用中构建方法论的合集,但是其也产生了极高的杠杆作用,即模型中的任何缺陷都会被所有应用模型盲目继承,包括幻觉问题和偏见问题等.

2.3.3.3 多模态化

基于多领域知识,构建统一的、跨场景、多任务的多模态基础模型已逐渐成为主流.模型多模态化虽然扩展了下游部署的应用范围,但同样增加了幻觉问题的复杂性.由于每个领域的的数据都有其独特结构和属性,这种差异可能意味着需要为每种模态设计独特的训练参数,而无法简单采用统一的参数策略.因此,随着训练规模的增加,大模型多模态趋势可能会导致更严重的幻觉现象.

2.3.3.4 提示工程

提示工程是大模型部署前必不可少的环节,其作用是让模型逐步学会人类的自然指令执行任务,而无需根据下游任务微调模型或更改模型参数.当前,指令微调技术和思维链技术是提示工程的代表性技术^[27,28],但在某些场景下可能引发幻觉问题.

指令微调技术由 Google 首创,旨在实现模型根据人类指令举一反三的效果^[27].其数据集通常由人工手写指令和语言模型引导的指令实例组成,但其中可能包括多个输入和输出实例,从而引发幻觉风险.例如,MiniGPT4 和

LLaVA等大模型采用合成指令数据进行微调。这些指令数据通常很长,并且可能涉及不存在的对象、情节或者关系,导致幻觉问题的产生。此外,数据指令模板的单一性也加剧了幻觉问题。现有模板多为正向的指令数据,而忽视了负面指令以及更多角度和语义程度的指令^[29]。

思维链技术旨在通过给模型提供推理提示,使其模仿人类逐步思考和推理流程,从而使模型掌握从简单任务到复杂任务的推理能力。然而,在推理过程中,由于难以评估推理过程的合理性和有效性,可能导致模型输出带有幻觉的答案。例如,Kojima等人^[30]通过实验证明,即使模型生成的最终答案是正确的,模型在推理过程中仍然可能使用了无效的推理链,导致模型幻觉的产生。

3 大模型幻觉的评估方法

精确评估大模型幻觉对于模型性能评价、幻觉缓解及适配下游任务等方面具有重要意义。然而,大多数传统度量指标不足以充分量化模型的幻觉水平^[16,31]。尤其是在从表到文本生成、摘要生成任务中,传统度量指标与人工判别的相关性十分微弱。为此,旨在准确量化模型幻觉的更为有效的评估指标被相继提出。根据大模型在数据层、模型层与应用层的幻觉成因,目前大模型幻觉的评估方法可被分为3类:基于数据文本的评估方法,基于模型的评估方法,以及基于多任务应用的评估方法。本节将从方法原理、研究进展及优缺点分析等方面梳理这3类评估方法。

3.1 基于数据文本的评估方法

基于数据文本的评估方法主要通过计算精确率、召回率等能够度量生成文本与参考文本之间信息匹配程度的统计指标来量化生成文本的幻觉程度。常用方法是利用词汇特征开展匹配计算^[16],以此为依据可将评估方法分为3类。

- 第1类方法是将目标文本作为参考文本设定统计指标。例如,Dhingra等人^[32]将蕴含精确率与蕴含召回率相结合共同度量(*F-score*)得分,提出针对从表到文本生成任务的PARENT指标,该方法克服了单纯依靠BLEU、ROUGE等指标存在的与人工判断结果的一致性较差的问题,提高了度量模型生成文本中幻觉的可靠性。Manakul等人^[33]利用WikiBio数据集评估事实陈述之间的一致性,通过设计AUC-PR分数检测文本生成段落与目标段落之间的幻觉。Wu等人^[34]利用RAG方法构建了包含18000个目标响应的幻觉语料库,通过统计幻觉响应数、幻觉响应率以及幻觉跨度数评估模型生成文本响应的幻觉程度。

- 第2类方法仅使用源文本作为参考。这一策略在评估时不需要目标文本,更加适应输出结果有多种可能性的场景。Wang等人^[35]提出了PARENT-T指标,该方法在PARENT指标基础上进行了优化,省略了关于目标文本的比较计算,提高了幻觉评估效率。Shuster等人^[36]针对基于知识的对话(KGD)任务提出了Knowledge F1,将传统F1指标用到的人类回复修改为数据收集过程中所基于的知识,将幻觉评估推广到训练数据之外的场景。

- 第3类方法并不局限于参考的文本,可以对多种文本进行扩展,并针对具体任务场景设计不同的评估方法。例如,Popović^[37]提出chrF分数评估方法,该方法基于字符n-gram对文本幻觉进行F-度量,在机器翻译任务中显示了优越的性能。Martindale等人^[38]设计了一种新的句子相似度(BVSS)评估方法,该方法利用输出结果与翻译参考的信息量来度量句子的恰当性,以帮助判断并量化幻觉,适用于机器翻译的幻觉评估。

基于文本的评估方法从数据层角度考察模型的幻觉程度,主要用于对内在幻觉、封闭域幻觉以及上下文矛盾幻觉进行评估,评估的精确度与指标的设计和数据的大小息息相关,其局限性在于难以从理解文本含义的角度进行评估。

3.2 基于模型的评估方法

为了能够理解生成文本和参考文本的含义从而更精确地评估幻觉,使用额外的模型来量化幻觉的评估方法被相继提出。这类方法大致可分为两步:先利用模型对文本进行某种学习;再根据学习结果判断生成文本中是否出现幻觉或计算幻觉分数。

3.2.1 基于模型信息提取的评估方法

利用模型进行信息提取 (IE) 的评估方法主要采用信息提取模型, 将生成文本和参考文本中的知识以某种方式 (如关系元组) 表示出来, 并进行比较验证, 从而评估幻觉, 如图 2 所示. 由于传统的 OpenIE^[39]方法无法利用知识库信息, 导致提取结果难以比较, Goodrich 等人^[40]比较了两类信息提取方法: 一类是两步提取法, 即先提取句子中所有的命名实体, 再将每一对实体的关系分类; 另一类是基于 Transformer 架构的端到端直接提取事实元组的方法, 该方法可以避免多步方法的错误组合.

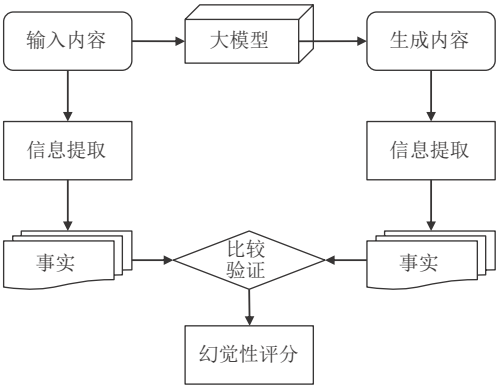


图 2 模型信息提取方法的评估流程图

虽然端到端模型的信息提取方法具有良好的效果, 但是以上两类方法在识别元组时仍可能出现错误. 由此, Nan 等人^[41]提出基于命名实体识别 (named-entity recognition, NER) 模型的评估方法. 该方法利用目标文本和生成文本中提取出的命名实体的匹配结果, 计算精确率与召回率, 从而得到 $F1$ 度量, 显示出更优的鲁棒性. Lee 等人^[42]也将此方法扩展应用于开放式文本生成任务的幻觉评估中. 此外, 在利用信息提取结果量化幻觉程度方面, 还存在多种度量指标. Dušek 等人^[43]使用了误时隙率; Wang^[20]则使用了内容选择、关系生成和内容排序等指标来度量幻觉.

基于 IE 的评估方法在度量指标上更为灵活和多样, 主要用于对事实矛盾幻觉、开放域幻觉进行评估, 但受限于模型本身的局限性, 可能在信息提取中出现错误, 并可能在下游任务评估中传播而影响评估精度.

3.2.2 基于模型推理的评估方法

考虑到幻觉问题研究早期所面临的带标签数据集缺失情况, 相关研究探索了基于模型推理的评估方法. 此类方法认为源知识参考应蕴含无幻觉生成结果的全部信息, 以自然语言推理 (natural language inference, NLI) 为基础, 通过判断假设 (生成文本) 与前提 (参考文本) 的关系 (蕴含、矛盾或中立) 来检测幻觉, 并用生成文本与参考文本蕴含、中立和矛盾次数的百分比, 即蕴含概率 (entailment probability) 量化幻觉^[16], 是当前主流的幻觉自动评估方法之一 (如图 3 所示).

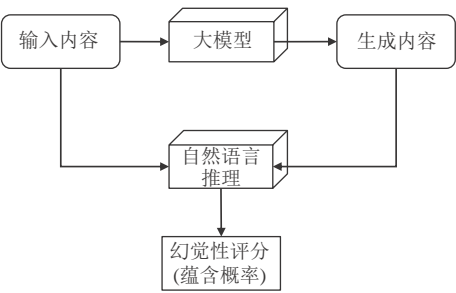


图 3 模型推理方法的评估流程图

Dušek 等人^[44]将此方法应用于数据到文本的生成任务中,并从两个方向检测输入数据与输出文本间的文本蕴含 (textual entailment) 情况:从输入数据中推断输出来检查幻觉情况,并从相反方向检查遗漏情况,该方法具有很高的精度.在此基础上,衍生出基于 MNLI (MultiNLI)^[45]、ANLI (Adversarial NLI)^[46]等不同数据集训练的 NLI 模型来评估蕴含情况的多种评估方法.

以上的研究集中在句子层面的推理,限制了 NLI 方法在下游任务的应用范围. Laban 等人^[47]为了将评估的级别从句级提升到文档级,设计了 SummaC_{Conv} 评估方法,该方法通过以句子为单位分割文本并聚合句子对之间的分数以评估幻觉,适用摘要生成的下游任务;而 Kryściński 等人^[48]提出 FactCC 方法,该方法可以直接从文档级计算分数;Yin 等人^[49]则构建了文档级的 NLI 数据集 DocNLI,在 DocNLI 预训练的模型在流行的句子级基准上也显示出很好的性能,并且可以很好地推广到依赖于文档粒度推理的域外 NLP 任务.此外,上述方法常用的 NLI 模型主要有 BERT^[50],以及在此基础上发展出的 RoBERTa^[51]、DeBERTa^[52]等.

针对某些 NLI 只能返回幻觉分数而难以精确定位文本幻觉所在的缺陷, Goyal 等人^[53]提出了依存级的 NLI 方法,通过关注输入和输出文本中的依存弧所表示的语义关系来检测并定位幻觉.另外,值得注意的是 NLI 模型的泛化能力也不够优秀^[54,55],因此需要针对不同的任务改进现有的 NLI 范式,例如使用合适的数据集对 NLI 模型进行微调^[56,57]等. Barrantes 等人^[57]针对摘要任务分别用 MNLI 和 ANLI 数据集对 BERT、RoBERTa 和 XLNet^[58]模型进行微调,通过人类和模型对抗方式生成的更具挑战性的数据可以帮助提高模型的推理能力,从而提高 NLI 方法的准确性; Dziri 等人^[59]针对 KGD 任务利用 NLI 范式的扩展构建 BEGIN 基准,将中立关系分为幻觉、与主题不相关和一般 (过于模糊而无法归类) 这 3 个子类,通过对抗性生成数据改进了现有的评估指标,提高了评估方法的泛化能力.

基于模型推理的评估方法适用于包含明确上下文的内在幻觉与封闭域幻觉.此外,该方法对于各种类型的矛盾幻觉,如输入矛盾幻觉、上下文矛盾幻觉以及事实矛盾幻觉均有较好的评估效果,但局限性在于如何扩展评估指标以提高其泛化性,以及模型推理能力的提升.

3.2.3 基于特定模型的评估方法

一类基于特定模型的方法通过使用两个在不同数据集上训练的模型 (条件语言模型和无条件语言模型) 来判断生成文本中每一个词例是否得到参考文本的支持.其中无条件语言模型 LM 只在目标文本上开展训练;而条件语言模型 LM_x 则增加了源文本开展训练. Filippova^[60]利用 LM_x 计算单一词例的损失关系,由此判断该词例是否出现幻觉.而 Cao 等人^[61]则利用两个模型来分别计算先验和后验概率,以此作为特征进行分类来评估幻觉情况. Yu 等人^[62]将此思路应用到提出的 KoLA 基准中,通过设计自我对比指标来度量幻觉,让同一模型分别在只根据上下文和加上先验知识 (人工编写的参考文本中的知识) 的情况下生成文本并计算其相似度,相似度越高说明模型抗幻觉能力越强.

另一类方法则通过模型直接计算某种分数来量化幻觉程度. Deng 等人^[63]通过引入信息对齐概念,训练自监督模型在词例维度上对幻觉进行度量,该方法在各种任务 (包括文本摘要、风格转换和基于知识的对话) 中表现出与人类判断相比更强或相当的性能. Zha 等人^[64]根据从文本到表对齐级别标签的映射,提出了对齐分数 (AlignScore) 评估方法,该方法能够评估任意两文本片段间的信息对齐模型,适用于多个场景的幻觉评估.特别对于处理长文本以及解决输入内容和生成文本不同的情形,提出将输入内容和生成文本分别拆分为粗粒度块和细粒度句子的拆分策略,通过聚合输入块和生成句子间的对齐分数得到最终的幻觉评估分数,该方法的效果不仅优于基于 GPT-4 的评估指标,同时减小了需要评估指标的数量集. Yue 等人^[65]设计了归因分数来评估模型生成结果的幻觉,分别定义了从查询、答案和参考到归因分数的映射,同时将归因错误分为矛盾错误和外推 (extrapolatory) 错误.通过提示大语言模型和微调较小的语言模型两种方式对问题回答、事实检查、自然语言推断和摘要生成等任务进行了评估,显示出较好的效果. Zhong 等人^[66]提出了一种多维评估器 UniEval,将不同维度的评估转换为布尔问答 (Boolean)^[67]问题,与其他常用的度量标准如 BERTScore^[68]、BARTScore^[69]、USR^[70]相比,该方法在文本摘要和对话响应任务中表现更佳,并对未知任务展示了强大的零学习能力. Wei 等人^[71]提出加权大模型的评估方法 FEWL,利用现成的 LLM 答案作为黄金标准答案的代理,设计 CHALE 可控指标和 Truthful-QA 指标共同评估模

型的幻觉程度. Su 等人^[72]提出无监督的训练框架 MIND, 通过在 LLM 的推理过程中基于每个令牌的上下文嵌入构建多层感知器模型, MIND 可以在实时过程中进行幻觉检测, 并有效减少计算开销和检测延迟. Zhang 等人^[73]提出了用于估计文本生成质量的双阶段可解释评估方法 DEE, 该方法基于 Llama 2 模型在初始阶段有效识别生成文本中的错误, 随后在第 2 阶段提供针对文本错误的全面诊断报告, 从而实现了评估效率、多样性与可解释性的有效结合.

基于特定模型方法的评估效果依赖于所使用模型的性能, 且其评估范围并不受限于幻觉的类型, 而更多与评估场景有关, 因此具有较大的普适性. 如何保证模型知识的实时更新以及训练数据集的完备对于此类评估方法极为重要.

3.3 基于多任务应用的评估方法

为了从应用层评估大模型的幻觉性问题, 一些研究提出基于多任务应用对大模型生成内容的评估方法. 这些方法按照任务种类可分为 3 类: 基于问答的评估方法、基于分类的评估方法和基于特定任务指令的评估方法.

3.3.1 基于问答的评估方法

基于问答 (question answering, QA) 的评估方法通过生成问题与答案的方式对模型幻觉水平开展评估. 该方法基于一种合理的假设, 即如果生成结果与参考文本事实一致, 就会从相同的问题中生成相似的回答 (如图 4 所示). 因此, 通过对答案进行验证便可以隐式地度量生成文本的幻觉水平.

