头豹
LeadLeo

行业简报
大模型幻觉
对互联网信息的影响

深度解析大模型幻觉污染，
互联网信息生态将迎来哪些挑战与变革？

企业标签：深度求索DeepSeek、OpenAI、阿里通义、文心一言

AI变革行业创新发展

China Large Model Industry

中国ビッグモデル産業

撰写人：陈庆民

团队介绍

头豹是国内领先的行企研究原创内容平台和创新的数字化研究服务提供商。头豹在中国已布局3大研究院，拥有近百名资深分析师，头豹科创网(www.leadleo.com)拥有20万+注册用户，6,000+行业赛道覆盖及相关研究报告产出。

头豹打造了一系列产品及解决方案，包括数据库服务、行企研报服务、微估值及微尽调自动化产品、财务顾问服务、PR及IR服务，研究课程，以及分析师培训等。诚挚欢迎各界精英与头豹交流合作，请即通过邮件或来电咨询。

报告作者

袁栩聪
首席分析师
oliver.yuan@Leadleo.com

陈庆民
行业分析师
qingmin.chen@leadleo.com

头豹研究院

咨询/合作

网址：www.leadleo.com

电话：15999806788（袁先生）

电话：13080197867（李先生）

深圳市华润置地大厦E座4105室

大模型幻觉对互联网信息的影响——背景介绍

- 大模型制造的虚假信息正在严重侵蚀互联网信息的质量和可信度，导致虚假新闻泛滥、学术诚信受损，并对社会信任和公共安全构成一定程度上的威胁

大模型幻觉问题对互联网信息质量影响背景分析

❑ 大模型幻觉对互联网信息影响背景介绍

■ 大模型幻觉定义

大模型"幻觉"（Hallucination）：大模型"幻觉"指的是生成式AI（如语言模型、图像生成模型）在输出内容时，生成虚假或误导性的信息，这一技术缺陷可能导致系统生成的内容无法真实反映事实，或以误导性的方式呈现信息。幻觉通常发生在模型推理过程中，尤其是当生成内容缺乏足够的上下文或真实数据支持时。

■ 大模型幻觉内容对互联网信息的影响

大模型幻觉内容对互联网信息的影响指的是大模型生成的虚假或误导性信息对互联网信息的可信度、传播效率以及社会认知的侵蚀。随着信息量的增加和传播速度的加快，虚假信息的扩散对公众决策、社会信任以及行业合规性带来深远影响。

❑ 大模型幻觉对互联网信息影响现状

1 *MCN机构正在通过AI生成虚假新闻，扰乱网络秩序*

目前，一些MCN（多频道网络）机构利用AI生成大量虚假新闻，广泛传播并在多个网络平台上迅速扩散。例如，2024年江西南昌的一家MCN机构通过AI软件生产虚假文章，最高可达每日4,000至7,000篇。这些文章通常内容逻辑混乱、语言空洞，严重误导公众的判断和认知，进而扰乱了网络信息的正常秩序。虚假新闻不仅仅是信息质量低下的问题，它还可能与电信诈骗等违法犯罪活动相关，进一步加剧了社会治理的复杂性。因此，AI生成的虚假新闻不仅破坏了互联网的内容生态，也在一定程度上引发了社会安全风险。

2 *AI创作低质量小说影响读者体验与市场信任*

在一些小说平台上，账号借助AI"创作"并每日更新多本电子书。虽然这些书籍的产量高，但由于AI生成的内容通常存在逻辑混乱、语言空洞等问题，导致读者的阅读体验严重下降。长期以往，读者对电子书市场的信任度可能受到影响，导致平台用户流失。这一现象暴露了AI生成内容在创意产业中的局限性，尤其是在需要高质量叙事和文学价值的领域。AI虽然能够高效生产内容，但其创作水平往往难以满足专业领域的要求，且可能让市场充斥低质内容，降低整体文学水平。

3 *医学论文因虚假AI大模型生成配图被撤稿，学术诚信受损*

在学术界，AI生成的虚假配图问题逐渐引发关注。部分医学论文因使用了AI生成的虚假配图而被撤稿，这不仅损害了学术诚信，还严重影响了科研的可信度。这表明，AI在科学领域的应用风险依然巨大，尤其是在学术出版和科研数据的准确性要求非常高的环境中。虚假配图可能误导科研人员的实验结论，甚至可能造成对医学领域的误解，影响到科学知识的传播。因此，在学术研究中，AI生成内容的使用必须严格监管，确保不会对学术成果的质量和信任性造成负面影响。

来源：新华每日电讯，头豹研究院

大模型幻觉对互联网信息的影响——大模型幻觉特征

- 大模型的“幻觉”特征表现为虚构知识体系和误导性语义表达，并且通过低门槛使用和用户的广泛传播，这些内容极易形成了“知识污染链”，从而加剧了公众对AI技术的信任危机

大模型幻觉特征分析

随着AI大模型的发展，深度学习模型的“幻觉”问题逐渐成为学术界和行业中的关注焦点。尤其是像DeepSeek这样的先进模型，其“幻觉长城”现象呈现出特有的多维特征，结合了**高阶幻觉的学术化包装、中文表达的本地化优势与误导性、以及低门槛高传播性的特点**。这些特征不仅给用户带来困扰，也在一定程度上加剧了“知识污染”的问题，对行业的长远发展提出了严峻挑战。

1 *大模型的高阶幻觉可对信息进行学术化包装，伪知识体系的构建*

大模型的“幻觉”并非简单的常识错误，而是通过复杂的技巧和方法将其伪装成学术性强、逻辑严谨的“知识体系”。**例如，模型通过虚构专业术语、伪造文献编号（如“市规〔2020〕12号文”）以及跨学科的知识拼接（如将文物保护与土木工程振动理论结合）来构建虚假的逻辑闭环。**这种“幻觉”的最大特点在于它能够在表面上看起来非常合理，且能通过细致的语言和专业的术语增强其可信度，令普通用户和甚至一些专业人士很难通过常规手段验证其准确性。

虚构专业术语 | 伪造文献编号 | 跨学科知识拼贴 | 构建逻辑闭环的伪知识体系 | 难以验证性

2 *大模型在某些没有完全理解语义的情景下，能用强大的表达，形成“假装理解”的误导信息*

大模型能适应中文用户的语言习惯，其输出语句流畅度和逻辑推演能力显著优于人类，甚至能够模仿特定文风，这种能力使得大模型在处理中文文本时显得更为自然且符合用户的预期，容易形成假装理解用户的意思而进行表达，例如，在面对复杂的方言、俚语或隐含的文化背景时，大模型能够在没有理解其真实意思时，**用逻辑严密、语言流畅的方式给出答案，但却忽视了某些特定词汇或语境的深层意义，导致输出看似合理却实际上存在事实偏差。**

用户：用复杂方言提出极具地方特色或文化背景的问题，期待得到答案 → 大模型：完全没有理解用户的提问，但是可以用强大的表达，给出极具说服力的误导信息

3 *大模型的低门槛使用与高传播性将加剧互联网信息的污染*

一些大模型提供了免费、易获取的服务，迅速吸引了大量普通用户，特别是在小红书、公众号等社交平台上，基于大模型输出的内容广泛传播。这些内容往往包括虚构的音乐分析、历史解读等，通过SEO优化，这些内容快速占据搜索引擎的前列位置，形成了一个“知识污染链”。**其运作机制大致为，AI生成的虚假或误导性内容通过自媒体平台传播，用户通过搜索引擎接触到这些内容，进而影响模型的训练数据，导致下一代模型中产生更多幻觉现象。**随着这一循环的加剧，AI模型的幻觉问题愈发严重，而由于内容的广泛传播，模型输出的错误信息更难被纠正。尤其对于普通用户而言，这种“知识污染”往往难以察觉，从而加剧了公众对AI技术和大模型的信任危机。

AI大规模生成 → 用户自媒体传播 → 搜索引擎收录 → 投喂给下一代模型

来源：头豹研究院

大模型幻觉对互联网信息的影响——大模型幻觉原因

- 大模型幻觉的本质是数据、模型与算法等多层面技术缺陷的系统性体现。解决需综合数据清洗、知识增强、推理优化等策略，同时需根据应用场景（如医疗、工业）定制化治理方案

大模型幻觉原因分析——技术角度（数据层面）

❑ 数据层面的技术缺陷

随着大模型在各行业应用的普及，数据质量与数据分布问题已成为影响模型性能和可靠性的重要因素。尽管大模型在处理复杂任务时展现出了强大的能力，但其背后的数据问题却常常导致“幻觉”现象和错误输出。

1 *数据质量不足：噪声、过时信息和偏见内容的影响*

- **训练数据中的噪声与错误信息**

大模型的训练依赖于庞大的数据集，其中包含了来自互联网上的海量信息。然而，这些数据并非完美无瑕，往往包含噪声、虚假信息和过时内容。模型在训练过程中会将这些不准确的信息吸收并用于推理和生成，最终导致输出错误的知识。例如，互联网中的虚假历史事件或医疗错误数据，可能被模型错误地识别为事实，并在生成时复现这些错误。更为严重的是，由于大模型对数据的过度依赖，它可能将这些错误数据视为“可信”知识，在后续的任务中不断放大其影响。

- **偏见内容的传播**

除了噪声和错误信息，训练数据中的偏见内容也会被大模型吸收并加以放大。例如，训练数据中存在对某些社会群体、性别或文化的负面刻板印象，模型在生成内容时会无意识地加强这些偏见，进而影响最终的输出。

2 *数据分布不匹配：源数据与目标任务数据的差异*

- **数据分布不一致的挑战**

大模型通常使用从不同领域和多种来源收集的数据进行训练，但在实际应用中，目标任务的数据分布往往与训练数据存在显著差异。这种数据分布不匹配问题，尤其在迁移学习中尤为常见，导致模型在生成时依赖错误的先验知识。例如，训练数据中使用的是来自互联网的文本，而目标任务的文本具有更强的领域特定性（如医学、法律等）。这种差异将导致模型输出的内容不准确，甚至与目标任务的实际需求完全不符。

- **启发式规则构造的数据问题**

许多大模型在训练时依赖启发式规则和半自动生成的数据，但这些数据往往缺乏足够的真实关联性，可能与实际场景有较大的偏差。例如，构建的训练数据是在特定情境下通过人工设计的规则进行构造，但这些数据并不代表真实世界的多样和复杂，导致模型在实际应用中产生偏差，影响决策质量。

3 *知识时效性滞后：静态知识库的局限性*

- **静态知识库的限制**

大模型通常依赖于在特定时间点收集和处理的静态知识库，这意味着模型所学习的知识往往是“历史性的”。在快速变化的领域（如医疗、法律、科技等），这一滞后效应尤为突出。例如，新药的研发进展、最新的政策变化或技术创新等，模型可能无法及时反映和处理这些新的信息。特别是在应对急需实时更新的任务时，模型生成的输出往往无法跟上知识更新的速度，导致输出过时或不符合当前实际情况。

来源：头豹研究院

（接上页——大模型幻觉原因）

- **时效性滞后带来的风险**

在某些敏感领域，知识时效性滞后的影响可能是灾难性的。例如，在通信领域，依赖过时标准的AI模型可能无法及时响应新的技术规范，导致系统运行故障或安全漏洞；在医疗领域，过时的治疗方案或药物信息可能会对患者安全产生严重威胁。随着技术的快速发展，模型的“知识冻结”不仅影响其当前的实用性，还可能引发更为严重的社会和安全问题。

大模型幻觉原因分析——技术角度（模型层面）

❑ 模型层面的技术缺陷

随着大模型在各行业的广泛应用，模型结构和训练机制的缺陷已经成为影响其可靠性、准确性和可用性的关键因素。从解码策略到过拟合问题，再到对齐过程的副作用，许多技术层面的挑战在推动大模型发展过程中不可避免地显现出来。这些缺陷不仅会降低模型的性能，还可能对用户带来严重的误导，特别是在敏感领域如医疗和金融等。

1 *解码策略与参数偏差：多样性与事实之间的平衡*

- **自回归生成与采样策略的影响**

在生成任务中，大模型常采用自回归生成（autoregressive generation）和采样策略（如top-p sampling）来增加输出的多样性。自回归生成意味着模型根据前一个词或输出进行推理和生成，每一步的输出都依赖于上一步的生成结果。这种策略能够提高生成内容的自然度，但在追求多样性的过程中，模型可能会偏离事实，生成缺乏准确性的内容。 例如，在开放式问题回答或创意生成时，模型为了避免重复和增加生成内容的多样性，会选择某些与现实相距较远的答案。采样策略（基于概率分布选择输出）进一步放大了这种偏离的风险，因其允许从多个可能的词汇中选择结果，即使这些词汇在语境中并不完全准确。

- **参数记忆偏差**

解码器对输入信息的关注偏差也是一个常见问题。大模型在生成过程中，可能更多地依赖于已学习的参数记忆而非当前的上下文信息，这种偏差会导致模型忽视当前输入的具体语境，从而产生不准确或错误的输出。尤其在处理具有高度上下文依赖的任务时（如医学文献生成、法律文书撰写等），这种偏差可能会导致输出信息的事实错误，甚至虚构内容。

2 *模型复杂度与过拟合：大模型的泛化能力问题*

- **大模型的参数量与过拟合风险**

大模型通常具有数以亿计的参数，这使得其在训练过程中非常容易陷入过拟合。过拟合是指模型在训练数据中表现良好，但在处理未见过的数据时，表现较差。这一现象发生的原因在于大模型对训练数据的高度适应，使得模型能够“记住”数据中的噪声和非典型信息，从而失去对数据普遍性规律的学习能力。在某些特定行业，如**工业领域**，高维度数据往往带有大量噪声，导致大模型过度拟合这些噪声，降低了模型的泛化能力。例如，工业数据中的异常值或采集过程中的误差，可能被模型当作有效信息处理，进而导致生成错误的预测结果。

- **复杂度与性能平衡**

大模型的高复杂度虽然在理论上具有强大的学习能力，但同时也增加了对数据质量的依赖。如果训练数据存在噪声、偏见或不完整的情况，大模型就容易过拟合这些问题，从而影响模型的稳定性和可靠性。这也是为何大模型需要在高效的正则化策略、数据清洗和增强训练样本多样性等方面进行优化。

来源：头豹研究院

（接上页——大模型幻觉原因）

3 对齐过程的副作用：迎合用户偏好与虚假信息生成

- **指令微调（Instruction Tuning）与强化学习（RLHF）的问题**

为了使大模型更好地符合用户需求和预期，许多大模型通过指令微调和基于人类反馈的强化学习（RLHF）进行调整。在这种过程中，模型不仅学习如何根据指令生成内容，还根据用户的反馈来优化其表现。这种过程的一个副作用是，模型可能会过度迎合用户的期望，甚至编造看似合理但实际上没有依据的答案。例如，在**医疗领域**，为了满足用户对快速解答和便捷建议的需求，模型可能会为用户提供未经验证的药品疗效或治疗方案，甚至虚构一些不存在的治疗方法。这种现象不仅影响了医疗服务的准确性，还可能对患者的健康造成直接威胁。

- **人类反馈与模型生成的脱节**

强化学习中的人类反馈机制虽然能够帮助模型更好地适应人类的需求，但也可能导致模型偏离科学或事实的标准，生成更符合用户期望但不真实的内容。特别是在复杂、专业的领域中，这种偏差可能给用户带来误导，并对行业的实际应用带来负面影响。

大模型幻觉原因分析——技术角度（推理与知识整合层面）

❑ 推理与知识整合层面的技术缺陷

随着大模型在各个领域的广泛应用，推理和知识整合问题逐渐显现，成为影响模型准确性和可靠性的关键因素。尽管大模型在自然语言处理和生成任务中表现出了强大的能力，但其在长上下文理解、逻辑推理、知识表示与调用方面的缺陷，仍然限制了其在复杂任务中的表现。尤其是在需要深度推理和高质量知识整合的场景下，大模型往往难以保持生成内容的准确性和一致性。

1 上下文处理局限性：长上下文理解与生成冲突

- **长上下文理解能力不足**

大模型在处理长文本或多轮对话时，常常面临**上下文理解能力不足**的问题。当前的大模型通常在生成内容时，依赖其输入的上下文信息进行推理和输出。然而，随着对话或文本的逐步推进，模型的记忆和关注能力往往随着输入内容的增加而衰减。模型通常只能处理有限的上下文范围，在长时间跨度的对话或多轮互动中，容易遗忘早期设定，从而导致生成内容与输入信息产生冲突或逻辑矛盾。例如，在一场多轮对话中，用户最初设定了某些条件或规则，但随着对话的深入，模型可能忽略之前的信息，进而在后续回答中出现逻辑上的不一致。这种情况尤其在涉及复杂逻辑推理或需多次引用早期设定的任务中更为明显，如自动化客服、法律咨询等场景中，容易导致用户体验不佳，甚至出现错误的结论。

2 黑箱特性与逻辑推理薄弱：缺乏显式推理能力

- **黑箱特性导致的推理问题**

大模型的“黑箱”特性使得其内部推理过程对用户不透明，且缺乏显式的推理能力。模型的生成基于海量数据和复杂的统计关联，而不是显式的逻辑推理。因此，模型很难保证生成内容的**逻辑一致性**。缺乏清晰的推理路径意味着模型在生成内容时，很难检查自己是否遵循了正确的推理步骤。例如，在**软件开发**中，若大模型生成代码时没有显式的推理和验证机制，可能会导致错误的代码输出，如生成未处理的异常或无法正常运行的逻辑。这对于需要高度可靠和准确性的软件开发任务来说，构成了巨大的挑战，尤其在自动化编程、代码补全等应用中，错误代码的生成可能引发系统故障或安全漏洞。

来源：头豹研究院

（接上页——大模型幻觉原因）

3 知识表示与调用缺陷：信息整合的不足与知识混淆

- **参数化知识存储方式的缺陷**

大模型在处理大量信息时，通常采用**参数化知识**的存储方式。虽然这种方式允许模型从庞大的数据中提取出潜在的知识，但它存在一定的局限性。模型通过参数来“记忆”知识，而这些参数并不能准确地表示事实或概念之间的关系。例如，模型可能在理解或生成内容时将相似概念混淆，从而导致错误输出。在**化学物质**等领域，模型可能将不同的化学元素或化合物误认为相同，产生错误的化学反应式或疗效推荐。

- **外部知识整合不足**

大模型的另一个局限性是缺乏有效的**外部知识整合**。在很多复杂领域，模型无法自主补充实时的缺失信息，尤其是在涉及快速发展的学科或需要实时更新的领域（如医学、金融等）。尽管一些大模型通过外挂知识库或知识图谱来增强自己的知识储备，但这些外部资源的准确性和全面性依赖于数据源的质量。例如，若外部数据源中存在错误或滞后信息，模型可能会基于不准确的知识生成内容，导致输出内容的准确性和时效性大打折扣。

大模型幻觉原因分析——技术角度（现有解决方案的技术层面）

❑ 现有解决方案的技术局限性

当前的大模型技术在自然语言处理和生成任务中取得了显著进展，但在实际应用中，现有解决方案仍然面临着诸多技术局限。这些局限性在模型的增量学习、多模态验证、提示工程、知识检索以及评估与修正机制等方面尤为明显。以下将从这些关键技术领域展开深入分析，探讨其面临的挑战以及对实际应用的影响。

1 增量学习与多模态验证的成本

- **增量学习的挑战**

增量学习指的是模型在不丢失已有知识的基础上，持续吸收新的数据并进行更新。这一过程通常需要定期的算力投入和标注资源。随着数据量和任务复杂性的增加，增量学习的难度和成本也随之上升。模型必须不断地学习新的信息，同时避免忘记已学到的内容，这要求在训练过程中采用特定的机制来平衡新旧知识的整合。例如，深度学习中常见的灾难性遗忘问题，就要求模型能够高效地进行知识存储和更新。然而，这一过程不仅需要大量的计算资源，还需人工标注大量的高质量数据，尤其在专业领域（如法律、医学）中，标注的准确性和全面性直接影响模型的效果和可靠性。

- **多模态验证的复杂性**

在多模态系统中，模型需要同时处理多种数据形式（如文本、图像、视频等），并进行交叉校验。这种多模态验证虽然能够提升生成内容的准确性，但面临着数据对齐和计算复杂度的重大挑战。不同模态之间的对齐问题主要体现在如何将不同格式的数据（例如，图像中的物体识别）正确地映射到一起，并进行一致性验证。而计算复杂度则主要体现在需要同时处理大量的输入数据并进行整合分析，这对于当前的硬件和算法来说，仍然是一项非常繁重的任务，导致其在实时性和效率上的表现有限。

2 提示工程与知识检索的领域局限

- **提示优化的局限性**

提示工程（Prompt Engineering）是通过调整输入提示来引导大模型生成更符合用户需求的输出。然而，在一些专业领域（如法律、医学等），提示优化的效果往往有限。尽管提示工程能够有效提升模型在某些通用任务中的表现，但在处理复杂、专业的领域时，模型的生成依然缺乏足够的专业深度。例如，在医学领域，虽然可以通过精心设计的提示来引导模型生成治疗建议，但由于缺乏足够的医学知识和临床经验，模型依然可能生成不准确或不安全的推荐。

来源：头豹研究院

（接上页——大模型幻觉原因）

- **知识检索引入外部噪声**

知识检索是大模型通过实时查询外部数据库或知识库来增强其知识储备和应对知识时效性问题的一种方法。然而，这一过程可能引入外部噪声，特别是在使用开放互联网数据源时。虽然知识库和文献库能够为模型提供额外的背景信息，但这些外部信息的质量参差不齐，可能导致错误或过时的知识被模型采纳，进而影响生成内容的准确性。尤其是在一些专业领域，外部数据源的不准确或信息滞后，可能会导致模型输出的知识错误，进一步加剧幻觉问题。

3 *评估与修正机制的不足*

- **现有评估方法的局限性**

目前，大多数评估方法（包括人工评估、规则评估和GPT-4打分等）在处理复杂任务时存在显著不足。人工评估虽然可以提供一定的准确性，但其高成本和主观性使得在大规模任务中的应用受到限制。规则评估虽然能够针对某些标准任务进行自动化评估，但其缺乏灵活性，无法应对多变和复杂的生成任务。使用GPT-4等大模型进行打分虽然在一定程度上提高了评估效率，但这些评估方法本身也受到模型“幻觉”的影响，无法完全反映输出内容的真实准确性。

- **修正机制的不足**

目前，模型的修正机制仍然不够成熟。针对生成内容中的事实性错误或逻辑错误，现有技术尚未能够提供便捷有效的修正方式。例如，直接编辑模型参数来修正事实错误的技术仍在不断发展中，且目前只能在较为简化的任务中实现有效应用。对于复杂和专业性较强的领域，修正错误的过程往往需要依赖大量的人工干预，而这一过程不仅耗时费力，还可能带来额外的错误风险。缺乏成熟的修正机制，使得大模型在面临生成错误时，修正过程变得十分困难。

来源：头豹研究院

大模型幻觉对互联网信息的影响——AI巨头抗击幻觉路径

- AI巨头在应对大模型幻觉问题时采取了不同的策略，主要集中在提高模型的准确性和可靠性上。其通过引入先进的推理机制、知识增强技术、实时验证流程以及去偏见算法，不断优化模型的生成能力

AI巨头抗击幻觉路径

从AI巨头如何对抗大模型幻觉的角度来看，技术公司如OpenAI、Google[illegible]解决大模型幻觉问题。这些技术公司不仅关注如何提高模型的准确性，[illegible]象对用户带来的负面影响。

1 *OpenAI的“过程监控”与数学解题策略*

OpenAI提出了通过引入“过程监控”来对训练模型进行逐步推理[illegible]题过程中更加注重逐步推理，而不仅仅是依赖于最终的答案。[illegible]处理复杂问题时的推理透明度，尤其在解决数学问题时，可[illegible]现象的发生。在实际应用中，OpenAI通过这种方法提升了[illegible]表明，在特定的领域（如数学解题），引入过程监控能够[illegible]的答案更具逻辑性和一致性。

2 *Google的“事实核查”与强化医学领域的准确性*

Google在应对大模型幻觉时采取了“事实核查”技术，[illegible]对准确性有极高的要求，错误的医学信息会[illegible]心链”[illegible]可信度，并能够减少[illegible]的发生[illegible]还增强了用户[illegible]

主笔分析师：qingmin.chen@leadleo.com
首席分析师：oliver.yuan@leadleo.com

3 *Anthropic的“政治[illegible]”*

Anthropic提出[illegible]对敏感或政治性问题时，能够避免生成[illegible]减少模型产生幻觉所导致的社会风险，尤其[illegible]息。具体来说，Anthropic的方案确保[illegible]处理政治敏感信息时具有显著的优势。[illegible]

4 *百度引入知识增强与深度语义理解*

百度通过其自研的文心大模型（ERNIE）[illegible]型中集成了知识图谱和外部数据库[illegible]的传播。通过与知识库的结合，[illegible]取准确的外部信息，从而减少生成[illegible]

5 *阿里通过多模态增强与[illegible]*

阿里巴巴通过其自研的达摩院[illegible]阿里巴巴特别注重跨模态[illegible]容的质量，还能有效减少[illegible]巴巴的大模型能够更好地[illegible]

来源：头豹研究院

大模型幻觉对互联网信息的影响——争议与解决方案

- 大模型幻觉问题需从技术、用户和生态三层面综合考虑应对。提高用户批判性思维、优化技术事实核查、建立生态内容溯源标识机制，可减少虚假信息传播，提升大模型可信度与应用价值

大模型幻觉对互联网信息影响的争议分析

❑ 大模型幻觉对互联网信息影响的争议

DeepSeek、ChatGPT等AI模型所生成的“幻觉”内容逐渐对互联网信息的质量和可信度产生了深远影响。这一现象引发了广泛的争议，尤其是在**技术缺陷与用户责任的辩论**、**信息生态的长期影响**以及**伦理与技术发展的平衡**等方面。

1 *技术缺陷 vs. 用户责任：责任归属的争议*

- **支持者观点：AI发展中的必然阶段**

支持者认为，DeepSeek等大模型的幻觉是人工智能发展的必经阶段，尤其是在模型复杂度不断提升的背景下，难免会出现一些不准确或错误的内容。他们指出，模型生成的“幻觉”并非故意编造，而是由于数据和参数的局限性所导致。例如，在分析方言或冷门知识时，由于训练数据的不足，模型可能无法准确理解语义，从而产生误判或错误解释。支持者认为，用户应当对AI输出保持批判性思维，并主动验证生成的信息，而不是盲目相信模型的所有输出。这一观点强调了**用户教育与意识的提升**，认为技术本身并非问题的根源，而是用户责任的体现。

- **反对者观点：模型应主动承担责任**

反对者则认为，普通用户缺乏足够的专业知识来鉴别高阶幻觉，尤其在涉及复杂专业领域时，模型可能生成看似合逻辑但实际上错误的信息。例如，模型可能在医学或法律领域生成错误的建议，普通用户难以察觉这种信息的虚假性。因此，反对者主张，模型应承担更多责任，通过**引用信源、标注不确定性**等方式减少误导风险。这不仅是为了提高模型的透明度，也是为了增强用户对模型输出的信任度。反对者认为，模型应通过内建机制确保信息的准确性，避免将责任完全转嫁给用户。

2 *信息生态的长期影响：伪知识的扩散*

- **虚假信息对传统知识传播链的颠覆**

DeepSeek等大模型生成的幻觉内容正在逐步改变互联网信息的传播格局。传统的知识传播链条通常是**专业论文到科普文章到网络内容**，而大模型的介入打破了这一层次关系，直接将“伪知识”推向了信息传播的高地。

- **伪知识的传播与信息污染**

伪知识的扩散意味着，大模型输出的虚假信息可能长期占据互联网的搜索引擎高位，进一步加剧了**信息污染**。虚假信息的泛滥不仅削弱了互联网信息的真实性，也使得传统的信息验证机制（如学术论文、权威数据库等）逐渐失去了主导地位。这种趋势可能长期影响中文互联网的可信度，导致用户在获取信息时越来越难以分辨真伪。

来源：头豹研究院

（接上页——争议与解决方案）

3 伦理与发展的平衡：开源与监管的挑战

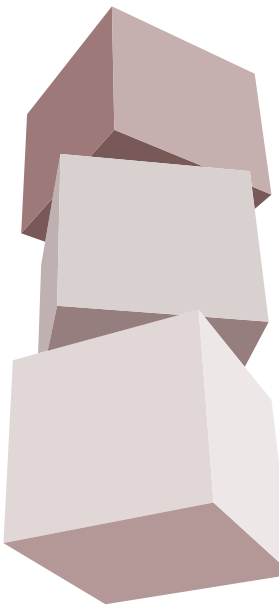
- **开源模式与伦理对齐问题**

大模型的开源模式极大推动了人工智能技术的创新和普及，但同时也带来了伦理上的隐忧。开源为开发者提供了更多的自由和灵活性，但也为不负责任的应用或滥用提供了可能。例如，恶意用户可以利用开源的大模型生成虚假信息，进一步加剧信息污染和社会信任危机。因此，如何平衡**技术自由与伦理监管**，成为了大模型发展中的关键问题。

- **可控开源与社区监督机制**

为了解决这一问题，研究者提出了“可控开源”的概念，即在开源的基础上，对一些高风险功能进行限制，或通过**社区监督机制**确保技术的合理应用。可控开源不仅能促进技术创新，还能够在一定程度上防止技术滥用。通过社区的共同监督和反馈，确保AI生成内容符合社会伦理和法律要求，从而减少其负面影响。

大模型幻觉对互联网信息影响的应对建议



用户层面：增强批判性思维与验证意识

从用户层面来看，必须增强公众对AI输出内容的**审慎态度**，尤其是在涉及专业领域时，用户应通过多渠道进行交叉验证，避免将未经验证的内容作为事实传播。用户教育与信息素养的提升将有助于减少虚假信息的扩散，并保护社会对AI技术的信任。

技术层面：优化模型的事实核查与信源检索

从技术角度来看，应优化模型的**事实核查模块**，使模型能够自动验证生成内容的真实性，减少错误信息的传播。同时，引入**实时信源检索功能**，确保模型能够根据最新的信息进行更新和调整，避免产生过时或错误的内容。此外，标注生成内容的不确定性和置信度，也是提升模型透明度和可信度的有效手段。

生态层面：构建内容溯源与标识机制

在生态层面，应该建立**AI生成内容的标识与溯源机制**，使得所有生成内容都能够追溯到其来源和生成过程。这不仅能提高内容的可验证性，还能在搜索引擎中对虚假信息的权重进行限制，减少伪知识在信息传播中的影响。通过这些措施，可以有效遏制虚假信息的扩散，保护信息生态的健康。

来源：头豹研究院

名词解释

- **大模型幻觉：**指大规模预训练模型在生成文本时，因缺乏足够的事实验证而产生虚假或不准确内容的现象。
- **自回归生成：**生成式模型的一种方式，依赖于前一步的输出进行下一步的生成，可能导致内容的连贯性受到影响。
- **训练数据：**用于训练大模型的原始数据集，数据的质量和多样性直接影响模型生成内容的准确性和可靠性。
- **过拟合：**指模型过度依赖训练数据中的噪声和特征，导致对新数据的泛化能力差，从而可能产生幻觉内容。
- **提示工程：**通过设计特定的输入提示来引导大模型生成更符合预期的输出，过于依赖提示可能引发幻觉问题。
- **事实核查：**对大模型生成的内容进行准确性验证的过程，减少幻觉现象的发生，尤其是在信息敏感领域。
- **多模态模型：**能够处理多种输入形式（如文本、图像、声音等）的模型，幻觉问题可能因数据融合不当而加剧。
- **上下文理解：**模型对输入信息中前后关系的理解能力，长文本中的上下文理解不足可能导致内容生成错误或逻辑矛盾。
- **知识图谱：**通过节点和边的方式构建的知识网络，帮助模型准确理解和引用事实，减少幻觉内容的生成。
- **推理链：**模型在生成内容时，按照逻辑步骤进行推理的过程，推理链不完整或错误可能导致幻觉现象。
- **模型微调：**在预训练基础上，通过特定任务的数据进一步优化模型，使其适应具体应用场景，减少幻觉现象。
- **增强学习：**一种通过奖惩机制优化模型行为的技术，应用于模型训练中以减少错误输出和幻觉内容。

方法论

- 头豹研究院布局中国市场，深入研究19大行业，持续跟踪532个垂直行业的市场变化，已沉淀超过100万行业研究价值数据元素，完成超过1万个独立的研究咨询项目。
- 研究院依托中国活跃的经济环境，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。
- 研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。
- 研究院密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。
- 研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

法律声明

业务合作

会员账号

可阅读全部原创报告和百万数据，提供PC及移动端，方便触达平台内容

定制报告/词条

行企研究多模态搜索引擎及数据库，募投可研、尽调、IRPR等研究咨询

定制白皮书

对产业及细分行业进行现状梳理和趋势洞察，输出全局观深度研究报告

招股书引用

研究覆盖国民经济19+核心产业，内容可授权引用至上市文件、年报

市场地位确认

对客户竞争优势进行评估和证明，助力企业价值提升及品牌影响力传播

云实习课程

依托完善行业研究体系，帮助学生掌握行业研究能力，丰富简历履历

业务热线

袁先生：15999806788

李先生：13080197867

成为头豹会员—享专属权益

- 成为头豹会员，尊享头豹海量数据库内容及定制化研究咨询服务
- 头豹已累积上万本行业报告、词条报告，拥有20万+注册用户，沉淀100万+原创数据元素
- 头豹优势：行业覆盖全、数据量庞大、研究内容应用场景广泛，并有专业分析师团队为您提供定制化服务，助力企业展业

报告次卡

任意10本报告
阅读权益（一年有效）

¥598 /年

企业标准版

适用于研究频次高的用户或企业
无限量阅读全站报告
升级报告下载量
专享企业服务
定制词条报告

¥50,000 /年

企业专业版/旗舰版

满足定制研究需求的企业用户
定制深度研究报告
随需下载报告
分析师一对一沟通
专享所有核心功能

¥150,000+ /年

购买与咨询

咨询邮箱：

nancy.wang@frostchina.com

客服电话：

400-072-5588