实证研究汉语言文学垂直AI对于消除AI幻觉的意义

郭棋慧　索广超　娄珠琳

（吉林大学）

【摘　要】通用大模型生成内容中的 AI 幻觉问题日益凸显，不仅会影响研究范式和学术生态，还可能对个人乃至社会产生恶劣影响。本项目针对汉语言文学专业特性，通过构建覆盖 10 万知识实体的专业知识图谱，利用增强检索生成技术推动垂直 AI 实践落地。实证研究采取对照实验方式，设计 85 组问答任务，涵盖基础认知、分析研究、实践应用三方面，以图表方式呈现实验结果并进行对比分析，深入论证汉语言文学专业垂直 AI 构建对于抑制 AI 幻觉的意义。

【关键词】AI 幻觉，垂直 AI，汉语言文学

大型语言模型（LLM）的问世标志着自然语言处理领域取得重大突破，推动了信息获取方式的根本性变革。但当面对超出训练数据覆盖范围或需要专业判断的任务时，LLM 常通过“合理编造”生成看似真实实则偏离问题本身意图甚至与现实世界知识相悖的虚假内容，这种现象被称为“AI 幻觉（Hallucination）[2]”。

本文通过阐释 AI 幻觉的定义、分类并剖析其产生的关键因素，针对 LLM 在汉语言文学领域存在的 AI 幻觉实证分析垂直 AI 如何通过 RAG 技术、语境约束强化等抑制 AI 幻觉，探讨其技术实现路径与应用价值。

一、AI 幻觉

1.定义

AI 幻觉是指大语言模型“在缺乏真实数据或有效约束的情况下，通过对其训练数据的过度泛化、不合理外推或创造性解释而生成的不准确、不真实、超现实甚至完全虚构的图像、文本、声音或视频内容[1]”。

2.分类

现有研究将 AI 幻觉分为两大类：内在幻觉和外在幻觉[2]。前者指 AI 生成的输出与输入内容相矛盾[3]，这种错误可以通过校验输入内容轻松识别；而后者指生成内容的正确性无法从输入内容中验证，通常是模型调用了输入内容之外的数据或文本的结果[3]，往往更为常见且难以识别。

然而，鉴于大语言模型高度重视以用户为中心的交互方式，且其产生的幻觉主要出现在事实层面，在此基础上研究者引入了更为细致的分类体系，即事实性幻觉和忠实性幻觉两种典型类别[2]。

事实性幻觉强调生成内容与可验证现实事实之间的差异，这种不一致性可进一步细分为事实矛盾（主要有实体错误和关系错误两大子类）和事实捏造，分别指 LLM 的输出内容存在与现实信息相矛盾的事实和输出结果无法利用现实世界知识进行检验。[2]忠实性幻觉聚焦于生成内容与用户输入的偏差，或生成内容内部缺乏自洽性，具体可细化为指令不一致（LLM 输出内容与用户指令存在偏差）、上下文不一致（LLM 输出内容与用户提供的上下文信息不符）、逻辑不一致（LLM 输出存在内部逻辑矛盾）。

3.产生原因

大语言模型 AI 幻觉现象的形成是多维因素共同作用的结果，贯穿其能力获取的全过程。

生成式人工智的核心在于经过大量数据投喂和训练的大语言模型，正是这些模型在数据、训练和推理环节所固有的局限性[4]，构成了引发 AI 幻觉的基本前提。预训练数据源不可避免地包含错误、过时、不完整或带有偏见的信息，这些数据的质量问题会直接影响 LLM 的准确性。同时，数据源知识分布的非均匀性使 LLM 在特定领域存在短板，且具有知识时效性不足、难以获取受版权保护的知识等问题。

4.现实危害与消除必要性

AI 幻觉目前已成为大语言模型研究的焦点。[5]其具备的逻辑自洽性、错误隐蔽性和时间动态性[6]不仅使普通公众难以辨别，还会随着训练数据规模的扩大提升识别打击工作的难度。如何破解 AI 幻觉技术难题，构建多维度治理体系，已成为当下网络舆论社会治理的新命题[7]。

《2025 年全球风险报告》将“错误和虚假信息”纳入全球面临的五大风险之一，而 AI 开始成为制造和放大错误、虚假信息的重要一环，不仅会影响公共话语的构建与社会认知的形成，还可能加剧数字鸿沟效应，威胁社会稳定与公平正义。同时，AI 幻觉也会对个体造成无形冲击：用户对 AI 生成内容的过度依赖会导致个体认知能力、批判思维的退化，知识民粹主义风险加大，甚至诱发个体信任危机并冲击传统道德观念。

二、汉语言文学垂直 AI

（一）项目现有成果

本项目依托 AI 教学中心平台，以不同版本教材、教学大纲为数据源搭建课程知识基础框架，同时将拓展文献资料（如学术论文、经典评论）、文化案例（如文学作品改编影视、文化活动）、相关人物（如作家、评论家）等资源，通过知识抽取、知识融合、知识存储等过程初步制作了中国现

