【大型语言模型幻觉研究综述】

——【A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Question】

1 相关资源

pdf: https://arxiv.org/pdf/2311.05232
ppt:
短视频:
数据集:
源码:
网站:
【除了网站，其他资源尽量下载】

2 论文属性

论文来源: AAAI-25
【给出具体会议名称和年限，不要仅仅写 ACM，IEEE】
论文类别: Large Language Model, Hallucination
【论文的类别，比如移动计算、轨迹处理、深度学习等】
论文关键字: LLMs, Hallucination
推荐程度：3 （其他说明可标注）
（5 非常棒，建议认真研读、小组讨论和复现；4 好，建议细读，考虑复现；3 可以，部分内容值得注意；2 一般，简单浏览即可；1 没有意义，不建议阅读）

3 工作团队

作者: LEI HUANG, WEIJIANG YU, WEITAO MA, WEIHONG ZHONG, ZHANGYIN FENG, HAOTIAN WANG, QIANGLONG CHEN, WEIHUA PENG, XIAOCHENG FENG*, BING QIN, TING LI
单位:

1. Harbin Institute of Technology, China
2. Huawei Inc., China

团队情况描述:

4 论文介绍

(1) 研究目的

【研究背景是什么？本文工作有什么用?】

大型语言模型在自然语言处理领域取得的显著进展，尤其是在语言理解、生成和推理方面。然而，LLMs 的一个主要问题是其生成内容中可能出现的“幻觉”现象，即生成的内容虽然看似合理，但与现实世界中的事实不符。这种幻觉现象对 LLMs 在信息检索系统中的实际应用构成了严重挑战，尤其是在需要高可靠性的场景中，如聊天机器人、搜索引擎和推荐系统等。

本文的目的是全面综述 LLMs 中的幻觉现象，包括其定义、分类、成因、检测方法、评估基准以及缓解策略，旨在为研究人员和实践者提供一个系统的框架，以更好地理解和解决 LLMs 中的幻觉问题，从而推动更可靠的信息检索系统的发展。

(2) 研究现状

【当前的最好研究做到什么程度了？存在的问题是什么？这里采信论文的说法，可以给出自己的评点】

当前关于 LLMs 幻觉的研究已经取得了一定的进展。研究者们提出了多种幻觉检测方法，包括基于事实核查的方法和基于模型内部不确定性的估计方法。此外，还开发了一些评估 LLMs 幻觉的基准数据集，用于量化 LLMs 生成内容的准确性。

然而，现有的研究仍存在一些问题。首先，幻觉的定义和分类尚未统一，尤其是在 LLMs 的开放性应用中，幻觉的边界更加模糊。其次，现有的幻觉检测方法大多依赖于外部知识源，这在某些情况下可能不可行或效率低下。此外，现有的缓解策略虽然在一定程度上有效，但往往需要大量的训练数据或计算资源，难以大规模应用。

本文认为，需要一个更细致的幻觉分类体系，以及更高效、更通用的幻觉检测和缓解方法。

(3) 本文解决的问题

【一句话概括本文解决的核心问题】

本文的核心问题是系统性地理解和解决 LLMs 中的幻觉现象，包括提出一个更细致的幻觉分类体系，分析幻觉的成因，并综述幻觉检测和缓解方法，以及评估基准，为未来的研究提供一个全面的框架。

(4) 创新与优势

【本文的创新之处是什么？新场景？新发现？新视角？新方法？请明确指出】

【本文工作的贡献或优点是什么?】

1. 提出了一个针对 LLMs 幻觉的全新分类体系，将幻觉分为“事实性幻觉”和“一致性幻觉”，并进一步细分为多个子类型。
2. 全面综述了幻觉检测和缓解方法，包括最新的技术进展和评估基准，为研究人员提供了一个系统的参考。
3. 为 LLMs 幻觉研究提供了一个全面的综述框架，有助于推动该领域的进一步发展。

(5) 解决思路

【本文是怎样解决问题的？包括方法、技术、模型等，以自己理解的方式表述清楚】

大模型的幻觉分类

大模型的训练主要分为三个阶段：预训练、监督微调和基于人类反馈的强化学习。分析这些阶段有助于理解 LLMs 产生幻觉的根源，因为每个阶段都赋予模型特定的能力。

大模型的幻觉主要分为两类：事实性幻觉和忠实性幻觉。前者强调生成内容与可验证现实事实的偏差，通常表现为事实矛盾；后者则关注生成内容与用户输入的偏离或内容自身的非一致性。
事实性矛盾又分为两类，事实矛盾和事实捏造：

- 事实矛盾指大语言模型输出中包含可追溯至现实世界信息但存在矛盾的事实。这类幻觉发生频率最高，其产生根源多样，涉及大语言模型对事实知识的捕获、存储和表达。根据矛盾错误类型，可进一步分为两个子类：实体错误幻觉和关系错误幻觉。
- 事实捏造指大语言模型输出中包含无法通过现有现实世界知识验证的事实。这又可细分为不可验证性幻觉和过度断言幻觉。