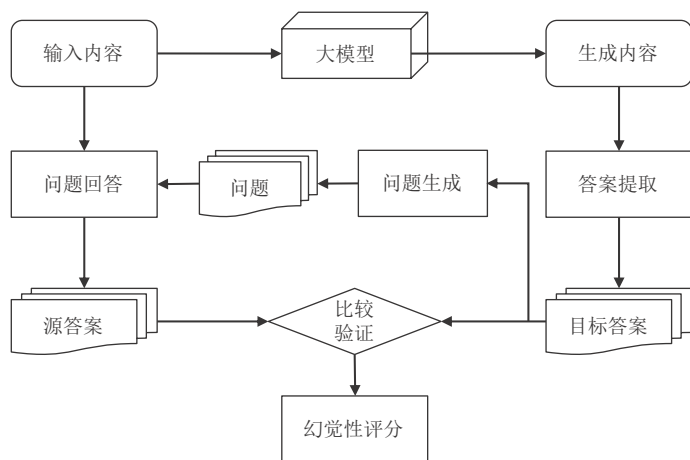


图4 问答评估方法流程图

基于问答的评估方法多用于摘要任务的幻觉评估^[31,74-76]. Durmus 等人^[31]提出了 FEQA 方法, 先屏蔽生成的摘要中重要的文本跨度并以此为答案生成问题, 再让 QA 模型根据源文本对此进行回答. Wang 等人^[74]提出了 QAGS 方法, 利用 QG 模型生成关于摘要的问题并对质量较低的问题加以过滤, 再让 QA 模型根据源文本和生成的摘要分别进行回答. Nan 等人^[75]克服 QAGS 计算代价高的缺陷, 提出了 QUALS, 并使用 QAGen^[76]生成“问题-答案对”的平均对数概率, 进而量化幻觉. Scialom 等人^[77]提出了 QuestEval 评估方法, 该方法同时利用摘要生成问题来计算精确度, 并利用源文本生成问题加权计算召回率, 最后结合精确率和召回率计算 F -度量量化幻觉, 在一致性、连贯性、流畅性和相关性上提高了与人类判断的评估相关性. Rebuffel 等人^[78]为满足数据到文本生成任务的评估需求, 将 QuestEval 方法改进为 Data-QuestEval, 该方法利用 QA 系统生成一组以源文档为条件的相关问题, 然后在其生成的摘要中询问这些问题, 其评估方式为: 如果 QA 系统提供的答案是正确的, 则认为摘要生成与其源文件一致. Yehuda 等人^[79]利用模型初始查询生成的答案对, 通过量化原始查询和重构查询之间的一致性水平, 确定模型生成的答案是否具有幻觉. Vu 等人^[80]提出了一种动态 QA 评估基准 FreshQA, 该基准提供了两种模式的评估程序: 首先利用 RELAXED 评估主要答案的正确性, 进一步利用 STRICT 评估答案中每个事实的准确性.

Zhang 等人^[81]提出一种利用语义感知交叉检查一致性 (Sac3) 来改进黑盒模型中幻觉的检测方法, 该方法利用 HotpotQA-halu 和 NQopen-halu 数据集对问题采样, 然后使用 GPT-3.5turbo 生成对应的答案, 最后对事实真相和相关知识来源进行比较检查。

基于 QA 的方法主要应用于摘要生成任务中生成与事实不一致的矛盾幻觉、外在幻觉以及开放域幻觉评估, 此类方法存在与基于 IE 的方法类似的缺陷, 不仅使用的模型可能存在潜在错误, 并且可能会向下游传播。

3.3.2 基于分类的评估方法

基于分类的评估方法主要在两方面加以改进: 一方面, 关于评估方法泛化能力较差的问题, 基于分类的方法构建了针对特定任务的数据集; 另一方面, 考虑到评估中的蕴含/中立与大模型的忠实并不完全相同, 基于分类的方法对于数据集做了关于忠实的分类, 更加贴合幻觉评估任务的要求。

维基百科向导 (wizard of Wikipedia, WoW)^[82]是 Dinan 等人针对 KGD 任务构建的分类数据集。Santhanam 等人^[83]对该数据集做了修改, 在该数据集人工回复的基础上增加了自动生成的事实不一致的回复, 提出了 ConvFEVER 方法, 用于检测事实不一致的模型。Honovich 等人^[55]则通过人工注释在 WoW 验证集上得到随机样本的事实一致性来构建检测的数据集, 进一步对幻觉进行评估。

Laban 等人^[47]针对摘要任务提出了 SUMMAC 基准, 通过对数据集进行标准化处理, 将其转化为含一致或不一致标签的二元分类数据集以对模型进行评估。Utama 等人^[84]提出了 Falsesum, 利用一个可控的文本生成模型来干扰人工注释的摘要, 通过设置可切换的输入控制代码允许生成含不同类型的输出以评估幻觉。此外, Cao 等人^[61]从另一角度提出了基于分类的幻觉检测法, 通过关注实体并建立实体幻觉与生成概率的联系, 根据预训练和经微调的掩码语言模型分别计算实体的先验和后验概率作为特征并进行分类来检测幻觉。Jiang 等人^[85]通过构建残差流到词汇空间的映射, 统计正确情况和幻觉情况之间输出令牌概率的动态变化, 对由已知事实引起的幻觉进行预测分类。与传统的基于分类的方法相比, 该方法能够更好地捕捉输出令牌概率的动态变化信息, 通过比较正确和幻觉情况下的概率变化差异提高幻觉预测的准确性。此外, 其专门针对已知事实产生的幻觉进行检测, 具有更明确的问题导向和应用场景。

基于分类的评估方法需要明确所验证数据集的幻觉类别。因此, 建立一个更加细致的幻觉分类体系对于该类评估方法的应用至关重要。

3.3.3 基于特定任务指令的评估方法

除了上述方法外, 还有研究者采用基于特定任务指令的方法开展幻觉评估。Li 等人^[86]提出了大模型幻觉评估基准 HaluEval, 该方法使用指令调整数据集来生成回复并自动生成样本数据以评估幻觉。Lin 等人^[87]提出了评估基准 TruthfulQA。通过精心设计的 817 个指令问答问题, 测试了 GPT-3、GPT-Neo/J、GPT-2 等模型的幻觉问题, 将评估扩展到金融、法律、健康等领域。Tang 等人^[88]、Chen 等人^[89]、Zhu 等人^[90]分别针对摘要任务、对话任务以及真实世界的任务下提出幻觉评估基准 TofoEval、DiaHalu、HaluEval-Wild, 其利用不同任务领域的指令生成参考答案对幻觉进行评估。此外, Chen 等人^[91]提出评估基准 FELM, 基于 ChatGPT 对世界知识、科学和技术、数学、写作和推荐以及推理等多个领域的事实错误进行评估。Yang 等人^[92]提出的 PHD 基准采用了基于反向验证的零资源方式自动检测事实错误, 在段落检测生成任务中展现出较高的效率和较优的成本效益。Lattimer 等人^[93]提出的评估基准 ScreenEval, 基于包含脚本和人工摘要的 SummScreen 数据集, 为 Longformer 和 GPT-4 在句子级别生成的摘要引入了事实不一致注释, 将评估范围扩展到长篇对话的幻觉问题。Muhlgay 等人^[94]基于语料库变换提出了 FACTOR 评估基准, 该方法自动将感兴趣的事实语料库转换为基准, 评估 LLM 生成的相似但不正确的陈述的倾向。Mündler 等人^[95]使用 aLM (analyzer LM) 来检测幻觉, 其使用零样本提示并结合思维链提示技术^[28]将 aLM 实例化, 让 aLM 在得出检测结果前先提供一个解释, 然后计算标准分类精度 (P)、召回率 (R) 和 F1 得分来评估幻觉。以上基于特定任务指令评估的方法适合于对单一任务在特定数据集下开展的评估, 具有一定的应用局限性。

相关研究者通过设计任务提示指令, 利用大模型对不同下游任务幻觉进行评估。Kocmi 等人^[96]通过对翻译任务的研究发现大模型是最先进的翻译质量评估工具, 提出了基于 GPT 的翻译质量评估指标 (GEMBA), 设计了任务提示指令对模型翻译质量进行评估。Gao 等人^[97]依据大语言模型学习上下文的能力, 通过指令调整帮助其与人

类评估行为对齐,使大模型在摘要任务中能够模仿人工评估. Liu 等人^[98]提出了 G-Eval 评估方法,该方法包含评估任务定义和标准的提示,并利用具有思维链的大模型框架根据返回词例的概率计算分数的评分函数,在文本摘要任务和对话生成任务评估中具有优异的表现. Min 等人^[99]针对长文本评估提出了 FActScore 方法,该方法先利用大模型将生成文本分解为一系列传达单一信息的短句,即原子事实 (atomic fact),再利用模型自动验证原子事实是否被可靠知识源支持以评估幻觉得分. Shafayat 等人^[100]在 FActScore 基础上提出 Multi-Fact 方法,专门用于多语言传记生成任务的真实性评估. 通过该项研究,发现了 LLM 事实生成中的地理偏见与多语言评估的必要性. Gekhman 等人^[101]针对摘要问题提出了 TrueTeacher,通过使用不同的模型生成一组候选摘要,再利用大模型对其事实一致性标注,该方法可以有效缓解数据集难以涵盖所有类型的事实错误的挑战.

利用大模型设计任务指令的幻觉评估依赖于对应的下游任务类型及任务要求,方法评估效率与准确性受限于大模型本身的能力,但对于缺乏人类参考的新任务评估具有较大的应用前景.

3.4 小 结

本节从数据、模型和多任务应用 3 个角度梳理了目前主流的评估方法,并对不同方法的原理进行了介绍和分析,相关的研究总结见表 1.

表 1 大模型幻觉评估方法分类

大类	子类	实例	评估类型
基于数据文本的 评估方法	目标文本	PARENT ^[32] , Manakul ^[33] , Wu 等人 ^[34]	内在幻觉
	源文本	PARENT-T ^[35] , Knowledge F1 ^[36]	封闭域幻觉
	文本扩展	Popović 等人 ^[37] , Martindale 等人 ^[38]	上下文矛盾幻觉
基于模型的 评估方法	信息提取	Goodrich 等人 ^[40] , Nan 等人 ^[41] , Lee 等人 ^[42] , Dušek 等人 ^[43] , Wang 等人 ^[20]	事实矛盾幻觉 开放域幻觉
	推理	Dušek 等人 ^[44] , Laban 等人 ^[47] , Kryściński 等人 ^[48] , Yin 等人 ^[49] , Goyal 等人 ^[53] , Barrantes 等人 ^[57] , Dziri 等人 ^[59]	内在幻觉 封闭域幻觉 输入矛盾幻觉、上下文矛盾幻觉、事实矛盾幻觉
	特定模型	两个模型: Filippova ^[60] 等人, Cao 等人 ^[61] , Yu 等人 ^[62] 分数计算: Deng 等人 ^[63] , AlignScore ^[64] , AttributionScore ^[65] , UniEval ^[66] , BERTScore ^[68] , BARTScore ^[69] , USR ^[70] , Wei 等人 ^[71] , Su 等人 ^[72] , Zhang 等人 ^[73]	不受限于任何幻觉类型 与评估场景相关
基于多任务应用的 评估方法	问答	FEQA ^[31] , QAGS ^[74] , QUALS ^[75] , QuestEval ^[77] , Data-QuestEval ^[78] , Yehuda 等人 ^[79] , FreshQA ^[80] , Sac3 ^[81]	外在幻觉 开放域幻觉 事实矛盾幻觉
	分类	数据集/基准方面: Conv-FEVER ^[83] , Honovich 等人 ^[55] , SUMMAC ^[47] , Falsesum ^[84] 分类方法: Cao 等人 ^[61] , Jiang 等人 ^[85]	不受限于任何幻觉类型 与分类标签相关
	特定任务	基准: HaluEval ^[86] , TruthfulQA ^[87] , TofoEval ^[88] , DiaHalu ^[89] , HaluEval-wild ^[90] , Felm ^[91] , PHD ^[92] , ScreenEval ^[93] , FACTOR ^[94] 方法: aLM ^[95] , GEMBA ^[96] , Gao 等人 ^[97] , G-Eval ^[98] , FActScore ^[99] , Multi-Fact ^[100] 数据集: TrueTeacher ^[101]	依赖于下游任务类型 及任务需求

基于以上分析,我们对比分析了 3 类方法的优点和局限性 (见表 2). 具体分析如下.

(1) 基于数据文本的评估方法简单直接,能够根据数据源设计相应的评估指标量化模型输出的幻觉程度,评估指标也可以根据评估精度进行调整,适用于对内在幻觉、封闭域幻觉以及上下文矛盾幻觉进行评估,易于操作和理解. 但此类方法也存在一些局限性,由于大多数基于数据文本的统计指标无法理解生成文本的含义,只能使用其中的字符和单词来统计幻觉,导致此类方法只能处理词汇信息而难以从句法和语义层面开展评估. 此外,基于数据文本进行评估也存在精确度不足的情况.

(2) 基于模型的评估方法可以克服数据文本不足、标签缺失的情况,同时适用于词汇和长文本幻觉评估,且对

各种幻觉类型均有较好地评估效果。相对于数据文本的评估方法, 模型评估方法更为精确, 评估指标设计与评估数据集设计更为灵活, 但也存在指标设计较复杂, 需从多角度量化幻觉问题的局限性。

(3) 基于多任务应用的评估方法易于对特定任务进行评估, 适用于任务种类较单一的模型。然而, 该方法一般利用模型生成数据集, 并进行对比评估, 容易造成模型的潜在偏误传播, 且对于幻觉评估而言可靠性不足。

表 2 大模型幻觉评估方法对比

分类	优点	局限性
基于数据文本的评估方法	简单、直接 易于操作和理解	长数据文本的理解较差 评估精确度差
基于模型的评估方法	克服数据文本标签缺失情况 适用于词汇和长文本幻觉评估	评估指标设计较复杂 无法完全量化幻觉问题
基于多任务应用的评估方法	易于评估特定的下游任务 适用于任务种类较单一的模型评估	易造成模型潜在错误传播 幻觉评估的可靠性不足

4 大模型幻觉的缓解方法

根据幻觉的主要来源, 常见的幻觉缓解方法可以分为 3 类: 数据层幻觉缓解方法、模型层幻觉缓解方法以及应用层幻觉缓解方法。下文将从方法原理、研究进展及优缺点分析等方面梳理这类模型幻觉的缓解方法。

4.1 数据层幻觉缓解方法

4.1.1 数据收集

考虑到不忠实的数据会引发模型幻觉, 学术界和产业界已提出多种人工构建忠实数据集的方法, 该方法可以缓解启发式数据收集方法不准确、不符合事实等问题产生的幻觉。一般常用于缓解内在幻觉、封闭域幻觉、输入矛盾幻觉和上下文矛盾幻觉。一种方法是从现有的知识库编写忠实的语料库。例如, Gardent 等人^[102]提出从现有知识库中半自动地创建“数据到文本”语料库的新框架, 创建过程主要包含: (1) 内容选择模块: 用于从数据库提取不同的、相关的且类型一致的数据单元。(2) 众包过程: 将数据单元与人类撰写的文本联系起来, 从而正确地捕捉到它们的意义。

另一种方法是重写来自现实数据的真实句子, 修改策略主要包括^[23]: (1) 短语删除: 删除例句中数据源不支持的短语; (2) 去语境化: 识别句子的主要主题, 解析共同引用, 并用源中的命名实体替换依赖于上下文的短语; (3) 语法修改: 使修改后的句子更加正确流畅。此外, 针对“数据到文本”相关任务, 还有研究提出了修改现有数据集的内容来提升数据保真度的缓解方法^[20]。

一些研究利用模型生成数据, 并标记这些输出是否包含幻觉^[55,103,104]。例如, Gabriel 等人^[103]对生成的细粒度事实错误的摘要进行注释, 以创建一个事实一致性诊断数据集。Honovich 等人^[55]通过自动问题生成和问答, 并使用自然语言推理比较答案范围, 进而确保基于知识的对话中的事实一致性。Dziri 等人^[104]通过编辑 WoW 基准测试中的幻觉反应, 创建用于信息检索对话的忠实数据集 FaithDial。由此可见, 利用模型生成数据的相关方法通常被用于构建幻觉评估数据集, 同时也可作为忠实数据集以降低模型幻觉。虽然这种方法比从重新收集数据的成本要低, 但它仍然需要耗费大量的人力和资源。

此外, 其他方法也被用于构建忠实的数据集^[105,106]。例如, Cheng 等人^[105]构建了一个用于问答系统和自然语言生成的层次表数据集 HiTab, 其中模型基于一组选定的单元格和运算符生成句子, 从而提升模型生成语言的忠实性和逻辑性。Chen 等人^[106]构建了涉及常见逻辑类型的大型数据集 Logic2Text, 用于从逻辑表生成高可信度的语言。其构建过程主要分为 3 个步骤: (1) 描述的组成和验证; (2) 逻辑表的注释和推导; (3) 逻辑表的执行和验证。为了提升模型的诚实性, 使其在面对超出能力范围的问题时能够坦诚地承认无法作答, 一种可行的解决方法是在数据收集阶段中引入一些体现诚实态度的样本。具体而言, Sun 等人^[107]在 Moss 项目的数据集中引入了诚实的样本, 依据其调整模型学会了在适当的情况下拒绝回答问题, 从而有助于减少幻觉现象的发生。

综上所述,通过数据收集提升训练数据集质量,是缓解大模型幻觉的一种直观且有效的方法.然而,数据收集方法一般是基于特定任务的,其可能缺乏泛化性,导致大模型在其他任务中仍存在幻觉性.

4.1.2 数据预处理

数据预处理方法主要包括数据清洗和数据增强两类方法,通过处理训练数据增强模型在输入和输出之间的对齐,一般常用于缓解内在幻觉、封闭域幻觉、输入矛盾幻觉和上下文矛盾幻觉.

4.1.2.1 数据清洗

数据清洗方法可用于解决由于数据重复或缺失、数据标注等问题引发的幻觉.在训练数据的构建过程中,难免会引入语义噪声,例如输出中的一些短语不能用输入来解释^[60].Raunak 等人^[108]研究表明,部分幻觉类型可以通过特定的语料库级噪声模式产生和解释,该模式决定了由模型产生的幻觉类型.为了缓解语义噪声问题,一些研究提出采用数据清洗手段过滤数据集中可能导致模型幻觉的样本,该方法适用于原始数据中存在低或中等噪声的情况^[60,109].例如,一种方法是通过幻觉评估指标,从现有的语料库中找到与输入无关或矛盾的信息,从而对数据过滤或修正^[60].一般而言,语料库过滤方法包括下列步骤.

- (1) 利用模型幻觉的评估指标,根据幻觉水平衡量训练样本的质量;
- (2) 将训练样本的幻觉性评分按降序排列;
- (3) 根据样本的幻觉性排序,选择并过滤出不可靠的样本.

一些工作已经在实例层面缓解了幻觉产生的问题,Liu 等人^[110]通过对每组“源-目标”使用一个评分,并过滤掉有幻觉产生的实例,从而降低模型的幻觉性表现.Shen 等人^[111]提出一种基于贝叶斯优化的数据过滤方法,通过线性组合多种属性的质量度量,量化实例的幻觉水平.

然而,实例级的过滤可能会导致信号丢失,其原因在于一些情况下幻觉发生在单词级,即目标句子的某部分忠于源输入,而其他部分则可能存在幻觉^[112].针对信号丢失问题,一些研究提出根据参考信息修正训练样本与模型输入数据^[109,113].例如,Nie 等人^[109]将一个用于数据细化的语言理解模块与自训练迭代集成起来,从而有效地诱导输入数据和对对应文本之间的强等价性,进而减少模型的幻觉.该方法的数据清洗步骤主要通过迭代重新标记过程实现,这种修正输入意义表示的方法可以增强输入和输出之间的语义一致性,而不需要删除部分数据集,从而缓解了信号丢失的问题.

针对具体任务,数据清洗有如下的应用实例.在文本摘要任务中,Nan 等人^[41]提出首先在参考摘要上应用 Spacy^[114]方法识别所有命名实体,并从摘要删去无法在源文档中找到匹配内容的句子,以确保数据集中没有幻觉.针对机器翻译任务,通过删除无效的“源-目标”对,使用启发式或滤波器的语料库级噪声过滤^[115,116]也能有效减少机器翻译中的幻觉.

4.1.2.2 数据增强

数据增强方法可用于缓解由于数据收集、数据标注等问题引发的幻觉.不同于数据清洗方法,数据增强技术主要基于有限的数据生成更多有效的数据,通过丰富数据的分布,提升模型泛化能力.针对大模型的幻觉性问题,由于外部知识、显式对齐和额外的训练数据等外部信息可以提高源与目标之间的相关性,帮助模型更好地学习与任务相关的特征,因而可通过增强输入数据的方式使其获得更好的源表示,进而缓解模型的幻觉问题.

在实体信息层面,Liu 等人^[110]提出合并辅助的实体信息增强模型训练过程.在事实描述提取层面,Cao 等人^[14]通过从源文档中提取的事实描述(如主谓宾关系三元组)增强数据,为摘要总结提供清晰且正确的指导.在预执行操作结果层面,Nie 等人^[117]提出通过预先操作输入数据提取信息,并将结果作为从输入数据中推断出的事实来指导生成.在合成数据层面,Wang 等人^[118]通过替换或扰动合成数据以增强负样本,并训练了一个受控生成模型.该模型可基于忠实性控制代码生成忠实或不忠实的摘要.Longpre 等人^[119]提出了知识冲突现象,即上下文信息与模型在训练过程中学习到的信息不符.

此外,为了避免模型过度依赖于历史知识,可通过替换修改语料库的训练示例来增强训练集.Chen 等人^[120]提出对比候选生成和选择模型.这是一种与模型无关的幻觉缓解技术.其中,对比候选生成步骤(数据增强)将生成摘要中的命名实体和数量替换为源文档中具有兼容语义类型的实体和数量.Biton 等人^[121]通过采用不同的采样策

略来选择图像说明中要替换的对象, 从而增强原训练数据. 在外部知识检索层面, Bi 等人^[122]提出允许模型利用不在输入中的相关外部知识, 并利用外部检索的知识增强模型生成内容的真实性.

针对具体任务, 数据增强有如下应用实例. 在对话生成系统中, Shuster 等人^[36]提出一种通过检索增强的架构, 其中对话基于检索到的知识生成. 针对文本摘要任务, Chen 等人^[120]通过替换或扰动生成增强数据, 提升了文本摘要的可信度.

综上所述, 数据预处理方法可以在数据层对大模型幻觉性问题进行有效缓解. 其中, 数据清洗方法从现有的并行语料库中找到与输入无关或矛盾的信息, 然后对数据进行过滤或修正, 但这种方法可能会造成信号丢失等问题. 数据增强方法强制要求在输入和输出之间进行更强的对齐. 然而, 原始信息源和增强信息之间的差距, 如语义差距和格式差异, 将增加数据增强方法的应用挑战.

为进一步清晰展示各种幻觉缓解方法的应用场景, 表 3 中将数据层幻觉缓解方法和幻觉产生的机理类型与任务类型相关联, 相关方法也被作为实例总结在表 3 中.

表 3 数据层幻觉缓解方法分类

幻觉成因	数据层幻觉缓解方法	任务类型	实例
数据收集问题	数据收集	摘要总结	Gabriel 等人 ^[103]
		对话	Honovich 等人 ^[55] , Dziri 等人 ^[104] , Sun 等人 ^[107]
		问答	Cheng 等人 ^[105]
		文本生成	Gardent 等人 ^[102] , Parikh 等人 ^[23] , Wang 等人 ^[20] , Chen 等人 ^[106]
数据重复或缺失 数据标注问题	数据清洗	摘要总结	Nan 等人 ^[41]
		对话	Shen 等人 ^[111]
		文本生成	Filippova ^[60] , Nie 等人 ^[117] , Liu 等人 ^[110] , Rebuffel 等人 ^[112] , Dušek 等人 ^[113]
		机器翻译	Raunak 等人 ^[108] , Junczys-Dowmunt 等人 ^[115] , Zhang 等人 ^[116]
数据收集、数据标注问题	数据增强	摘要总结	Cao 等人 ^[14] , Wang 等人 ^[118] , Chen 等人 ^[120]
		对话	Shuster 等人 ^[36]
		问答	Longpre 等人 ^[119] , Bi 等人 ^[122]
		文本生成	Liu 等人 ^[110] , Nie 等人 ^[117]
		视觉-语言	Biten 等人 ^[121]

4.2 模型层幻觉缓解方法

4.2.1 模型结构

对大模型架构, 特别是编码器和解码器部分进行修改, 有助于缓解由于编码与解码缺陷等问题引发的模型幻觉. 下文将从编码器和解码器两个角度分别介绍相关的模型层幻觉缓解方法.

4.2.1.1 编码器

对文本进行编码是机器理解语言的核心. 在语言模型中, 编码器将一个可变长度的序列从输入文本编码为一个固定长度的向量表示. 由于模型对输入缺乏语义解释, 进而导致幻觉产生, 因此一些工作从模型角度出发, 通过修改编码器架构, 使其与输入文本更兼容. 改进编码器的结构可以使模型学习更好的文本表示, 使其对文本间的内在联系有更好的理解, 进而缓解大模型的幻觉性问题. 因此, 基于编码器的方法一般常在缓解内在幻觉、封闭域幻觉、输入矛盾幻觉和上下文矛盾幻觉类型中展现出较强的适用性. 例如, Lim 等人^[123]提出了一种基于多编码器的候选评分系统, 包括候选项编码器和上下文编码器等部分, 用于捕获上下文输入和候选项之间的语义相似性, 进而在知识选择器和角色选择器中选取最合适的源.

针对具体任务, 编码器优化方法有如下的应用实例. 在文本摘要任务中, Zhu 等人^[124]使用显式图神经网络对从源文档中提取的事实元组进行编码. 在对话生成系统中, Wang 等人^[35]采用了掩码语言模型 (MLM) 任务^[125]推

断输入词是否被正确编码,该任务可以提高对整体输入的建模能力,并给出准确和完整的表示。

综上所述,改进的编码器可以提升对输入文本的理解,并将其编码到有意义的表示。提升编码器的理解能力,并避免其对训练数据的不同部分之间产生错误相关性,能够有效缓解大模型的幻觉性问题。

4.2.1.2 解码器

在语言模型中,解码器将一个固定长度的向量表示(自然语言的编码)转化为可变长度的文本序列,通常用给定输入表示的自然语言生成最终输出。如前文所述,错误的解码会导致模型幻觉。因此,众多研究致力于修改解码器结构以减轻幻觉。由于基于解码器的幻觉缓解方法直接作用于输出端,这种方法不受限于任何幻觉类型,而是与缓解场景高度相关。

Liu 等人^[126]提出一种用于受控文本生成的解码方法,它将预先训练的语言模型与“专家”和“反专家”语言模型相结合。Rebuffel 等人^[112]提出多分支解码器,其中每个控制因素(即内容、幻觉或流畅性)均通过单一的解码模块进行建模,最终的输出表示可以根据各分支期望的重要性进行加权,从而减轻推理步骤中的幻觉。考虑到幻觉概率与预测的不确定性呈正相关,Xiao 等人^[127]提出一种不确定性感知解码器,通过在模型解码过程中惩罚预测的不确定性来减少生成语言的幻觉。Song 等人^[128]提出一种由序列解码器和基于树的解码器组成的双解码器。其中序列解码器主要用于生成新的摘要词,而基于树的解码器主要用于预测部分摘要单词之间的依赖关系,从而将源序列转换为摘要序列的线性化解析树,可以在改善句子语法的同时提升忠实性。Balakrishnan 等人^[129]提出一种约束解码方法,该解码器使用具有词汇或结构限制的约束,并利用这种表示方式来提高语义的正确性。

这些改进的解码器通过找出标记之间的隐性差异和受显式限制的依赖性,提高了忠实标记的可能性,同时减少了在推理过程中产生幻觉的可能性。但由于改进后的解码器可能更难生成流畅或多样化的文本,因此还需要在大模型的语言生成性能和幻觉水平之间寻求恰当的平衡。

4.2.2 模型训练与微调

在模型训练与微调阶段引入幻觉缓解策略有助于缓解由预训练知识偏好、模型知识更新局限等因素引发的幻觉性问题。这类方法主要用于内在幻觉、封闭域幻觉、输入矛盾幻觉和上下文矛盾幻觉的缓解。下文将从优化损失函数、引入辅助任务两个角度介绍模型层中训练与微调相关的幻觉缓解方法。

4.2.2.1 损失函数

损失函数是用来度量模型的预测值与真实值的差异程度的运算函数,语言模型通常通过最小化损失函数来匹配大尺度语料库的分布特性。正则化则通过在损失函数的末尾添加额外的惩罚项来帮助防止模型过度拟合。Lee 等人^[130]通过实验分析了常用的正则化项在缓解模型幻觉性问题方面的作用。此外,Kang 等人^[131]提出,常用的损失函数(如对抗损失等)虽然易于优化,但这种方法迫使模型重现数据集中的所有变化,包括噪声和无效的引用,从而导致模型幻觉。因此,一些工作通过修改模型训练中的损失函数来缓解幻觉问题。

针对对话生成任务,Yoon 等人^[132]提出引入文本幻觉的正则化损失,该正则化项根据语言模型和幻觉模型之间的互信息求出,最小化该项有助于提高对话性能。针对预训练的视觉-语言模型,Dai 等人^[133]提出对象掩码语言模型,通过屏蔽在图像中出现文本所指对象,该方法可以在生成过程中增强文本标记和视觉对象之间的对齐和限制。为了减轻噪声数据集的影响,Li 等人^[134]提出倾斜的经验风险最小化方法,通过引入倾斜超参数,直接扩展传统的经验风险目标函数,灵活地调整单个损失的影响。

针对机器翻译任务,Wang 等人^[135]提出最低风险训练方法,其目标函数可以表示为关于后验分布的预期损失,从而有效避免传统对数损失函数在训练过程中与模型推断不匹配的问题。针对表到文本生成任务,Wang 等人^[135]通过引入新的表格-文本最优传输匹配损失以及一个表格-文本嵌入相似性损失作为内容匹配约束的方法,提升了生成文本的忠实性。

综上所述,作为一般的训练方法,针对损失函数进行优化,如正则化和损失重建,对幻觉问题的缓解有重要作用。在实际应用中,基于改进损失函数训练的模型可能更难生成流畅或多样化的文本,因此还需要在大模型的语言生成性能和幻觉水平之间寻求恰当的平衡。

4.2.2.2 辅助任务

幻觉问题可能源于训练过程对单个数据集或任务的依赖, 导致模型无法学习到实际的任务特征. 因此, 通过在训练过程中添加适当的额外任务和目标, 可以提升模型的综合性能, 并减少模型幻觉. 基于辅助任务的方法将大模型和事实正确性相关的额外任务结合起来, 以隐式的方式提高大模型性能. 其中, 强化学习和多任务学习是常用的将辅助任务纳入大模型相关任务的方法.

由于词级最大似然训练会导致暴露偏差的幻觉性问题^[136], 一些研究采用强化学习来解决幻觉问题, 如图 5 所示. 强化学习的目的是让智能体学习一个最优策略, 使从环境中积累的奖励最大化, 将实际任务和与正确性相关的辅助任务结合起来, 以提高生成性能. 其中, 奖励函数对强化学习至关重要, 如果设计得当, 它可以提供训练信号, 帮助模型实现其减少幻觉的目标.

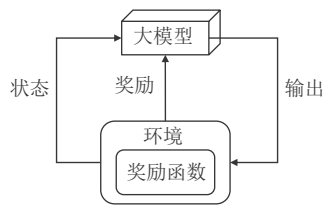


图 5 强化学习方法示意图

Cao 等人^[61]提出一种用于区分实体事实和非事实性幻觉的检测方法, 并将检测器评分当作离线强化学习算法的奖励信号, 从而降低生成内容的幻觉性. Mesgar 等人^[137]通过自然语言推断模型获得人格一致性奖励, 可以量化生成内容和角色之间的一致性以及语义合理性, 以减少幻觉问题. Song 等人^[138]提出基于强化学习的 RCDG 模型, 用于生成角色一致的对话. 类似于生成对抗性神经网络, RCDG 由一个生成器和两个评估器组成, 分别用于估计生成话语的质量和一致性. Wang 等人^[35]扩展了基于注意力的 Transformer 的极大似然损失, 通过度量输入表和输出文本的嵌入向量之间的距离和实体匹配情况, 缓解数据到文本生成的幻觉. 基于这些奖励函数的强化学习方法可以直接优化生成文本的忠实度. 为了提升模型的诚实性, 并使其在超出训练数据范围的任务时能够准确判断自身的能力边界, Schulman 等人^[139]提出了诚实导向的强化学习方法. 其核心理念是鼓励模型表达不确定性, 并通过学习特殊设计的奖励函数促使模型在超出能力范围时予以坦诚承认. 该方法不再需要大量的人工标注数据, 同时也消除了对注释者准确猜测模型知识边界的依赖.

还有一些研究通过多任务学习处理大模型在不同自然语言任务中的幻觉问题, 其方法如图 6 所示. 在多任务学习的训练范式中, 一个共享模型同时在多个任务上进行训练, 以学习不同任务的共性. 对于大模型的多任务学习框架, 一般将在共享权重编码器的基础上构建一个特定于任务的层. 由此, 大模型和辅助不同模型可以具有相同的语义表示, 但具有不同的学习目标.

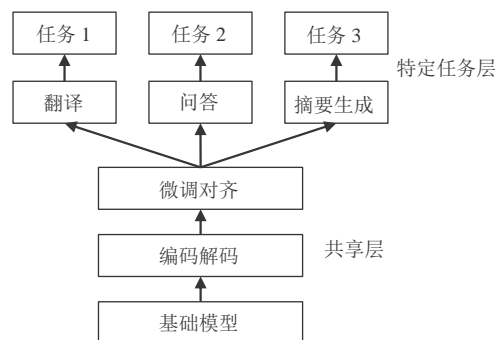


图 6 多任务学习方法示意图

Weng 等人^[140]将单词对齐任务纳入翻译模型, 通过从训练集中抽取一个子集翻译获得误翻译片段. 随后, 在编码器和解码器上采用多任务学习范式指导语言模型正确翻译这些片段, 由此提高输入和输出之间的对齐精度, 从而缓解模型幻觉. Li 等人^[141]将基本原理提取任务整合到语言生成模型中, 在编码过程中将与问题最相关的输入片段作为答案的基本原理. 基于提取的基本原理和原始输入, 解码器可以产生一个高可信度的答案. Wang 等人^[35]提出了具有两个辅助任务的多任务学习范式方法: 首先, 在编码器层面引入掩码语言模型任务, 用于推断输入的词是否被正确地翻译, 以提高对输入的建模和表示能力. 在解码器方面, 引入单词对齐任务来提高编码器, 解码器和交叉注意的对齐精度, 以帮助解码器捕获正确的上下文表示. Nan 等人^[41]提出在文本摘要任务的训练过程中添加判断实体是否值得总结的分类任务, 并采用联合实体和摘要的生成方法, 从而进一步改进了模型在实体级别的事实一致性.

综上, 基于辅助任务的训练方法可以有效缓解大模型的幻觉性问题. 其中, 基于强化学习的训练方法有助于根据训练信号, 帮助模型实现其减少幻觉的目标. 在实际应用中, 其难点在于奖励函数的设计与选取. 多任务学习的训练方法具有提高数据效率、减少过拟合等优点. 在大模型的应用中, 其难点在于共同学习的附加任务的选择. 此外, 同时学习多个任务对模型的设计和优化提出了新的挑战.

4.2.3 后处理输出内容

基于后处理输出内容的方法是一类通用性的幻觉缓解方法. 这类方法可用于缓解多种因素引发的幻觉, 如编码与解码缺陷、预训练知识偏好、模型知识更新局限等. 由于后处理方法直接作用于模型生成的内容, 因此其适用范围广泛, 不受特定幻觉类型的限制. 在对应的缓解场景下, 只需开发相关的后处理方法即可有效地改善模型输出质量. 这种灵活性使得后处理方法成为一个有前景的研究方向. 通过对生成内容的智能化分析和修正, 可以在不改变原有模型架构的情况下, 显著提升其抗幻觉能力和实际应用价值.

前文所述方法需要通过额外的样本构建过程、修改模型结构或训练过程, 以提高生成内容与事实的一致性, 这可能会影响模型的输出性能. 基于后处理的方法通过对生成内容添加校正器以纠正幻觉 (如图 7 所示). 特别是对于在产生幻觉的噪声数据集上训练的模型, 建模校正可以作为缓解幻觉的有效方法. 这种方法一般将生成的内容视为草稿, 并采用删除、重写等方法纠正事实错误, 以形成最终的输出文本. 此过程与人类写作中审查和编辑初稿的过程非常相似, 是常用的后处理方法.

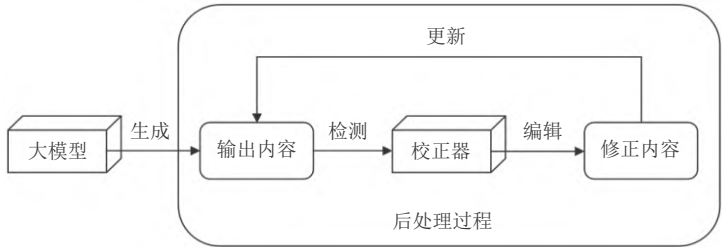


图 7 后处理方法示意图

Cao 等人^[142]提出了后处理校正器模块的方法, 通过识别和纠正生成文本中的事实错误来解决幻觉. 其中, 神经校正器模型的预训练过程在人工实例上进行. 这些训练数据是在参考摘要上应用一系列启发式变换而创建的. Chen 等人^[120]提出一种与模型无关的后处理技术, 通过对比候选生成和选择的后处理方法减轻生成内容的幻觉性. 其通过命名实体和数量替换生成候选摘要学习并生成判别矫正模型, 将生成摘要中的命名实体和数量替换为源文档中具有兼容语义类型的实体和数量. 然后使用该模型选择最佳的备选内容作为最终输出摘要. Dong 等人^[143]提出考虑实体级修正并迭代进行的 SpanFact. 该方法使用两种事实修正模型, 利用从问题回答模型中学到的知识, 通过跨度选择和校正来纠正系统生成的摘要. Song 等人^[144]提出了一个基于后处理的对话模型, 通过生成、删除和重写 3 个步骤改正生成话语中的事实错误: (1) 生成: 正常的文本生成过程; (2) 删除: 识别并删除冲突词; (3) 重写: 通过生成模块恢复被删除的单词. Dziri 等人^[145]对基于知识的对话系统应用了类似的策略, 通过后处理模块对

基于知识图的内容进行重写. 在删除生成文本中潜在的错误实体后, 基于图神经网络的后处理模块在基础知识图中检索正确的实体并进行改进.

综上所述, 作为一个独立模块, 后处理方法可以在保持信息量的同时有效地缓解大模型幻觉. 在实际应用中, 后处理校正步骤可能会导致不符合语法的文本, 但这种方法允许研究人员利用在其他属性 (如多样性、流畅性) 方面表现最佳的模型, 然后通过使用少量训练数据校正正确结果.

4.2.4 (反) 专家模型

另一种常用的幻觉缓解方法是通过专家模型或反专家模型. 这种方法可以用于缓解由模型知识更新局限产生的幻觉. 其中专家模型主要引导模型转向积极行为, 而反专家模型主要引导模型远离消极行为. 专家模型通常基于忠实样本子集训练, 反专家模型通常基于幻觉性的训练数据. 只需在对应的缓解场景收集相关的幻觉数据, 并据此训练专家模型和反专家模型, 就可以将这一方法应用于任何幻觉类型.

Qiu 等人^[146]采用源摘要数据集训练一个基本的模型, 其次用忠实的子集进行微调, 以获得一个专家模型, 并用幻觉的子集进行微调, 以获得一个反专家模型. Ilharco 等人^[147]提出任务向量否定方法. 该方法通过从大模型中减去反专家模型的任务向量来减轻幻觉, 并使用超参数控制二者融合程度. Choubey 等人^[148]认为仅从大模型中减去反专家任务向量会造成模型性能的潜在损失, 因此提出对比参数组合方法, 建议添加专家模型参数, 并使用超参数控制 3 种模型之间融合程度. 与直接操纵模型参数的任务向量否定方法和对比参数组合方法相反, Liu 等人^[126]使用专家模型和反专家模型在每个解码步骤中修改预测内容的分数. 同样, 其通过使用超参数控制解码过程中的融合程度实现对生成内容的调整.

综上, (反) 专家模型是引导模型行为的重要方法, 可以有效缓解幻觉. 在实际部署过程中, 其难点之一是确定不同模型参数之间的融合程度, 从而实现可控生成.

为了更清晰、直观地展示各种幻觉缓解方法的应用场景, 表 4 中将模型层的幻觉缓解方法和幻觉产生的不同机理与任务类型相关联, 相关方法也被作为实例总结在表 4 中.

表 4 模型层幻觉缓解方法分类

幻觉成因	模型层幻觉缓解方法	任务类型	实例
编码缺陷与解码缺陷	模型结构	摘要总结	Zhu 等人 ^[124] , Huang 等人 ^[125] , Song 等人 ^[128]
		对话	Lim 等人 ^[123] , Balakrishnan 等人 ^[129]
		文本生成	Wang 等人 ^[35] , Liu 等人 ^[126] , Rebuffel 等人 ^[112] , Xiao 等人 ^[127] , Lee 等人 ^[44]
预训练知识偏好、模型知识更新局限	模型训练与微调	摘要总结	Kang 等人 ^[131] , Cao 等人 ^[61] , Huang 等人 ^[125] , Nan 等人 ^[41]
		对话	Mesgar 等人 ^[137] , Song 等人 ^[138]
		问答	Li 等人 ^[141]
		文本生成	Kang 等人 ^[131] , Wang 等人 ^[35]
		机器翻译	Lee 等人 ^[130] , Li 等人 ^[134] , Wang 等人 ^[135] , Weng 等人 ^[140]
		视觉-语言	Yoon 等人 ^[132] , Dai 等人 ^[133]
编码与解码缺陷、预训练知识偏好、模型知识更新局限	后处理输出内容	摘要总结	Cao 等人 ^[142] , Chen 等人 ^[120] , Dong 等人 ^[143]
		对话	Song 等人 ^[144] , Dziri 等人 ^[145]
模型知识更新局限	(反)专家模型	摘要总结	Qiu 等人 ^[146] , Choubey 等人 ^[148]
		文本生成	Liu 等人 ^[126]

4.3 应用层幻觉缓解方法

4.3.1 提示工程

提示工程是一种在不更新模型参数的前提下, 通过精心设计输入文本等方式引导大模型的方法, 其旨在指导模型行为, 引导其生成所需的结果. 提示工程方法可以缓解由模型同质化、模型输入提示合理性以及有效性和多样性不足引发的幻觉性问题. 当大模型的参数给定后, 在模型部署应用的过程中, 提示工程是影响模型生成质量的

关键环节. 许多研究通过提示工程方法缓解模型幻觉, 提高输出的忠实性. 由于提示工程的多样性与灵活性, 该方法可以针对任意缓解场景设计相关指令, 不受限于任何幻觉类型.

针对大模型幻觉, Lightman 等人^[149]通过对比根据最终结果提供反馈的结果监督方法与根据中间推理步骤提供反馈的过程监督方法发现, 在解决来自 MATH 数据集的问题时, 过程监督方法明显优于结果监督方法, 且主动学习显著提高了过程监督的有效性. Li 等人^[150]提出了一种基于轮询的对象探测评估方法. 通过轮询方式, 其将图片标注问题转变为一系列单独的分类问题, 即判断特定目标是否存在, 从而避免了指令设计和标题长度带来的偏差. Kumar^[151]认为, 提供足够的上下文可以让模型为专业问题生成相关且准确的内容, 通过在提示工程中加入一系列思维链提示方法, 可提升大模型推理能力. 该研究还提出可以利用行动-观察-思维框架解决综合性工程问题. Zhang 等人^[152]提出过程监督方法可能存在幻觉滚雪球现象, 在此基础上, 通过提示使模型回溯其错误, 可以提高模型在相关任务上的性能. Martino 等人^[153]还提出了知识注入技术, 通过将任务相关的实体的上下文数据从知识图映射到文本空间, 并将其包含在模型输入的提示符中以缓解模型幻觉.

基于提示工程的方法可以在不改变模型参数和结构的前提下, 缓解大模型幻觉. 然而, 由于提示工程是影响模型生成质量的关键环节, 这种方法的难点在于设计合理的输入指令, 引导模型生成具有较低幻觉性的结果.

4.3.2 事实指导

事实指导是提高大模型输出可靠性和信息量的一种直观、有效的方法. 事实指导方法可以用于缓解由下游任务领域专业化、模型同质化等问题带来的模型幻觉. 输入模型的引导信号可以定义为对原输入的附加内容. 在事实指导方法中, 关键在于向模型输入信息的类型以及输入信息的方法. 引导信号可以是关键词句、外部知识或其他结构, 如关系图或语义图. 其中, 关键词指导主要用于缓解内在幻觉、封闭域幻觉、输入矛盾幻觉以及上下文矛盾幻觉. 而外部知识指导引入了额外事实信息, 可以缓解外在幻觉、开放域幻觉和事实矛盾幻觉.

在关键词指导层面, 针对抽象摘要任务, Li 等人^[154]提出了结合提取方法的语言生成模型指导方法, 通过关键信息指南网络的方式将关键词编码为关键信息表示, 并以协同注意机制和指针机制中的关键词表示指导生成过程. Saito 等人^[155]进一步将预训练模型与显著性标记模型相结合, 通过显著性标记模型为每个令牌产生一个分数, 实现对生成内容重要的字符片段的选择.

在外部知识指导层面, Dong 等人^[156]提出, 通过利用从源链接的外部知识库可以提高摘要的忠实度. 在对话生成任务中, Shuster 等人^[156]提出检索-增强的神经结构, 基于检索到的相关知识生成对话, 以提升生成内容的忠实性. 在知识问答系统中, Zhang 等人^[157]提出一种用户和知识库交互的框架, 通过使用语言模型实现自动问题-知识对齐, 以使存储的信息与用户的问题相关联. 针对其他领域的专业任务, Bran 等人^[158]提出由于大模型缺乏对外部知识来源的获取, 限制了其在科学应用场景的有效性. 通过将大模型与化学专家工具结合, 大模型原本欠缺的化学性能获得了增强, 并涌现出了将化学相关任务自动化的能力.

综上所述, 事实指导是提高大模型忠实度和信息性的一种直观、有效的方法. 在事实指导框架中, 关键的技术挑战在于确定向模型中输入信息的类别以及输入方式.

4.3.3 多模态应用

在通用人工智能发展的背景下, 实现多模态、多任务的智能体已经成为人工智能研究的新趋势. 由于大模型具有解决自然语言处理、机器视觉、自主决策等多模态任务的能力, 其可以作为涵盖各种输入模态的下游任务的支柱, 成为通用智能体的核心组件. 同时, 大模型在某一领域的专业能力也十分重要. 因此, 在模型部署过程中, 基于多模态应用幻觉缓解方法依赖于下游任务类型及任务需求, 需要通过对不同下游任务进行适配, 提升大模型在特定领域的专业能力, 并可以解决多种类型的幻觉性问题.