当代文学、欧美文学、古代汉语、现代汉语、语言学概论、文学理论等核心课程的知识图谱。该知识图谱系统包含 8 类 10 万的知识实体，可以实现概念阐释、文本解读、文献检索等多类问题的问答。

（二）抑制 AI 幻觉的机制

目前学术界比较公认的减少 AI 幻觉的方法主要有模型优化、数据过滤、增强检索生成等[2]。

本研究通过引入团队自主制作的专业课程知识图谱，对接外部数据库以拓展知识边界，使 LLM 不再受限于预训练数据的范围，提升准确性与专业性；同时，通过知识图谱实时动态更时解决知识时效性不足问题，并从结构化、半结构化数据库等多个渠道获取知识，构建多元化的数据源体系，促进数据多样性。不仅如此，本研究采用循环检索和事后检索方法，既通过接入知识图谱让 LLM 在文本生成过程中通过多次检索迭代实现动态信息整合[2]，还在生成答案后对 LLM 的生成内容进行优化和事实核查，促进 AI 自主学习、持续积累知识。

（三）实证研究

1.实验设计

本实验采取对照实验方式，由 85 个问答任务组成，分别对通用大模型和垂直大模型进行提问并统计 AI 幻觉的类型、出现次数和概率。大语言模型采用 eepseek、豆包等通用大模型，垂直大模型采用依托指导教师虚拟教研室平台开发的垂直 AI。

2.AI 幻觉评估标准

评估采用三档评分：“无幻觉”（完全符合事实与逻辑）、“轻度幻觉”（存在<3 处非关键性错误）、“重度幻觉”（关键性错误>=3 处，或存在主要观点判断错误、核心逻辑断裂、大量内容虚构）。统计时将“轻度”与“重度”合并为“存在幻觉”。

3.数据收集与分析

(1)对比分析

由于回答界面无法以图片形式堆积呈现，本实验采用对比表格形式，先从 85 个样本中抽取 15 个样本回答，提取 LLM 输出内容中的 AI 幻觉原文并进行分析，再与垂直大模型的输出内容进行对比。二者在汉语言文学领域的 AI 幻觉主要存在于分析研究和实验应用领域（均>50%），以事实性错误为主。

（2）典型案例呈现

以下抽取文献检索与考据人任务类型中的一例样本进行对比分析。

1.问题：解释“北冥有鱼，其名为鲲”这句话中“冥”的意思。给出相应的文献考据依据的原文及其出处。

通用大模型（deepseek）：

“北冥有鱼，其名为鲲”出自《庄子·内篇·逍遥游》，其中“冥”的含义需结合古籍注疏及文字学进行考证。其核心释义如下：

“冥”即“溟”：幽暗深邃之海。

1. 通假关系与核心释义

“冥”在此处为“溟”的通假字，指幽暗无涯的深海。此说最早见于西晋郭象《庄子注》：“溟，谓南北海也。夫巨海深弘，不可穷测。”（郭象《庄子注》）唐代成玄英《庄子疏》进一步阐释：“溟，犹海也。取其溟漠无涯，故谓之溟……北冥，北海也。窅冥无极，故谓之冥。”

2. 文字学依据

《说文解字》释“冥”：“幽也。从日从六，冖声。”段玉裁注：“窈窕深黑之谓。”其本义指幽暗不明，与深海“黑不见底”的特征契合。《尔雅·释水》亦载：“溟，海也。”印证“溟”为海的专称。

3. 语境印证

庄子后文曰：“南冥者，天池也。”以“冥”对称南北，皆指浩瀚水域。东晋支遁《逍遥论》释：“北溟有鲲鱼，其广数千里”，明确将“溟”对应为鲲鹏栖居的广漠海域。

4. 文化意象延伸

汉代《淮南子·览冥训》言：“穷溟之南”，高诱注：“溟，海也。”此用法承袭庄子，强化“溟”作为幽邃海洋的文学意象，暗喻超越认知的宏大境界。

结论

“北冥”之“冥”，实通“溟”，特指北方幽暗无极的深海（北海）。其释义源自：

1. 郭象、成玄英等权威注疏的直训
2. 《说文》《尔雅》的字义关联
3. 《庄子》文本的南北对称结构
4. 汉晋文献的承袭用法

参考文献原文节录

1. 郭象《庄子注》（《庄子集释》引）：“溟，谓南北海也。”
2. 成玄英《庄子疏》：“溟，犹海也……北冥，北海也。窅冥无极，故谓之冥。”
3. 《说文解字·冥部》：“冥，幽也。”
4. 《尔雅·释水》：“溟，海也。”
5. 《淮南子·览冥训》高诱注：“穷溟之南……溟，海也。”

垂直大模型（AI 助教）：

“北冥有鱼，其名为鲲”出自《庄子·逍遥游》。其中“冥”有多重含义，根据文献考据可以理解为“海”或“幽暗、深远之地”。以下是关于“冥”的具体释义及原文出处。

释义：

冥作“海”解释：

古代典籍中，“冥”通“溟”，常用来指代宽广的海或深

远的水域。

如《博物志·地》:“海之言昏晦无所睹。”

《释名·释水第四》:“海，晦也，主承秽浊，其水黑如晦也。” 以上两则解释将海的属性（昏暗、深远）和冥相连，因此“北冥”可译为北方无涯的海。

冥作“幽暗”或玄远之地解释：

《庄子·知北游》:“夫昭昭生于冥冥，有伦生于无形。”这里冥与昭（光明）相对，代表幽暗深远之地。

明释德清在《庄子内篇注》中解释:“‘北冥’，即北海，以旷远非世人所见之地，以喻玄冥大道。”此注强调北冥象征哲学意义上的玄远与深邃。

冥与“溟”的通用解释：

《经典释文·庄子音义》唐陆德明:“‘冥’,本亦作‘溟’。嵇康云：取其溟漠无涯也。”

郭象《庄子注》中记载:“北冥，海也，取其溟溟，无涯也。”

总结考据原文及出处：

《博物志·地》：海之言昏晦无所睹。

《释名·释水第四》：海，晦也，主承秽浊，其水黑如晦也。

《经典释文·庄子音义》(唐陆德明):‘冥’,本亦作‘溟’，嵇康云：取其溟漠无涯也。

郭象《庄子注》：北冥，海也，取其溟溟，无涯也，水黑色谓乏。

明释德清《庄子内篇注》:‘北冥’，即北海，以旷远非世人所见之地，以喻玄冥大道。

由此可见,“冥”既有物理意义上的“海”，也有象征意义上的“幽暗深远之地”，结合上下文是指广阔无际、玄远的境域或空间。

对比分析：

幻觉辨析：DeepSeek 在答案生成过程中进行了深度思考和联网搜索，所调用知识源覆盖 30 余个网络页面资源。但其数据来源多而不精，以光明数字报、第一范文网、搜狗问问等非结构化或半结构化网络数据库为主体,同时包含微信公众号推文、微博博文等社交媒体平台用户生成内容。此类数据源普遍具有组织形式非标准化、数据描述不准确、传播路径碎片化等特征，需审慎评估其知识权威性。

虽然 AI 助教输出内容中无参考文献的版本、网址等详细信息，但列出了原文的详细出处和完整句段，便于研究者追溯验证（经查证 AI 幻觉 0 处）；且结合古代汉语、古代文学、文献学等多学科领域知识，通过多维度交叉验证和严格文献筛选显著降低 AI 幻觉，使结论更加准确可信。

综上，与 LLM 相比，垂直大模型在文献精准性、逻辑严谨性、专业深度、错误规避等方面对 AI 幻觉的消除具有不可替代的价值。

结语

尽管更专业、颗粒度更细的数据集不能从根本上彻底解决 AI 幻觉问题，但通过为模型提供包含更丰富、精确的信息，能增强 LLM 的泛化能力，显著改善 AI 幻觉问题。并且，垂直领域知识图谱的发展，对于促进领域知识的发展具有明显的作用，也是一个领域发展成熟的重要标志。

团队制作的汉语言文学专业知识图谱通过结构化的知识关联与多维度的语义网络,支持从基础事实查询到深度理论应用、从单一文本解读到跨域综合研究的全链条知识问答需求，为汉语言文学专业的教学、研究与学习提供智能化支撑。并通过检索增强生成技术，拓展通用大模型知识边界，降低幻觉风险，实现垂直 AI 实践落地，有效解决现有平台面临的信息庞杂、结构性差、专业性欠缺、检索效率低下等问题，满足对知识结构化、可视化的需求。同时推动虚拟教研室的建设发展，为汉语言文学专业教学赋能助力，提升教学质量，助推学术创新，满足个性化学习需求。

参考文献

[1]裴宏娇,郭瑞宇,杨志和.图书馆 AI 语言生成中的信息幻觉：信息素养视角下的风险规避策略[J].图书情报导刊,2024,9(12):36-43+53.

[2] HUANG L,YU W,MA W,et al.A survey on hallucination in large language models:Principles,taxonomy,challenges,and open questions[J].arXiv preprint,2023,arXiv:2311.05232.

[3]经羽伦,张殿元.生成式 AI 幻象的制造逻辑及其超真实建构的文化后果[J].山东师范大学学报(社会科学版),2024,69(05):113-126.

[4]陈万球,罗一人.生成式人工智能的“知识幻觉”及其风险治理探论[J].上海市社会主义学院学报,2024,(04):38-51.

[5]张诗瑶,冯家照,高建进.释放数据潜能引领智慧未来[N].光明日报,2024-05-26(003).

*基金项目：本文为 2025 年度吉林大学“大学生创新创业训练计划”创新训练项目“基于虚拟教研室的汉语言文学垂直 AI 开发与研究”（项目编号：202510183009）。

作者简介：

郭棋慧（2005.03- ），女，汉族，籍贯：山东省东营市，本科在读，研究方向：中国现当代文学。

娄珠琳（2004.06- ），女，汉族，籍贯：河南省新乡市，本科在读，研究方向：中国现当代文学。

索广超（2004.04－），男，汉族，籍贯：河北省任丘市，本科在读，研究方向：汉语言文字学。