忠实性幻觉又分为三类，指令不一致、上下文不一致和逻辑不一致：

- 指令不一致指大语言模型输出偏离用户指令的情况。
- 上下文不一致指大语言模型输出与用户提供语境信息不忠实的情况。
- 当大语言模型输出出现内部逻辑矛盾时，逻辑不一致性尤为凸显，这在推理任务中经常被观察到。这种不一致既体现在推理步骤之间，也存在于推理步骤与最终答案之间。

幻觉成因

大语言模型幻觉的根本原因主要归为三个关键方面：数据、训练、推理三个方面。训练大语言模型的数据包含两个主要组成部分：预训练数据（模型通过该数据获得通用能力和事实知识）和对齐数据（用于教导模型遵循用户指令并与人类价值观对齐）。尽管这些数据不断拓展大语言模型的能力边界，却也无意中成为模型幻觉的主要源头，主要体现在三个方面：

- 缺陷预训练数据源中的错误信息与偏见：大模型的固有记忆能力在对抗幻觉时是把双刃剑：一方面表明大语言模型具备获取深层世界知识的潜力；另一方面当预训练数据存在错误信息和偏见时，可能无意中被放大，表现为模仿性虚假和社会偏见的强化。
- 预训练数据范围固化的知识边界：尽管海量预训练语料赋予大语言模型广泛的事实性知识，它们本质上仍存在知识边界。这种边界主要源于两方面：(1) LLM 无法记忆预训练时接触的所有事实性知识，尤其是低频长尾知识；(2) 预训练数据本身的固有边界，其不包含快速更新的世界知识或受版权限制的内容。因此当 LLM 遇到超出其有限知识边界的信息时，更容易产生幻觉。
- 劣质对齐数据诱发的幻觉：预训练阶段后，LLM 已在参数中嵌入大量事实性知识，形成明确知识边界。在监督微调阶段，LLM 通常使用人工标注的指令对进行训练，这可能引入超出预训练知识边界的新事实知识。

不同训练阶段赋予 LLM 不同能力：预训练侧重获取通用表征和世界知识，对齐阶段使 LLM 更契合用户指令与偏好。虽然这些阶段对塑造 LLM 卓越能力至关重要，但任一阶段的缺陷都可能无意中引发幻觉。

- 预训练阶段的幻觉现象：该阶段采用因果语言建模目标，模型仅通过从左到右的单向方式学习预测后续标记。虽然这种设计有利于高效训练，但本质上限制了捕捉复杂上下文依赖的能力，可能增加幻觉产生的风险。
- 监督微调引发的幻觉：LLM 在预训练阶段已建立固有能力边界，监督微调旨在通过指令数据及对应响应来释放这些预获取能力。但当标注指令要求超出模型预设能力边界时，LLM 会被训练生成超越其真实知识边界的响应。
- 基于人类反馈的强化学习导致的幻觉：研究表明 LLM 的激活状态包含其生成陈述真实性的内部信念，但这些信念与输出间可能存在偏差。即便经过人类反馈优化，LLM 仍可能产生背离内部信念的输出。

解码策略在 LLM 预训练和对齐后能力展现中至关重要，但某些解码缺陷会导致幻觉产生。

- 不完美的解码策略：解码随机性带来的多样性代价是幻觉风险正相关提升。采样温度升高会使标记概率分布更均匀，增加从分布尾部采样低频标记的可能性，这种对罕见标记的采样倾向会加剧幻觉风险。
- 过度自信问题：尽管主要采用因果语言模型架构的大语言模型已广泛应用，这种过度自信现象依然存在。在生成过程中，下一个词的预测同时受语言模型上下文和已生成文本影响。
- Softmax 瓶颈：大多数语言模型利用作用于最后一层表征的 softmax 层，结合词嵌入来计算词汇预测的最终概率。然而基于 softmax 的语言模型存在一个公认限制——Softmax 瓶颈，即 softmax 与分布式词嵌入的结合会制约给定上下文下输出概率分布的表现力，导致语言模型无法输出理想分布。
- 推理失败：若问题间存在多重关联，受限于推理能力仍可能无法给出准确结果。有研究人员还揭示了大语言模型中特定的推理缺陷——逆转诅咒现象：当问题表述为"A 是 B" 时模型能正确回答，但反向提问"B 是 A" 时却出现逻辑推导失败。这种推理差异不仅存在于简单推论中。

幻觉检测与基准测试

事实性幻觉检测：需评估大语言模型输出是否符合现实世界事实，典型方法主要分为两类：事实核查——通过可信知识源验证生成回答的事实准确性；不确定性估计——通过内部不确定性信号检测事实不一致性。