Li 等人^[150]首次对大型视觉语言模型中的物体幻觉开展研究. 其研究揭示输入视觉指令中出现的物体可能会导致模型幻觉, 为避免评价标准受到输入指令和生成风格的影响, 基于轮询的物体幻觉评估方法可能有效缓解该问题. Yoon 等人^[132]发现基于视频的对话系统可能出现以输入中不加选择地复制文本为表现形式的文本幻觉问题. 根据信息理论文本幻觉测量方法, 他们提出了文本幻觉正则化损失, 相应的文本幻觉缓解框架显著提升了系统的忠实性和可解释性. Biten 等人^[121]针对图像字幕中的物体幻觉问题, 采用均匀抽样、逆多项式抽样和更新共存矩阵等不同的策略选择要替换的对象, 从而开展数据增强. 该方法显著减少模型对幻觉指标的物体偏差, 同时减少

了对视觉特征的依赖性。

针对条件语言的生成任务, Xiao 等人^[127]发现在图像字幕和数据到文本生成任务中, 预测不确定性与幻觉发生几率正相关, 从而提出了不确定性感知解码器的方法, 通过在模型解码过程中惩罚预测的不确定性来减少幻觉。Dai 等人^[133]研究了大规模视觉语言预训练模型在基于视觉信息生成文本时的幻觉问题。他们发现基于图像块(patch)的图像编码会产生较低的幻觉性, 且更小的图像块分辨率可以显著减少模型关于物体的幻觉性。

此外, 鉴于词(token)级的图像文本对齐和受控生成对减少幻觉的重要性, 对象掩码语言模型被提出。通过屏蔽在图像中出现文本的所有对象, 该方法通过增强文本标记和视觉对象之间的对齐和限制减少幻觉的产生^[159]。

为了更清晰地展示基于应用层的幻觉缓解方法的使用场景, 表 5 中将应用层幻觉缓解方法和不同产生幻觉的机理与任务类型联系起来, 相关方法也被作为实例总结在表 5 中。

表 5 应用层幻觉缓解方法分类

幻觉成因	应用层幻觉缓解方法	任务类型	实例
模型同质化、提示工程问题	提示工程	问答	Zhang 等人 ^[152]
		文本生成	Martino 等人 ^[153]
		视觉-语言	Li 等人 ^[150]
		专业任务	Lightman 等人 ^[149] , Kumar ^[151]
下游任务领域专业化、模型同质化	事实指导	摘要总结	Li 等人 ^[154] , Saito 等人 ^[155] , Dong 等人 ^[156]
		对话	Shuster 等人 ^[36]
		问答	Zhang 等人 ^[157]
		专业任务	Bran 等人 ^[158]
多模态化问题	多模态应用	文本生成	Xiao 等人 ^[127]
		视觉-语言	Li 等人 ^[150] , Yoon 等人 ^[132] , Biten 等人 ^[121] , Dai 等人 ^[133] , Ullah 等人 ^[159]

4.4 小 结

本节分别从数据层、模型层、应用层这 3 个角度详细梳理了大模型幻觉的缓解方法, 并对不同方法的原理进行了介绍和分析, 相关的研究总结见表 6。

通过以上分析, 大模型幻觉的缓解方法较为丰富, 但也存在一些局限性, 具体对比分析如下(见表 7)。

(1) 数据层幻觉缓解方法集中在通过构建真实数据集或修正幻觉数据集以提高数据质量从而减少幻觉的产生。常用的策略包括数据清洗和数据增强等方法, 亦有研究者关注对抗学习和检索增强的架构。目前数据层方法的局限性在于针对于个别数据集或者下游任务设计缺乏泛化性, 同时缺乏小样本学习方法以进一步提升效率。

(2) 模型层幻觉缓解方法涵盖从模型初始结构、模型训练与微调过程、模型后处理方法等各个方面, 灵活且多样, 并易于根据不同幻觉成因进行迭代优化。同时, 所提出方法也面临解释性差, 缺乏保障模型时效性, 难以平衡生成性能和幻觉水平的局限。

表 6 大模型幻觉缓解方法分类

大类	子类	实例	幻觉类型
数据层幻觉缓解方法	数据收集	Gardent 等人 ^[102] , Parikh 等人 ^[23] , Wang 等人 ^[20] , Gabriel 等人 ^[103] , Honovich 等人 ^[55] , Dziri 等人 ^[104] , Cheng 等人 ^[105] , Chen 等人 ^[106] , Sun 等人 ^[107]	内在幻觉 封闭域幻觉 输入矛盾幻觉 上下文矛盾幻觉
	数据预处理	Filippova ^[60] , Raunak 等人 ^[108] , Nie 等人 ^[109] , Liu 等人 ^[110] , Shen 等人 ^[111] , Rebuffel 等人 ^[112] , Dušek 等人 ^[113] , Nan 等人 ^[41] , Junczys-Dowmunt 等人 ^[115] , Zhang 等人 ^[116] , Cao 等人 ^[114] , Nie 等人 ^[117] , Wang 等人 ^[118] , Longpre 等人 ^[119] , Chen 等人 ^[120] , Biten 等人 ^[121] , Bi 等人 ^[122] , Shuster 等人 ^[36]	内在幻觉 封闭域幻觉 输入矛盾幻觉 上下文矛盾幻觉

表 6 大模型幻觉缓解方法分类 (续)

大类	子类	实例	幻觉类型
模型层幻觉缓解方法	模型结构	编码器 Lim等人 ^[123] , Zhu等人 ^[124] , Wang等人 ^[35] , Huang等人 ^[125]	内在幻觉 封闭域幻觉 输入矛盾幻觉 上下文矛盾幻觉
		解码器 Liu等人 ^[126] , Rebuffel等人 ^[112] , Xiao等人 ^[127] , Song等人 ^[128] , Balakrishnan等人 ^[129]	不受限于任何幻觉类型 与缓解场景相关
	模型训练与微调	损失函数 Lee等人 ^[130] , Kang等人 ^[131] , Yoon等人 ^[132] , Dai等人 ^[133] , Li等人 ^[134] , Wang等人 ^[135]	内在幻觉 封闭域幻觉 输入矛盾幻觉 上下文矛盾幻觉
		辅助任务 Cao等人 ^[61] , Mesgar等人 ^[137] , Song等人 ^[138] , Wang等人 ^[35] , Schulman等人 ^[139] , Weng等人 ^[140] , Li等人 ^[141] , Nan等人 ^[41]	内在幻觉 封闭域幻觉 输入矛盾幻觉 上下文矛盾幻觉
	后处理输出内容 Cao等人 ^[142] , Chen等人 ^[120] , Dong等人 ^[143] , Song等人 ^[144] , Dziri等人 ^[145]		不受限于任何幻觉类型 与缓解场景相关
	(反)专家模型 Qiu等人 ^[146] , Ilharco等人 ^[147] , Choubey等人 ^[148] , Liu等人 ^[126]		不受限于任何幻觉类型 与缓解场景相关
	提示工程 Lightman等人 ^[149] , Li等人 ^[150] , Kumar ^[151] , Zhang等人 ^[152] , Martino等人 ^[153]		不受限于任何幻觉类型 与缓解场景相关
	应用层幻觉缓解方法	事实指导 Li等人 ^[154] , Saito等人 ^[155] , Dong等人 ^[156] , Shuster等人 ^[36] , Zhang等人 ^[157] , Bran等人 ^[158]	关键词指导: 内在幻觉、封闭域幻觉、输入矛盾幻觉、上下文矛盾幻觉 外部知识指导: 外在幻觉、开放域幻觉、事实矛盾幻觉
		多模态应用 Li等人 ^[150] , Yoon等人 ^[132] , Biten等人 ^[121] , Xiao等人 ^[127] , Dai等人 ^[133] , Ullah等人 ^[159]	依赖于下游任务类型及任务需求

表 7 大模型幻觉缓解方法对比

分类	优点	局限性
数据层幻觉缓解方法	方便构建对应任务的忠实数据集 大幅度提高数据标记效率和质量	缺乏泛化性 缺乏小样本学习方法
模型层幻觉缓解方法	灵活且具有多样性 易于迭代优化	缺乏一定的解释性 缺乏保障模型时效性的方法 难以平衡生成性能和幻觉水平
应用层幻觉缓解方法	效率高 易于实现	缺乏与下游任务的精确适配 依赖于外部知识

(3) 应用层幻觉缓解方法能够直接设计相关提示或者对齐指令缓解特定任务幻觉的产生, 效率高且易于实现。但仍然存在方法的颗粒度高, 与下游任务适配不精确等局限性。此外, 外部知识是否完备也是影响方法有效性的重要原因。

5 大模型幻觉问题的未来展望

5.1 大模型幻觉问题影响的系统认知

在开发真实可靠的大模型时, 如何平衡创造力与真实性是处理幻觉问题的巨大挑战^[160,161]。大模型的幻觉问题具有两面性^[162,163]。一方面, 幻觉被视为模型的缺陷, 需要通过技术手段予以纾解。尤其当大模型用于法律、医疗等关键应用时, 不正确或虚构的信息可能导致严重后果, 破坏信任和可靠性。但与此同时, 幻觉也可以带来创新和创

造性,生成出人意料的、新颖的想法。例如,在创意写作、产品设计、营销广告等应用场景中,幻觉能够激发设计师的灵感,协助探索多种可能的解决方案,生成引人注目的广告语和活动策划,打破传统思维定式的束缚。

已有学者认识到幻觉的两面性特点及其带来的相互作用,从认知科学的角度评估其价值,旨在将幻觉风险降至最低的同时,评估和利用其创造潜力,使大模型幻觉的价值最大化。Sam Altman 认为大模型的幻觉是释放人工智能创造潜力的基石^[164]。Lee 等人^[165]通过严格的数学分析,从概率论和信息论的角度为理解大模型幻觉和创造力的相互联系提供了理论参考。Wang 等人^[166]通过对多模态 AGI 模型的分析表明幻觉与创造之间的相互作用是有益的。Rawte 等人^[167]则将幻觉模型描述为协作创意伙伴,在创造性和艺术的背景下,幻觉被视为是创新思维的催化剂。在科学研究领域,大模型的创造性有助于拓展人类知识边界,协助人类研究者取得突破性进展^[168]。具体而言,Gupta 等人^[169]基于微调的大模型实现了精确的分子能量预测。Völker 等人^[170]开发了 Text2Concrete,通过大模型对实验进行优先级排序,从而加快混凝土材料设计进程。Hong 等人^[171]利用科学文章训练 ScholarBert 模型,并通过上下文发现与储能应用相关的氢载体分子。Abramson 等人^[172]提出 AlphaFold3 模型,并利用其准确预测蛋白质、DNA、RNA 以及配体等生命分子的结构及其相互作用方式。综上所述,由于大模型具有自然语言输入的灵活性和上下文学习能力,因此可以非常有效地构建专业领域模型,并以前所未有的方式进行知识发现与创新。这些研究工作充分展现了大模型在科学研究中的巨大潜力,也为进一步探索幻觉与创造力的平衡提供了实践基础。

当前的研究工作倾向于减少幻觉,一定程度上忽略了其创造性。以教育领域为例,保证人工智能的诚实输出是教学和培训顺利开展的重要指标;然而,另一方面,利用幻觉生成的模拟教学案例和多样化场景能够增强学生的学习体验和理解能力,培养其批判性思维^[173]。因此,评估和利用幻觉现象与缓解幻觉问题同样重要,需要给予均衡的关注。Liu 等人^[174]通过调研发现,确保模型输出的真实性无可厚非,因为错误的信息对于实际问题的解决可能带来重大影响,但幻觉的评估不应仅局限于诚实和有用这两个维度。对于幻觉问题,可控性意味着模型能够控制幻觉水平,并在忠实性和多样性之间取得平衡。可控生成将幻觉视为一种可控属性,成为权衡幻觉和多样性的一种重要方法^[121,145]。这种方法通过受控重采样^[175]和提供控制代码,可以对幻觉水平进行手动^[60,175,176]或自动^[176]的控制。考虑到幻觉的两面性特点,可以进一步调整可控生成方法来改变幻觉的程度,以满足不同现实应用对忠实性和多样性的需求。

总之,在未来的工作中,通过更细致的幻觉分类兼顾事实准确性和知识灵活性具有重要的现实意义。具体而言,可通过对有益幻觉的注释和识别,使大模型能够有能力区分有害和有价值幻觉,通过价值对齐保证大模型的诚实性和真实性^[177,178]。随着大模型应用场景的不断扩展,不同应用领域的用户对于幻觉中创造力的价值具有不同的标准,对于幻觉两面性的深入理解与评估至关重要^[179]。未来的研究应继续致力于改进幻觉的检测和缓解技术,在不同应用场景下实现忠实性与创造力的动态平衡,最大限度地发挥大模型的价值。

5.2 大模型幻觉评估展望

尽管目前的幻觉评估方法已经在许多任务中取得了不错的效果,但仍存在一些不足与挑战。有关幻觉评估方法潜在的研究方向将体现于下列方面。

5.2.1 细粒度 (fine-grained)

一方面,许多幻觉评估方法是通过统计指标度量的,虽可估计生成文本的整体幻觉水平,但很难精确定位幻觉。而幻觉定位对于成因剖析、寻求缓解以及下游部署策略等具有重要意义,因此需要从更细致的层面来展开。另一方面,绝大多数方法在评估时未能区分内在幻觉与外在幻觉,无法为人们提供足够的信息探究幻觉来源与纾解方法。设计细粒度的评估方法需要两步:先从更精细的层级精确定位幻觉,如使用依存级的依存弧^[53];再对检测出的幻觉分别进行分类,这需要能够对幻觉进行自动分类的方法加以指引。因此,对该方法的实现是研究开展的未来方向,其挑战包括:用何种方式提取文本关键信息来定位幻觉、如何自动判断生成文本中幻觉部分与源文本的关系以确定幻觉类别以及如何保证两步方法的适配度等。

5.2.2 内在幻觉的解释与外在幻觉的事实验证

考虑到内在幻觉的成因,如果设计的幻觉评估方法能够提供适当的关于内在幻觉的解释,例如提供与生成文本中幻觉相矛盾的源文本中的关键和精确信息,则可以更好地判断幻觉评估的正确性并更为精确地分析幻觉。这

不仅可以指导幻觉评估方法的设计,还能够为幻觉评估方法的应用增加可靠性。

而对于外在幻觉而言,仅检测外部幻觉难以满足需求,还需要使用世界知识对其进行事实验证。由于人工验证耗费通常需要大量成本,如何设计能够自动进行事实验证的方法是未来的研究方向。Ji 等人^[16]指出事实验证包含两个子任务:知识证据选择与声明验证。对于前一子任务,主要挑战在于如何从世界知识中获取证据,尤其是世界知识的完整性与正确性。如果能够准确地选择知识证据,便可更加精确地完成事实验证。该子任务的输出结果也可以提供给人类使用者,帮助其判断验证结果的合理性。而对于后一子任务,所用的验证模型鲁棒性一般较差,例如易受对抗性攻击以及易受否定词、数字和比较词的影响等^[180]。因此,如何提高其鲁棒性以及两个子任务之间的协同成为了重要的研究方向。

5.2.3 训练数据的自动合成

基于模型的评估方法需要足够的数据训练评估模型,而传统方法均是在已有数据集的基础上添加人工扰动,这不仅耗费人力,还可能导致添加扰动后的幻觉类型与模型产生的幻觉类型不一致,同时出现难以覆盖所有幻觉类型的情况,影响幻觉评估效果^[89]。因此,自动合成幻觉数据的方法被提出,旨在解决简单的噪声添加方法所呈现的模式单一以及难以满足训练要求的技术局限。Gekhman 等人^[101]提出了 TrueTeacher 方案,通过利用模型生成数据,并通过大模型标注幻觉。此研究方向的挑战在于如何保证自动生成数据的质量,包括幻觉类别覆盖情况、正负样本的占比,以及如何确保数据标注的正确性从而提升评估模型性能等问题。如果未来能够很好地解决这些挑战,则自动合成训练数据的方法将会为基于模型的方法的发展起到积极作用。

5.2.4 泛化能力

现有评估方法多针对特定的自然语言生成任务,难以直接应用到其他任务场景中,这限制了幻觉评估方法的发展。因此,如何设计能够适用于多种任务(多模态、多语言等)的自动评估方法将成为重要的研究方向。这包括了不同任务数据的标准化整合以及能够应用于多种任务的评估模型的设计。Zha 等人^[64]设计的 AlignScore 就是一种有益尝试。此外,为不同生成任务的幻觉评估设计标准化指标也十分重要,在这个过程中需要了解不同任务的联系,以避免牺牲一些任务的特有性质。因此,如何在泛化能力与特定任务中的表现之间做权衡也是该方向的一大挑战。

5.2.5 幻觉评估方法的可解释性

幻觉评估方法的可解释性对于大模型性能提升具有重要意义,可从模型内部和模型外部分别进行。从自解释性角度,可直接设计可解释性强的幻觉评估模型方法。例如,可参考已有设计,修改模型结构从而提升模型可解释性^[181,182],或者在评估时展示模型方法对于生成文本中幻觉部分的识别与评价过程以提供更多参考^[77]。从事后解释性角度,可在评估结果基础上执行额外的操作为其提供解释。基于迁移学习的思想,Wang 等人^[183]指出模型的可解释性具有可迁移性,可通过局部替代模型和模型蒸馏两类方法展开。前者用可解释强的模型来拟合目标模型的局部来辅助解释,代表性工作为 LIME^[184];后者则利用结构简单的学生模型来模拟相对复杂的教师模型,在尽可能保证性能的前提下压缩模型,从而提升解释模型决策过程^[57,185,186]。

此外,还可结合幻觉评估问题的特点,参考基于反向传播^[187,188]、基于注意力机制^[189]以及基于样例驱动^[190]等方法提高评估模型方法的可解释性。Michaud 等人^[191]提出了利用模型增强其自身可解释性的方法,这对提高幻觉自动评估方法的可解释性有所启发:可以用模型为自动评估方法提供安全性证明,再通过人工检查证明的正确性。

5.3 大模型幻觉缓解方法展望

5.3.1 数据层展望

如前所述,海量的优质数据是训练可信大模型的重要基石,而训练数据也往往是大模型幻觉的根本成因之一。因此,从数据角度入手缓解大模型幻觉问题是机器学习未来发展的重要方向之一。Villalobos 等人^[192]通过研究揭示,机器学习将可能在 2026 年前耗尽高质量的语言数据,并可能在 2060 年前耗尽图像数据。这表明,如果数据使用效率并未大幅提高或开发出新的优质数据源,依赖大量数据集的机器学习模型的发展趋势可能会放缓。因此,如何提升数据利用效率和质量,开拓更先进的数据增广方法和更有效的小样本学习方法,以及知识蒸馏架构等提升

数据利用效率的机器学习方法,将成为未来的重要发展方向.此外,由于尚不存在对各类人工智能任务有效的通用方法^[193],通用人工智能的发展受到实质性限制.因此,开发通用且稳健的处理方法以缓解大模型幻觉将成为未来发展的重要方向.

5.3.2 模型层展望

近年来,大模型的兴起和其在各领域的广泛应用在很大程度上源于其涌现能力,即当训练数据的量与模型规模超过某个阈值的时候,模型的性能突然大幅提升.然而,大模型的涌现特性是欠鲁棒的,且不具备可解释性.由于大模型训练基于概率分布进行,这就导致大模型生成的信息具有基于概率分布的随机性,无法在理论上保证生成信息的真实性和准确性.因此,改进模型结构,通过优化输入内容表示能力、数据推理能力以及因果推理能力,从根源上缓解模型幻觉成为重要的发展方向.

在大模型需要进行多模态输入输出任务的前提下,学习跨模态表示是提高大模型表达性,进而提高其忠实性的一个重要方向^[194].数据推理能力主要涉及数学问题解答和生成文本中数字的正确性,通过思维链等基于提示工程的推理方法,以及在数值建模中增加推理能力是缓解大模型幻觉问题的关键方法^[195].因果推理是根据一个结果发生的条件对因果关系得出结论的过程,与人类的思维过程类似.因此,将因果推理纳入大模型可以为解决模型的幻觉性问题提供新的思路^[196],从而提升机器学习模型可信性.

提升大模型的可解释性,特别是提供关于模型输出的结果解释问题^[197],对缓解大模型幻觉至关重要.首先,可以进一步将知识库技术与大模型相融合,在提升模型可解释性的同时,利用知识库中的真实知识缓解模型幻觉^[198].基于溯源的方法也可以为模型的输出提供可解释性,这种方法通过展示关于模型输出的预测推导过程,为最终的输出结果提供推理依据.当模型输出是一系列推理步骤的结果时,这是一种直观而有效的可解释性技术^[199,200].此外,对可解释性评估过程和指标的扩展,也可以作用于缓解模型幻觉性的过程中.这些指标可以作为损失函数的一部分纳入机器学习模型的训练过程中,或作为强化学习过程中的奖励函数.在可解释性指标的指导下,模型的可信性会进一步提升,其幻觉性问题也有机会得到有效缓解.

此外,虽然关于减轻幻觉的研究已经非常丰富,但大多数研究并未着重区分内在幻觉和外在幻觉.如前所述,当前研究的主要重点是处理内在幻觉,而外在幻觉则通常被忽视.由于大模型在事实性上的表现与其训练数据量和时效性的息息相关,模型的知识参数也需要不断更新^[201].因此,应对模型时效性等因素导致的外部幻觉也是面临的重要挑战.未来应当对内在和外在幻觉系统探索多元化缓解方法.

5.3.3 应用层展望

模型幻觉问题导致其实际部署面临着可信性不足、传播误导性信息等严峻挑战.推进大模型广泛落地的关键在于解决前沿技术与真实应用场景之间的适配问题.首先,应建立能够与应用场景相衔接的大模型体系.其次,需要配套平台、工具,尽可能降低非专家用户的应用门槛.最后,大模型需要完整的生态支持,包括应用生态、硬件生态的建设等.这些措施有助于对大模型的实际应用提供全流程、端到端的支持,增强其在真实应用场景中的可信度^[202].

大模型具有数据量大、参数规模大、通用性高等优势,但直接部署利用大模型却难以处理实际应用场景中多样化的应用需求,过多的通用知识可能会对专业问题产生幻觉性影响.因此,在通用模型基础上,还需要根据具体下游任务和专业领域创建专家模型.其中,针对特定任务,如自然语言处理领域的抽象摘要、对话等任务以及视觉领域的图文搜索、文档图像理解等,需要对大模型进行适配,以减少其幻觉性问题.同样,对于特定的行业或专业领域,应当以通用大模型为支柱,挖掘高质量的行业领域数据,将其与知识引入模型中,提高其生成内容的忠实性.

在用户层面,还可通过多样化的正确指令,对模型识别复杂性问题的能力做出指引,在提示工程中尽可能提供正确的外部知识,降低模型输出幻觉的概率.与此同时,以无害性、真实性、专业性为对齐标准,尽可能提供公正、无偏的人类反馈,优化和改进奖励模型,通过对齐技术进一步降低幻觉问题的产生也是值得关注的重要方向.

随着生成式人工智能产业的不断壮大,大型模型幻觉问题已经显现为制约其广泛应用的重要障碍.解决这一问题不仅需要数据层、模型层和应用层系统上展开综合性的努力,还需要广泛的技术社群、用户群体等多元利

益相关者的积极协同合作. 只有通过多维度的探索和合作, 形成跨领域的技术解决方案, 才能够切实推动可信人工智能的健康和持续发展.

References:

- [1] Yu TR, Jin R, Han XZ, Li JH, Yu T. Review of pre-training models for natural language processing. *Computer Engineering and Applications*, 2020, 56(23): 12–22 (in Chinese with English abstract). [doi: 10.3778/j.issn.1002-8331.2006-0040]
- [2] Bao SQ, He H, Wang F, Wu H, Wang HF. PLATO: Pre-trained dialogue generation model with discrete latent variable. In: *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL*, 2020. 85–96. [doi: 10.18653/v1/2020.acl-main.9]
- [3] Chen WH, Su Y, Yan XF, Wang WY. KGPT: Knowledge-grounded pre-training for data-to-text generation. *arXiv:2010.02307*, 2020.
- [4] Liu YH, Gu JT, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L. Multilingual denoising pre-training for neural machine translation. *Trans. of the Association for Computational Linguistics*, 2020, 8: 726–742. [doi: 10.1162/tacl_a_00343]
- [5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: *Proc. of the 31st Int'l Conf. on Neural Information Processing Systems*. Long Beach: Curran Associates Inc., 2017. 6000–6010.
- [6] Zhao WX, Zhou K, Li JY, Tang TY, Wang XL, Hou YP, Min YQ, Zhang BC, Zhang JJ, Dong ZC, Du YF, Yang C, Chen YS, Chen ZP, Jiang JH, Ren RY, Li YF, Tang XY, Liu ZK, Liu PY, Nie JY, Wen JR. A survey of large language models. *arXiv:2303.18223*, 2023.
- [7] Pagnoni A, Balachandran V, Tsvetkov Y. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In: *Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics*, 2021. 4812–4829. [doi: 10.18653/v1/2021.naacl-main.383]
- [8] Liu PJ, Saleh M, Pot E, Goodrich B, Sepassi R, Kaiser L, Shazeer N. Generating Wikipedia by summarizing long sequences. In: *Proc. of the 6th Int'l Conf. on Learning Representations*. 2018. 1–18.
- [9] Wiseman S, Shieber S, Rush A. Challenges in data-to-document generation. In: *Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing*. Copenhagen: Association for Computational Linguistics, 2017. 2253–2263. [doi: 10.18653/v1/D17-1239]
- [10] Zhou CT, Neubig G, Gu JT, Diab M, Guzmán F, Zettlemoyer L, Ghazvininejad M. Detecting hallucinated content in conditional neural sequence generation. In: *Proc. of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics*, 2021. 1393–1404. [doi: 10.18653/v1/2021.findings-acl.120]
- [11] Zhang C, Lee G, D'Haro LF, Li HZ. D-Score: Holistic dialogue evaluation without reference. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2021, 29: 2502–2516. [doi: 10.1109/TASLP.2021.3074012]
- [12] Zhang XC, Ghorbani AA. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 2020, 57(2): 102025. [doi: 10.1016/j.ipm.2019.03.004]
- [13] Huang YC, Feng XC, Feng XC, Qin B. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv:2104.14839*, 2023.
- [14] Cao ZQ, Wei FR, Li WJ, Li SJ. Faithful to the original: Fact aware neural abstractive summarization. In: *Proc. of the 32nd AAAI Conf. on Artificial Intelligence*. Palo Alto: AAAI, 2018. 4784–4791. [doi: 10.1609/aaai.v32i1.11912]
- [15] Sobieszek A, Price T. Playing games with ais: The limits of GPT-3 and similar large language models. *Minds & Machines*, 2022, 32(2): 341–364. [doi: 10.1007/s11023-022-09602-0]
- [16] Ji ZW, Lee N, Frieske R, Yu TZ, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023, 55(12): 248. [doi: 10.1145/3571730]
- [17] OpenAI: GPT-4 System Card. 2023. <https://cdn.openai.com/papers/gpt-4-system-card.pdf>
- [18] Evans O, Cotton-Barratt O, Finnveden L, Bales A, Balwit A, Wills P, Righetti L, Saunders W. Truthful AI: Developing and governing AI that does not lie. *arXiv:2110.06674*, 2021.
- [19] Zhang Y, Li YF, Cui LY, Cai D, Liu LM, Fu TC, Huang XT, Zhao EB, Zhang Y, Chen YL, Wang LY, Luu AT, Bi W, Shi F, Shi SM. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv:2309.01219*, 2023.
- [20] Wang HM. Revisiting challenges in data-to-text generation with fact grounding. *arXiv:2001.03830*, 2020.
- [21] Lee K, Ippolito D, Nystrom A, Zhang CY, Eck D, Callison-Burch C, Carlini N. Deduplicating training data makes language models better. *arXiv:2107.06499*, 2022.
- [22] Jo ES, Gebru T. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In: *Proc. of the 2020 Conf. on Fairness, Accountability, and Transparency*. Barcelona: ACM, 2020. 306–316. [doi: 10.1145/3351095.3372829]
- [23] Parikh A, Wang XZ, Gehrmann S, Faruqui M, Dhingra B, Yang DY, Das D. ToTT: A controlled table-to-text generation dataset. In: *Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, 2020.

- 1173–1186. [doi: 10.18653/v1/2020.emnlp-main.89]
- [24] Tian R, Narayan S, Sellam S, Parikh AP. Sticking to the facts: Confident decoding for faithful data-to-text generation. arXiv:1910.08684, 2020.
- [25] Longpre S, Perisetla K, Chen A, Ramesh N, DuBois C, Singh S. Entity-based knowledge conflicts in question answering. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021. 7052–7063. [doi: 10.18653/v1/2021.emnlp-main.565]
- [26] Bommasani R, Hudson DA, Adeli E, *et al.* On the opportunities and risks of foundation models. arXiv:2108.07258, 2022.
- [27] Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV. Finetuned language models are zero-shot learners. arXiv:2109.01652, 2022.
- [28] Wei J, Wang XZ, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 24824–24837.
- [29] Liu FX, Lin K, Li LJ, Wang JF, Yacoob Y, Wang LJ. Mitigating hallucination in large multi-modal models via robust instruction tuning. arXiv:2306.14565, 2024.
- [30] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 22199–22213.
- [31] Durmus E, He H, Diab M. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 5055–5070. [doi: 10.18653/v1/2020.acl-main.454]
- [32] Dhingra B, Faruqui M, Parikh A, Chang MW, Das D, Cohen W. Handling divergent reference texts when evaluating table-to-text generation. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 4884–4895. [doi: 10.18653/v1/P19-1483]
- [33] Manakul P, Liusie A, Gales MJF. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. arXiv:2303.08896, 2023.
- [34] Niu C, Wu YH, Zhu J, Xu SL, Shum K, Zhong R, Song JT, Zhang T. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. arXiv:2401.00396, 2024.
- [35] Wang ZY, Wang XY, An B, Yu D, Chen CY. Towards faithful neural table-to-text generation with content-matching constraints. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 1072–1086. [doi: 10.18653/v1/2020.acl-main.101]
- [36] Shuster K, Poff S, Chen MY, Kiela D, Weston J. Retrieval augmentation reduces hallucination in conversation. In: Proc. of the Association for Computational Linguistics: EMNLP. Punta Cana: Association for Computational Linguistics, 2021. 3784–3803. [doi: 10.18653/v1/2021.findings-emnlp.320]
- [37] Popović M. chrF: Character n-gram F-score for automatic MT evaluation. In: Proc. of the 10th Workshop on Statistical Machine Translation. Lisbon: Association for Computational Linguistics, 2015. 392–395. [doi: 10.18653/v1/W15-3049]
- [38] Martindale M, Carpuat M, Duh K, McNamee P. Identifying fluently inadequate output in neural and statistical machine translation. In: Proc. of Machine Translation Summit XVII: Research Track. Dublin: European Association for Machine Translation, 2019. 233–243.
- [39] Banko M, Cafarella MJ, Soderland S, Broadhead M, Etzioni O. Open information extraction from the Web. In: Proc. of the 20th Int'l Joint Conf. on Artificial Intelligence. Hyderabad: Morgan Kaufmann Publishers Inc., 2007. 2670–2676.
- [40] Goodrich B, Rao V, Liu PJ, Saleh M. Assessing the factual accuracy of generated text. In: Proc. of the 25th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. Anchorage: ACM, 2019. 166–175. [doi: 10.1145/3292500.3330955]
- [41] Nan F, Nallapati R, Wang ZG, dos Santos CN, Zhu HH, Zhang DJ, McKeown K, Xiang B. Entity-level factual consistency of abstractive text summarization. In: Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, 2021. 2727–2733. [doi: 10.18653/v1/2021.eacl-main.235]
- [42] Lee N, Ping W, Xu P, Patwary M, Fung P, Shoenybi M, Catanzaro B. Factuality enhanced language models for open-ended text generation. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 34586–34599.
- [43] Dušek O, Novikova J, Rieser V. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. Computer Speech & Language, 2020, 59: 123–156. [doi: 10.1016/j.csl.2019.06.009]
- [44] Dušek O, Kasner Z. Evaluating semantic accuracy of data-to-text generation with natural language inference. In: Proc. of the 13th Int'l Conf. on Natural Language Generation. Dublin: Association for Computational Linguistics, 2020. 131–137. [doi: 10.18653/v1/2020.inlg-

1.19]

- [45] Williams A, Nangia N, Bowman S. A broad-coverage challenge corpus for sentence understanding through inference. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers). New Orleans: Association for Computational Linguistics, 2018. 1112–1122. [doi: 10.18653/v1/N18-1101]
- [46] Nie YX, Williams A, Dinan E, Bansal M, Weston J, Kiela D. Adversarial NLI: A new benchmark for natural language understanding. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 4885–4901. [doi: 10.18653/v1/2020.acl-main.441]
- [47] Laban P, Schnabel T, Bennett PN, Hearst MA. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. Trans. of the Association for Computational Linguistics, 2022, 10: 163–177. [doi: 10.1162/TACL_A_00453]
- [48] Kryściński W, McCann B, Xiong CM, Socher R. Evaluating the factual consistency of abstractive text summarization. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 9332–9346. [doi: 10.18653/v1/2020.emnlp-main.750]
- [49] Yin WP, Radev D, Xiong CM. DocNLI: A large-scale dataset for document-level natural language inference. In: Proc. of the 2021 Association for Computational Linguistics. Association for Computational Linguistics, 2021. 4913–4922. [doi: 10.18653/v1/2021.findings-acl.435]
- [50] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805, 2019.
- [51] Liu YH, Ott M, Goyal N, Du JF, Joshi M, Chen DQ, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.
- [52] He PC, Li XD, Gao JF, Chen WZ. DeBERTa: Decoding-enhanced BERT with disentangled attention. arXiv:2006.03654, 2021.
- [53] Goyal T, Durrett G. Evaluating Factuality in generation with dependency-level entailment. In: Proc. of the Association for Computational Linguistics: EMNLP. Association for Computational Linguistics, 2020. 3592–3603. [doi: 10.18653/v1/2020.findings-emnlp.322]
- [54] Falke T, Ribeiro LFR, Utama PA, Dagan I, Gurevych I. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 2214–2220. [doi: 10.18653/v1/P19-1213]
- [55] Honovich O, Choshen L, Aharoni R, Neeman E, Szpektor I, Abend O. Q²: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021. 7856–7870. [doi: 10.18653/v1/2021.emnlp-main.619]
- [56] Fabbri AR, Wu CS, Liu WH, Xiong CM. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In: Proc. of the 2022 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle: Association for Computational Linguistics, 2022. 2587–2601. [doi: 10.18653/v1/2022.naacl-main.187]
- [57] Barrantes M, Herudek B, Wang R. Adversarial NLI for factual correctness in text summarisation models. arXiv:2005.11739, 2020.
- [58] Yang ZL, Dai ZH, Yang YM, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized autoregressive pretraining for language understanding. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 5735–5763.
- [59] Dziri N, Rashkin H, Linzen T, Reitter D. Evaluating groundedness in dialogue systems: The BEGIN benchmark. arXiv:2105.00071, 2022.
- [60] Filippova K. Controlled Hallucinations: Learning to generate faithfully from noisy data. In: Proc. of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, 2020. 864–870. [doi: 10.18653/v1/2020.findings-emnlp.76]
- [61] Cao M, Dong Y, Cheung J. Hallucinated but factual! Inspecting the factuality of hallucinations in abstractive summarization. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Dublin: Association for Computational Linguistics, 2022. 3340–3354. [doi: 10.18653/v1/2022.acl-long.236]
- [62] Yu JF, Wang XZ, Tu SQ, Cao SL, Zhang-Li D, Lv X, Peng H, Yao ZJ, Zhang XH, Li HM, Li CY, Zhang ZY, Bai YS, Liu YT, Xin A, Lin NY, Yun KF, Gong LL, Chen JH, Wu ZL, Qi YJ, Li WK, Guan Y, Zeng KS, Qi J, Jin HL, Liu JX, Gu Y, Yao Y, Ding N, Hou L, Liu ZY, Xu B, Tang J, Li JZ. KoLA: Carefully benchmarking world knowledge of large language models. arXiv:2306.09296, 2024.
- [63] Deng MK, Tan BW, Liu ZZ, Xing E, Hu ZT. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021. 7580–7605. [doi: 10.18653/v1/2021.emnlp-main.599]
- [64] Zha YH, Yang YC, Li RC, Hu ZT. AlignScore: Evaluating factual consistency with a unified alignment function. arXiv:2305.16739,

- 2023.
- [65] Yue X, Wang BS, Chen ZR, Zhang K, Su Y, Sun H. Automatic evaluation of attribution by large language models. arXiv:2305.06311, 2023.
- [66] Zhong M, Liu Y, Yin D, Mao YN, Jiao YZ, Liu PF, Zhu CG, Ji H, Han JW. Towards a unified multi-dimensional evaluator for text generation. In: Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing. Abu Dhabi: Association for Computational Linguistics, 2022. 2023–2038. [doi: 10.18653/v1/2022.emnlp-main.131]
- [67] Clark C, Lee K, Chang MW, Kwiatkowski T, Collins M, Toutanova K. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 2924–2936. [doi: 10.18653/v1/N19-1300]
- [68] Zhang TY, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating text generation with BERT. arXiv:1904.09675, 2020.
- [69] Yuan WZ, Neubig G, Liu PF. BARTScore: Evaluating generated text as text generation. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 27263–27277.
- [70] Mehri S, Eskenazi M. USR: An unsupervised and reference free evaluation metric for dialog generation. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 681–707. [doi: 10.18653/v1/2020.acl-main.64]
- [71] Wei JH, Yao YS, Ton JF, Guo HY, Estornell A, Liu Y. Measuring and reducing LLM hallucination without gold-standard answers. arXiv:2402.10412, 2024.
- [72] Su WH, Wang CY, Ai QY, Hu YR, Wu ZJ, Zhou YJ, Liu YQ. Unsupervised real-time hallucination detection based on the internal states of large language models. arXiv:2403.06448, 2024.
- [73] Zhang SY, Li Y, Wu R, Huang XT, Chen YR, Xu WH, Qi GL. DEE: Dual-stage explainable evaluation method for text generation. arXiv:2403.11509, 2024.
- [74] Wang A, Cho K, Lewis M. Asking and answering questions to evaluate the factual consistency of summaries. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 5008–5020. [doi: 10.18653/v1/2020.acl-main.450]
- [75] Nan F, dos Santos CN, Zhu HH, Ng P, McKeow K, Nallapati R, Zhang DJ, Wang ZG, Arnold AO, Xiang B. Improving factual consistency of abstractive summarization via question answering. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing (Vol. 1: Long Papers). Association for Computational Linguistics, 2021. 6881–6894. [doi: 10.18653/v1/2021.acl-long.536]
- [76] Shakeri S, dos Santos CN, Zhu HH, Ng P, Nan F, Wang ZG, Nallapati R, Xiang B. End-to-end synthetic data generation for domain adaptation of question answering systems. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 5445–5460. [doi: 10.18653/v1/2020.emnlp-main.439]
- [77] Scialom T, Dray PA, Lamprier S, Piwowarski B, Staiano J, Wang A, Gallinari P. QuestEval: Summarization asks for fact-based evaluation. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021. 6594–6604. [doi: 10.18653/v1/2021.emnlp-main.529]
- [78] Rebuffel C, Scialom T, Soulier L, Piwowarski B, Lamprier S, Staiano J, Scouttheeten G, Gallinari P. Data-QuestEval: A referenceless metric for data-to-text semantic evaluation. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021. 8029–8036. [doi: 10.18653/v1/2021.emnlp-main.633]
- [79] Yehuda Y, Malkiel I, Barkan O, Weill J, Ronen R, Koenigstein N. In search of truth: An interrogation approach to hallucination detection. arXiv:2403.02889, 2024.
- [80] Vu T, Iyyer M, Wang XZ, Constant N, Wei J, Wei J, Tar C, Sung YH, Zhou D, Le Q, Luong T. FreshLLMs: Refreshing large language models with search engine augmentation. arXiv:2310.03214, 2023.
- [81] Zhang JX, Li ZH, Das K, Malin BA, Kumar S. SAC3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. arXiv:2311.01740, 2024.
- [82] Dinan E, Roller S, Shuster K, Fan A, Auli M, Weston J. Wizard of Wikipedia: Knowledge-powered conversational agents. arXiv:1811.01241, 2019.
- [83] Santhanam S, Hedayatnia B, Gella S, Padmakumar A, Kim S, Liu Y, Hakkani-Tur D. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. arXiv:2110.05456, 2022.
- [84] Utama P, Bambrick J, Moosavi N, Gurevych I. Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization. In: Proc. of the 2022 Conf. of the North American Chapter of the Association for Computational

- Linguistics: Human Language Technologies. Seattle: Association for Computational Linguistics, 2022. 2763–2776. [doi: 10.18653/v1/2022.naacl-main.199]
- [85] Jiang C, Qi BQ, Hong XY, Fu DY, Cheng Y, Meng FD, Yu M, Zhou BW, Zhou J. On large language models' hallucination with regard to known facts. arXiv:2403.20009, 2024.
- [86] Li JY, Cheng XX, Zhao WX, Nie JY, Wen JR. HaluEval: A large-scale hallucination evaluation benchmark for large language models. arXiv:2305.11747, 2023.
- [87] Lin S, Hilton J, Evans O. TruthfulQA: Measuring how models mimic human falsehoods. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Dublin: Association for Computational Linguistics, 2022. 3214–3252. [doi: 10.18653/v1/2022.acl-long.229]
- [88] Tang LY, Shalymov I, Wong AWM, Burnsky J, Vincent JW, Yang YA, Singh S, Feng S, Song H, Su H, Sun LJ, Zhang Y, Mansour S, McKeown K. TofuEval: Evaluating hallucinations of LLMs on topic-focused dialogue summarization. arXiv:2402.13249, 2024.
- [89] Chen KD, Chen Q, Zhou J, He YS, He L. DiaHalu: A dialogue-level hallucination evaluation benchmark for large language models. arXiv:2403.00896, 2024.
- [90] Zhu ZY, Yang YM, Sun ZQ. HaluEval-Wild: Evaluating hallucinations of language models in the wild. arXiv:2403.04307, 2024.
- [91] Chen SQ, Zhao YR, Zhang JH, Chern IC, Gao SY, Liu PF, He JX. FELM: Benchmarking factuality evaluation of large language models. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 44502–44523.
- [92] Yang SP, Sun RL, Wan XJ. A new benchmark and reverse validation method for passage-level hallucination detection. arXiv:2310.06498, 2023.
- [93] Lattimer BM, Chen P, Zhang XY, Yang Y. Fast and accurate factual inconsistency detection over long documents. arXiv:2310.13189, 2023.
- [94] Muhlgay D, Ram O, Magar I, Levine Y, Ratner N, Belinkov Y, Abend O, Leyton-Brown K, Shashua A, Shoham Y. Generating benchmarks for factuality evaluation of language models. arXiv:2307.06908, 2024.
- [95] Mündler N, He JX, Jenko S, Vechev M. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. arXiv:2305.15852, 2024.
- [96] Kocmi T, Federmann C. Large language models are state-of-the-art evaluators of translation quality. arXiv:2302.14520, 2023.
- [97] Gao MQ, Ruan J, Sun RL, Yin XJ, Yang SP, Wan XJ. Human-like summarization evaluation with chatgpt. arXiv:2304.02554, 2023.
- [98] Liu Y, Iter D, Xu YC, Wang SH, Xu RC, Zhu CG. G-Eval: NLG evaluation using GPT-4 with better human alignment. arXiv:2303.16634, 2023.
- [99] Min S, Krishna K, Lyu XX, Lewis M, Yih WT, Koh PW, Iyyer M, Zettlemoyer L, Hajishirzi H. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. arXiv:2305.14251, 2023.
- [100] Shafayat S, Kim E, Oh J, Oh A. Multi-FAct: Assessing multilingual LLMs' multi-regional knowledge using FActScore. arXiv:2402.18045, 2024.
- [101] Gekhman Z, Herzig J, Aharoni R, Elkind C, Szepkter I. TrueTeacher: Learning factual consistency evaluation with large language models. arXiv:2305.11171, 2023.
- [102] Gardent C, Shimorina A, Narayan S, Perez-Beltrachini L. Creating training corpora for NLG micro-planning. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Vancouver: Association for Computational Linguistics, 2017. 179–188. [doi: 10.18653/v1/P17-1017]
- [103] Gabriel S, Celikyilmaz A, Jha R, Choi Y, Gao JF. GO FIGURE: A meta evaluation of factuality in summarization. In: Proc. of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021. 478–487. [doi: 10.18653/v1/2021.findings-acl.42]
- [104] Dziri N, Kamalloo E, Milton S, Zaiane O, Yu M, Ponti EM, Reddy S. FaithDial: A faithful benchmark for information-seeking dialogue. Trans. of the Association for Computational Linguistics, 2022, 10: 1473–1490. [doi: 10.1162/tacl_a_00529]
- [105] Cheng ZJ, Dong HY, Wang ZR, Jia R, Guo JQ, Gao Y, Han S, Lou JG, Zhang DM. HiTab: A hierarchical table dataset for question answering and natural language generation. arXiv:2108.06712, 2022.
- [106] Chen ZY, Chen WH, Zha HW, Zhou XY, Zhang YK, Sundaresan S, Wang WY. Logic2Text: High-fidelity natural language generation from logical forms. arXiv:2004.14579, 2020.
- [107] Xu L, Li AQ, Zhu L, Xue H, Zhu CT, Zhao KK, He HN, Zhang XW, Kang QY, Lan ZZ. SuperCLUE: A comprehensive Chinese large language model benchmark. arXiv:2307.15020, 2023.
- [108] Raunak V, Menezes A, Junczys-Dowmunt M. The curious case of hallucinations in neural machine translation. In: Proc. of the 2021

- Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021. 1172–1183. [doi: 10.18653/v1/2021.naacl-main.92]
- [109] Nie F, Yao JG, Wang JP, Pan R, Lin CY. A simple recipe towards reducing hallucination in neural surface realisation. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 2673–2679. [doi: 10.18653/v1/P19-1256]
- [110] Liu TY, Zheng X, Chang BB, Sui ZF. Towards faithfulness in open domain table-to-text generation from an entity-centric view. In: Proc. of the AAAI Conf. on Artificial Intelligence. Virtually: AAAI Press, 2021. 13415–13423. [doi: 10.1609/aaai.v35i15.17583]
- [111] Shen L, Zhan HL, Shen X, Chen HS, Zhao XF, Zhu XD. Identifying untrustworthy samples: Data filtering for open-domain dialogues with Bayesian optimization. In: Proc. of the 30th ACM Int'l Conf. on Information and Knowledge Management. Virtual Event: ACM, 2021. 1598–1608. [doi: 10.1145/3459637.3482352]
- [112] Rebuffel C, Roberti M, Soulier L, Scouteeten G, Cancelliere R, Gallinari P. Controlling hallucinations at word level in data-to-text generation. Data Mining and Knowledge Discovery, 2022, 36(1): 318–354. [doi: 10.1007/s10618-021-00801-4]
- [113] Dušek O, Howcroft DM, Rieser V. Semantic noise matters for neural natural language generation. In: Proc. of the 12th Int'l Conf. on Natural Language Generation. Tokyo: Association for Computational Linguistics, 2019. 421–426. [doi: 10.18653/v1/W19-8652]
- [114] Honnibal M, Montani I. spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. Neural Machine Translation. Proc. of the Association for Computational Linguistics. 2017. 688–697.
- [115] Junczys-Dowmunt M. Dual conditional cross-entropy filtering of noisy parallel corpora. arXiv:1809.00197, 2019.
- [116] Zhang BL, Nagesh A, Knight K. Parallel corpus filtering via pre-trained language models. arXiv:2005.06166, 2020.
- [117] Nie F, Wang JP, Yao JG, Pan R, Lin CY. Operation-guided neural networks for high fidelity data-to-text generation. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 3879–3889. [doi: 10.18653/v1/D18-1422]
- [118] Wang TS, Ladhak F, Durmus E, He H. Improving faithfulness by augmenting negative summaries from fake documents. In: Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing. Abu Dhabi: Association for Computational Linguistics, 2022. 11913–11921. [doi: 10.18653/v1/2022.emnlp-main.816]
- [119] Longpre S, Perisetla K, Chen A, Ramesh N, DuBois C, Singh S. Entity-based knowledge conflicts in question answering. arXiv:2109.05052, 2022.
- [120] Chen SH, Zhang F, Sone K, Roth D. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021. 5935–5941. [doi: 10.18653/v1/2021.naacl-main.475]
- [121] Biten AF, Gómez L, Karatzas D. Let there be a clock on the beach: Reducing object hallucination in image captioning. In: Proc. of the 2022 IEEE/CVF Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2022. 2473–2482. [doi: 10.1109/WACV51458.2022.00253]
- [122] Bi B, Wu C, Yan M, Wang W, Xia JN, Li CL. Incorporating external knowledge into machine reading for generative question answering. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 2521–2530. [doi: 10.18653/v1/D19-1255]
- [123] Lim J, Kang M, Hur Y, Jung S, Kim J, Jang Y, Lee D, Ji H, Shin D, Kim S, Lim H. You truly understand what I need: Intellectual and friendly dialogue agents grounding knowledge and persona. arXiv:2301.02401, 2023.
- [124] Zhu CG, William H, Xu RC, Zeng QK, Zeng M, Huang XD, Jiang M. Enhancing factual consistency of abstractive summarization. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021. 718–733. [doi: 10.18653/v1/2021.naacl-main.58]
- [125] Huang LY, Wu LF, Wang L. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 5094–5107. [doi: 10.18653/v1/2020.acl-main.457]
- [126] Liu A, Sap M, Lu XM, Swayamdipta S, Bhagavatula C, Smith NA, Choi Y. DExperts: Decoding-time controlled text generation with experts and anti-experts. arXiv:2105.03023, 2021.
- [127] Xiao YJ, Wang WY. On hallucination and predictive uncertainty in conditional language generation. In: Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, 2021. 2734–2744. [doi: 10.18653/v1/2021.eacl-main.236]
- [128] Song KQ, Lebanoff L, Guo QP, Qiu XP, Xue XY, Li C, Yu D, Liu F. Joint parsing and generation for abstractive summarization. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 8894–8901. [doi: 10.1609/aaai.v34i05.6419]

- [129] Balakrishnan A, Rao JF, Upasani K, White M, Subba R. Constrained decoding for neural NLG from compositional representations in task-oriented dialogue. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 831–844. [doi: 10.18653/v1/P19-1080]
- [130] Lee K, Firat O, Agarwal A, Fannjiang C, Sussillo D. Hallucinations in neural machine translation. In: Proc. of the 2019 Int'l Conf. on Learning Representations. 2019.
- [131] Kang D, Hashimoto T. Improved natural language generation via loss truncation. arXiv:2004.14589, 2020.
- [132] Yoon S, Yoon E, Yoon HS, Kim J, Yoo CD. Information-theoretic text hallucination reduction for video-grounded dialogue. arXiv:2212.05765, 2022.
- [133] Dai WL, Liu ZH, Ji ZW, Su D, Fung P. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. arXiv:2210.07688, 2023.
- [134] Li T, Beirami A, Sanjabi M, Smith V. Tilted empirical risk minimization. In: Proc. of the 9th Int'l Conf. on Learning Representations. Virtual Event, 2021.
- [135] Wang CJ, Sennrich R. On exposure bias, hallucination and domain shift in neural machine translation. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 3544–3552. [doi: 10.18653/v1/2020.acl-main.326]
- [136] Ranzato MA, Chopra S, Auli M, Zaremba W. Sequence level training with recurrent neural networks. In: Proc. of the 4th Int'l Conf. on Learning Representations. San Juan, 2016.
- [137] Mesgar M, Simpson E, Gurevych I. Improving factual consistency between a response and persona facts. In: Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, 2021. 549–562. [doi: 10.18653/v1/2021.eacl-main.44]
- [138] Song HY, Zhang WN, Hu JW, Liu T. Generating persona consistent dialogues by exploiting natural language inference. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 8878–8885. [doi: 10.1609/aaai.v34i05.6417]
- [139] Schulman J. Reinforcement learning from human feedback: Progress and challenges. 2023. <https://eecs.berkeley.edu/research/colloquium/230419-2/>
- [140] Weng RX, Yu H, Wei XP, Liu WH. Towards enhancing faithfulness for neural machine translation. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 2675–2684. [doi: 10.18653/v1/2020.emnlp-main.212]
- [141] Li CL, Bi B, Yan M, Wang W, Huang SF. Addressing semantic drift in generative question answering with auxiliary extraction. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. Association for Computational Linguistics, 2021. 942–947. [doi: 10.18653/v1/2021.acl-short.118]
- [142] Cao M, Dong Y, Wu JP, Cheung JCK. Factual error correction for abstractive summarization models. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 6251–6258. [doi: 10.18653/v1/2020.emnlp-main.506]
- [143] Dong Y, Wang SH, Gan Z, Cheng Y, Cheung JCK, Liu JJ. Multi-fact correction in abstractive text summarization. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 9320–9331. [doi: 10.18653/v1/2020.emnlp-main.749]
- [144] Song HY, Wang Y, Zhang WN, Liu XJ, Liu T. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 5821–5831. [doi: 10.18653/v1/2020.acl-main.516]
- [145] Dziri N, Madotto A, Zaïane O, Bose AJ. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021. 2197–2214. [doi: 10.18653/v1/2021.emnlp-main.168]
- [146] Qiu YF, Ziser Y, Korhonen A, Ponti EM, Cohen SB. Detecting and mitigating hallucinations in multilingual summarisation. arXiv:2305.13632, 2023.
- [147] Ilharco G, Ribeiro MT, Wortsman M, Gururangan S, Schmidt L, Hajishirzi H, Farhadi A. Editing models with task arithmetic. arXiv:2212.04089, 2023.
- [148] Choubey PK, Fabbri AR, Vig J, Wu CS, Liu WH, Rajani NF. CaPE: Contrastive parameter ensembling for reducing hallucination in abstractive summarization. arXiv:2110.07166, 2022.
- [149] Lightman H, Kosaraju V, Burda Y, Edwards H, Baker B, Lee T, Leike J, Schulman J, Sutskever I, Cobbe K. Let's verify step by step. arXiv:2305.20050, 2023.

- [150] Li YF, Du YF, Zhou K, Wang JP, Zhao WX, Wen JR. Evaluating object hallucination in large vision-language models. arXiv:2305.10355, 2023.
- [151] Kumar K. Geotechnical Parrot Tales (GPT): Harnessing large language models in geotechnical engineering. arXiv:2304.02138, 2023.
- [152] Zhang MR, Press O, Merrill W, Liu A, Smith NA. How language model hallucinations can snowball. arXiv:2305.13534, 2023.
- [153] Martino A, Iannelli M, Truong C. Knowledge injection to counter large language model (LLM) hallucination. In: Proc. of the Semantic Web: ESWC 2023 Satellite Events. Herssonissos: Springer, 2023. 182–185. [doi: 10.1007/978-3-031-43458-7_34]
- [154] Li CL, Xu WR, Li S, Gao S. Guiding generation for abstractive text summarization based on key information guide network. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.2 (Short Papers). New Orleans: Association for Computational Linguistics, 2018. 55–60. [doi: 10.18653/v1/N18-2009]
- [155] Saito I, Nishida K, Nishida K, Tomita J. Abstractive summarization with combination of pre-trained sequence-to-sequence and saliency models. arXiv:2003.13028, 2020.
- [156] Dong Y, Wieting J, Verga P. Faithful to the document or to the world? Mitigating hallucinations via entity-linked knowledge in abstractive summarization. arXiv:2204.13761, 2022.
- [157] Zhang S, Pan LM, Zhao JZ, Wang WY. The knowledge alignment problem: Bridging human and external knowledge for large language models. arXiv:2305.13669, 2024.
- [158] Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. ChemCrow: Augmenting large-language models with chemistry tools. arXiv:2304.05376, 2023.
- [159] Ullah N, Mohanta PP. Thinking hallucination for video captioning. In: Proc. of the 16th Asian Conf. on Computer Vision. Macao: Springer, 2022. 623–640. [doi: 10.1007/978-3-031-26316-3_37]
- [160] Li JY, Chen J, Ren RY, Cheng XX, Zhao WX, Nie JY, Wen JR. The dawn after the dark: An empirical study on factuality hallucination in large language models. arXiv:2401.03205, 2024.
- [161] Xu ZW, Jain S, Kankanhalli M. Hallucination is inevitable: An innate limitation of large language models. arXiv:2401.11817, 2024.
- [162] Bi BL, Liu SH, Wang YW, Mei LR, Cheng XQ. Is factuality decoding a free lunch for LLMs? Evaluation on knowledge editing benchmark. arXiv:2404.00216, 2024.
- [163] Huang L, Yu WJ, Ma WT, Zhong WH, Feng ZY, Wang HT, Chen QL, Peng WH, Feng XC, Qin B, Liu T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv:2311.05232, 2024.
- [164] Jiang XH, Tian YX, Hua FR, Xu CJ, Wang YZ, Guo J. A survey on large language model hallucination via a creativity perspective. arXiv:2402.06647, 2024.
- [165] Lee M. A mathematical investigation of hallucination and creativity in gpt models. Mathematics, 2023, 11(10): 2320. [doi: 10.3390/math11102320]
- [166] Wang F. Lighthouse: A survey of AGI hallucination. arXiv:2401.06792, 2024.
- [167] Rawte V, Sheth A, Das A. A survey of hallucination in large foundation models. arXiv:2309.05922, 2023.
- [168] Jablonka KM, Ai QX, Al-Feghali A, *et al.* 14 examples of how LLMs can transform materials science and chemistry: A reflection on a large language model hackathon. Digital Discovery, 2023, 2(5): 1233–1250. [doi: 10.1039/D3DD00113J]
- [169] Gupta AK, Raghavachari K. Three-dimensional convolutional neural networks utilizing molecular topological features for accurate atomization energy predictions. Journal of Chemical Theory and Computation, 2022, 18(4): 2132–2143. [doi: 10.1021/acs.jctc.1c00504]
- [170] Völker C, Rug T, Jablonka KM, Kruschwitz S. LLMs can design sustainable concrete—A systematic benchmark. 2024. [doi: 10.21203/rs.3.rs-3913272/v1]
- [171] Hong. tuhz/Molecule-discovery-by-context: LLM hackthon release (0.1). Zenodo. 2023. [doi: 10.5281/zenodo.8122087]
- [172] Abramson J, Adler J, Dunger J, *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024, 630(8016): 493–500. [doi: 10.1038/s41586-024-07487-w]
- [173] Bleumink AG, Shikhule A. Keeping AI honest in education: Identifying GPT-generated text. Edukado AI Research, 2023: 1–5.
- [174] Liu R, Summers TR, Dasgupta I, *et al.* How do large language models navigate conflicts between honesty and helpfulness? arXiv:2402.07282, 2024.
- [175] Rashkin H, Reitter D, Tomar GS, *et al.* Increasing faithfulness in knowledge-grounded dialogue with controllable features. arXiv:2107.06963, 2021.
- [176] Wu ZQ, Galley M, Brockett C, Zhang YZ, Gao X, Quirk C, Koncel-Kedziorski R, Gao JF, Hajishirzi H, Ostendorf M, Dolan B. A controllable model of grounded response generation. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. Virtually: AAAI, 2021. 14085–14093. [doi: 10.1609/aaai.v35i16.17658]
- [177] Yu L, Cao M, Cheung JCK, Dong Y. Mechanistic understanding and mitigation of language model non-factual hallucinations.

- arXiv:2403.18167, 2024.
- [178] Yang YQ, Chern E, Qiu XP, Neubig G, Liu PF. Alignment for honesty. arXiv:2312.07000, 2023.
 - [179] Tonmoy SMTI, Zaman SMM, Jain V, Rani A, Rawte V, Chadha A, Das A. A comprehensive survey of hallucination mitigation techniques in large language models. arXiv:2401.01313, 2024.
 - [180] Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. Evaluating adversarial attacks against multiple fact verification systems. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 2944–2953. [doi: 10.18653/v1/D19-1292]
 - [181] Zhang QS, Wu YN, Zhu SC. Interpretable convolutional neural networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8827–8836. [doi: 10.1109/CVPR.2018.00920]
 - [182] Chen CF, Li O, Tao DF, Barnett AJ, Su J, Rudin C. *This looks like that*: Deep learning for interpretable image recognition. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 8930–8941.
 - [183] Wang DL, Yang S, Ouyang WL, Li BP, Zhou Y. Explainability of artificial intelligence: Development and application. Computer Science, 2023, 50(6A): 220600212 (in Chinese with English abstract). [doi: 10.11896/jsjx.220600212]
 - [184] Ribeiro M, Singh S, Guestrin C. “Why should I trust you?”: Explaining the predictions of any classifier. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016. 1135–1144. [doi: 10.1145/2939672.2939778]
 - [185] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.
 - [186] Zhao L, Peng X, Chen YX, Kapadia M, Metaxas DN. Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 6527–6536. [doi: 10.1109/CVPR42600.2020.00656]
 - [187] Sundararajan M, Taly A, Yan QQ. Axiomatic attribution for deep networks. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: JMLR.org, 2017. 3319–3328.
 - [188] Smilov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: Removing noise by adding noise. arXiv:1706.03825, 2017.
 - [189] Lin ZH, Feng MW, dos Santos CN, Yu M, Xiang B, Zhou BW, Bengio Y. A structured self-attentive sentence embedding. In: Proc. of the 5th Int'l Conf. on Learning Representations. 2017.
 - [190] Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, Guidotti R, Del Ser J, Díaz-Rodríguez N, Herrera F. Explainable artificial intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, 2023, 99: 101805. [doi: 10.1016/j.inffus.2023.101805]
 - [191] Michaud EJ, Liu ZM, Girit U, Tegmark M. The quantization model of neural scaling. arXiv:2303.13506, 2024.
 - [192] Villalobos P, Ho A, Sevilla J, Besiroglu T, Heim L, Hobbahn M, Ho A. Will we run out of data? Limits of LLM scaling based on human-generated data. arXiv:2211.04325, 2024.
 - [193] Li BH, Hou YT, Che WX. Data augmentation approaches in natural language processing: A survey. arXiv:2110.01852, 2022.
 - [194] Li W, Gao C, Niu GC, Xiao XY, Liu H, Liu JC, Wu H, Wang HF. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing (Vol. 1: Long Papers). Association for Computational Linguistics, 2021. 2592–2607. [doi: 10.18653/v1/2021.acl-long.202]
 - [195] Thawani A, Pujara J, Ilievski F, Szekely P. Representing numbers in NLP: A survey and a vision. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021. 644–656. [doi: 10.18653/v1/2021.naacl-main.53]
 - [196] Kıcıman E, Ness R, Sharma A, Tan C. Causal reasoning and large language models: Opening a new frontier for causality. arXiv:2305.00050, 2023.
 - [197] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Computing Surveys, 2018, 51(5): 93. [doi: 10.1145/3236009]
 - [198] Liang Z, Wang HZ, Dai JJ, Shao XY, Ding XO, Mu TY. Interpretability of entity matching based on pre-trained language model. Ruan Jian Xue Bao/Journal of Software, 2023, 34(3): 1087–1108 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/6794.htm> [doi: 10.13328/j.cnki.jos.006794]
 - [199] Amini A, Gabriel S, Lin SC, Koncel-Kedziorski R, Choi Y, Hajishirzi H. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 2357–2367. [doi: 10.18653/v1/N19-1245]

- [200] Zhou MT, Huang ML, Zhu XY. An interpretable reasoning network for multi-relation question answering. In: Proc. of the 27th Int'l Conf. on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018. 2010–2022.
- [201] Li W, Wu WH, Chen MY, Liu JC, Xiao XY, Wu H. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. arXiv:2203.05227, 2022.
- [202] Li G, Peng X, Wang QX, Xie T, Jin Z, Wang J, Ma XX, Li XD. Challenges from LLMs as a natural language based human-machine collaborative tool for software development and evolution. Ruan Jian Xue Bao/Journal of Software, 2023, 34(10): 4601–4606 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/7008.htm> [doi: 10.13328/j.cnki.jos.007008]

附中文参考文献:

- [1] 余同瑞, 金冉, 韩晓臻, 李家辉, 郁婷. 自然语言处理预训练模型的研究综述. 计算机工程与应用, 2020, 56(23): 12–22. [doi: 10.3778/j.issn.1002-8331.2006-0040]
- [183] 王冬丽, 杨珊, 欧阳万里, 李抱朴, 周彦. 人工智能可解释性: 发展与应用. 计算机科学, 2023, 50(6A): 220600212. [doi: 10.11896/jsjcx.220600212]
- [198] 梁峥, 王宏志, 戴加佳, 邵心玥, 丁小欧, 穆添愉. 预训练语言模型实体匹配的可解释性. 软件学报, 2023, 34(3): 1087–1108. <http://www.jos.org.cn/1000-9825/6794.htm> [doi: 10.13328/j.cnki.jos.006794]
- [202] 李戈, 彭鑫, 王千祥, 谢涛, 金芝, 王戟, 马晓星, 李宣东. 大模型: 基于自然交互的人机协同软件开发与演化工具带来的挑战. 软件学报, 2023, 34(10): 4601–4606. <http://www.jos.org.cn/1000-9825/7008.htm> [doi: 10.13328/j.cnki.jos.007008]



刘泽垣(2001—), 男, 博士生, CCF 学生会员, 主要研究领域为可信机器学习, 大模型技术研究.



张欣(1987—), 女, 博士, 教授, 博士生导师, 主要研究领域为人工智能治理.



王鹏江(1994—), 男, 博士, 助理研究员, 主要研究领域为大模型技术及其应用.



江奔奔(1987—), 男, 博士, 副教授, 博士生导师, 主要研究领域为可信机器学习.



宋晓斌(2001—), 男, 博士生, 主要研究领域为计算智能.