忠实性幻觉检测：确保大语言模型遵循上下文或用户指令的忠实性，对其在信息检索应用中的实用价值至关重要。作者将现有面向忠实性的幻觉检测指标分为：

- 基于事实: 测量生成内容与源内容关键事实的重合度
- 基于分类器的指标: 利用训练好的分类器判别生成内容与源内容的蕴含关系
- 基于问答的指标: 采用问答系统验证源内容与生成内容的信息一致性
- 不确定性估计: 通过测量模型对生成结果的置信度评估忠实性
- 基于提示的指标: 诱导大语言模型作为评估者，通过特定提示策略判断生成内容的忠实性。

鉴于大语言模型擅长记忆高频常识知识，当前幻觉评估基准主要针对长尾知识和易引发模仿性虚假的难题。评估方式通常采用选择题 QA（以准确率衡量性能）或生成式 QA（通过人工评判或代理模型打分）。

幻觉缓解策略

作者根据幻觉的根本原因对这些方法进行系统分类：针对数据相关幻觉的解决方案、训练相关幻觉和推理相关幻觉，每种方案都针对特定成因的固有挑战提供定制化对策。

数据相关幻觉的本质源于预训练数据中的错误信息、偏见和知识缺口。现有缓解方法主要分为三类：(1) 数据过滤——筛选高质量数据避免引入错误和偏见；(2) 模型编辑——通过修改模型参数注入最新知识；(3) 检索增强生成——利用外部非参数化数据库补充知识。

训练相关幻觉通常源于 LLM 架构固有局限及采用的训练策略。（1）减轻预训练相关幻觉的重要研究方向聚焦于模型架构固有的局限性，特别是单向表征和注意力机制缺陷。针对此问题，大量研究致力于设计专门解决这些缺陷的新型模型架构。（2）对齐阶段产生的幻觉通常源于能力错位和信念错位。然而界定大语言模型的知识边界存在挑战，难以弥合模型固有能力与人工标注数据知识之间的鸿沟。目前针对能力错位的研究有限，焦点主要集中在信念错位方面。

大型语言模型的解码策略对生成内容的真实性与忠实度具有决定性作用。（1）真实性增强解码：该方法通过优先保证生成信息的事实准确性，提升大型语言模型输出的可靠性。这类方法致力于使模型输出与现实世界既定事实紧密对齐，从而降低传播错误或误导性信息的风险。（2）忠实性增强解码：该方法优先考虑与给定语境的对齐，同时强调增强生成内容的内在一致性。本节我们将现有工作归纳为语境一致性与逻辑一致性两类。

检索增强生成中的幻觉现象

RAG 中的幻觉现象具有高度复杂性，表现为事实错误或误导性输出，其成因包括生成内容与现实不符、未能准确反映用户查询、或缺乏检索依据。这些幻觉主要源于两大因素：检索失败与生成瓶颈。

检索失败：检索环节是 RAG 框架的关键初始步骤，负责为信息查询获取最相关内容。因此，检索阶段的失误会对整个 RAG 流程产生严重连锁反应，最终导致幻觉。这些失败通常源于三个核心环节：用户查询的表述、检索源的可靠性与覆盖范围、以及检索器的有效性。

生成瓶颈：检索流程结束后，生成阶段成为关键环节，负责产出忠实反映检索信息的内容。但该阶段可能出现导致幻觉的严重瓶颈。我们总结了与这些瓶颈密切相关的 LLMs 两大核心能力：上下文感知与上下文对齐，二者对保障 RAG 系统可靠性至关重要。

(6) 可改进的地方

【本文工作的局限性是什么？你觉得可以从哪些方面改进工作?】

本文提出的分类很合理，总结的方法也很全面，没有什么可改进的地方。

(7) 可借鉴的地方

【你觉得本文哪些方面可以借鉴？比如思路、方法、技术等】

本文的幻觉分类体系和成因分析提供了一个清晰的研究框架，可以借鉴用于其他类型的生成模型或任务。此外，本文综述的幻觉检测和缓解方法提供了丰富的参考，可以启发在自己的研究中探索新的方法或改进现有方法。最后，本文对评估基准的总结也提供了选择和设计评估数据集的参考。

(8) 其他收获

【你有什么其他收获吗？比如了解了哪些团队和大牛在某领域做得很好，某类问题通常用什么技术解决，某些技术之间存在什么样的关联，某些会议和期刊在某领域很知名……】

本文让我了解到，幻觉问题通常涉及模型架构、训练数据和解码策略等多个方面，需要综合考虑才能有效解决。在技术上，基于事实核查和模型内部不确定性的估计方法是当前幻觉检测的主流技术，而检索增强生成和模型编辑则是缓解幻觉的有效策略。

5 评阅人

姓名:

时间: