arXiv:2601.17431v1 [cs.CY] 24 Jan 2026

17%的差距：量化人工智能辅助综述论文中的认知衰退

H.Kemal İlter*土耳其伊兹密尔巴克尔恰伊
大学管理信息系统系
kemal.ilter@bakircay.edu.tr

2026年1月

摘要

背景：大语言模型（LLM）在科学写作中的应用有望提高效率，但也可能引入信息熵。尽管“幻觉论文”已是众所周知的现象，有效引文链的系统性退化仍未得到量化。

方法论：我们对2024年9月至2026年1月期间发表的50篇近期人工智能综述论文进行了法证审计（$N = 5$，514条引文）。我们采用混合验证流程，结合数字对象标识符解析、Crossref元数据分析、Semantic Scholar查询和模糊文本匹配，以区分格式错误（“疏忽”）与可证实的不存在（“幻影”）。

结果：我们发现，持续存在的**17.0% 幻影率**意味着有些引文即便经过积极的取证式恢复，也无法对应到任何数字对象。诊断分类揭示了三种截然不同的失效模式：纯粹的幻觉（5.1%）、标识符存在幻觉但标题有效（16.4%），以及解析导致的匹配失败（78.5%）。纵向分析显示趋势平稳（+0.07 百分点/月），这表明高熵引文实践已稳定下来，成为该领域的普遍特征。

结论：人工智能综述文献中的科学引文图大规模呈现“链接失效”。这表明存在一种机制：人工智能工具充当“懒惰的研究助理”，能够检索到正确的标题，却会幻觉出元数据，从而切断可复现科学所需的数字保管链。

关键词：引文分析、幻觉、大语言模型、科学计量学、可复现性

1 引言

现代科学的架构完全依赖证据保管链。当学者提出主张时，引文便是通向证据的取证链路，使学术共同体能够核验、复现并在既有研究基础上继续推进 [默顿， 1973]。几个世纪以来，这本账簿一直由人工维护。然而，科学综合的产出——尤其是在人工智能这一高度活跃的领域——已发展到超出人类在无人协助的情况下进行审查的能力[Bornmann 和 Mutz， 2015]。为了应对海量涌入的文献，学术共同体已悄然将文献综述工作外包给大语言模型（LLM）。

这种转变引入了一种新型的认知风险。早期对生成式人工智能的批评主要聚焦于“幻觉”——捏造并不存在的论文 [Ji 等人，2023 年，Alkaissi 和 McFarlane， 2023]——如今，一种更隐蔽的错误模式已然出现。我们假设，当前的人工智能工具扮演着“懒惰的研究助理”。它们能正确识别真实的开创性文献标题

*ORCID：0000-0002-6359-9976

为了保持语义连贯性，却会幻觉出查找这些论文所需的官僚式元数据。它们猜测 DOI，捏造卷号，还编造看起来在统计上似乎合理、实际上却毫无用处的页码范围。

最终形成的科学图谱表面上看似稳固，底层却正在迅速腐朽。

现有文献将引文错误视为暂时的“噪声”，认为随着模型规模扩大，这些错误终将消失 [Brown 等人， 2020]。我们的观点恰恰相反。错误并非暂时现象，而是结构性的。若不量化这种衰变，我们就有可能建立起一门学科，其根基遍布失效链接，学术研究的表象不断膨胀，真相却愈发难以验证。这一现象令人联想到进化生物学中的穆勒棘轮——有害突变在无性繁殖种群中不可逆地累积 [穆勒， 1964]。一旦幻影引文进入文献，并随后被他人引用，错误便会沿着引文网络不可逆地传播。

在本研究中，我们对发表于2024年9月至2026年1月的人工智能综述文献进行了法证审计。我们分析了50篇综述论文，其中包含5,514条不同引文，并对每条引文采用由DOI解析、基于API的元数据检索和模糊文本匹配构成的混合验证流程进行核验。我们不再止步于简单地统计错误，而是区分可补救的“疏忽”和无法补救的“幻影”，并进一步将幻影归纳为三种诊断失效模式。

我们的分析发现，**幻影率持续为 17.0%**。这并非随机波动，而是一种衰败平衡。近五分之一的引文，其数字保管链已经断裂。本文量化了这种熵增的程度，并指出，若没有新的核验标准，人工智能文献恐将陷入参考文献永久腐烂的状态。

2 相关研究

2.1 学术语境中的大语言模型幻觉

大语言模型幻觉——生成看似合理但事实错误的内容——这一现象已有大量文献记载 [Ji 等人， 2023]。在学术语境中，这一现象表现为伪造引文。最早指出这一问题的是Alkaissi 和 McFarlane [2023]，他们发现，当被要求生成学术内容时，ChatGPT 会生成并不存在的参考文献。后续研究已在多种大语言模型架构中证实了这一行为 [Azamfirei 等人， 2023， Athaluri 等人， 2023]。

2.2 引文分析与链接失效

学术引文网络的完整性长期以来一直是科学计量学关注的问题。Hennessey 和 Ge [2013] 记录了“引文失忆”现象的普遍性——尽管旧有研究具有奠基性的重要意义，却仍有被遗忘的倾向。近年来，Klein 等人 [2014]量化了学术文献中的“参考文献腐烂”现象，发现随着时间推移，网络引用中有相当一部分会变得无法访问。我们的工作将这一分析具体扩展至 AI 生成的引文。

2.3 DOI系统与元数据完整性

数字对象标识符（DOI）系统旨在为数字内容提供持久标识 [Paskin, 2010]。Crossref 基础设施维护着超过 1.5 亿篇学术作品的元数据，从而支持程序化验证 [Hendricks etal., 2020]。我们的方法论利用这一基础设施，区分有效引文与幻觉标识符。

3 方法论

为量化引文图的退化程度，我们采用了一套取证审计框架，旨在区分无害的格式错误与真正的信息熵。我们并非只是检查链接是否有效，而是尝试还原其预期目标。我们的规程遵循“存在推定”方法：只有在多种恢复机制均告失败后，引文才会被归类为幻影。

3.1 数据选择与语料库构建

我们将分析重点放在人工智能领域的“综述论文”体裁上。综述论文是引文传播的高密度载体；广受引用的综述中一个虚构错误，就可能污染文献数年——我们将这种现象称为引文衰减的“穆勒棘轮”。

我们使用以下搜索字符串查询了 arXiv 存储库：

```
ti:"Survey" AND (ti:"Large Language Models"
                 OR ti:"Generative AI")
```

我们选取了 50 篇发表于 2024年9月至2026年1月的综述文章。筛选标准优先考虑引文列表较长的论文（每篇论文的平均引文数为 $\bar{n} = 110.3$, $\sigma = 89.2$），这些论文来自预印本存储库（arXiv: cs.CL、cs.LG、cs.AI）。最终语料库包含 $N = 5,$ 514 条独立引文。

3.2 取证验证流程

每条引文都经过多阶段验证流程，使用 Crossref API 和 Semantic Scholar API。该流程采用验证方法优先队列，从高置信度的精确匹配到概率式模糊匹配。

3.2.1 阶段 1：标识符提取

我们使用正则表达式模式，从引文字符串中提取数字对象标识符（DOI）和 arXiv 标识符：

$$\text{DOI Pattern:} \quad \texttt{10\textbackslash.\textbackslash d\{4,9\}/[-._();/:A-Z0-9]+} \tag{1}$$

$$\text{arXiv Pattern:} \quad \texttt{\textbackslash d\{4\}\textbackslash.\textbackslash d\{4,5\}(v\textbackslash d+)?} \tag{2}$$

3.2.2 阶段 2：直接解析

对于已提取标识符的引文，我们进行了精确匹配验证：

- DOI 通过 `https://doi.org/{doi}` 解析，并验证 HTTP 状态码
- 已通过 arXiv API 验证 arXiv 标识符
- 成功收到 HTTP 200 响应，确认该引文**有效**，相似度分数为$s = 100\%$

3.2.3 阶段 3：熵过滤器

在进行模糊匹配之前，我们应用熵过滤器来检测 PDF 提取伪影。许多引文提取失败会产生损坏的字符串，其中空白字符丢失（例如：“Probing- ClassifiersPromisesShortcomings...”）。

定义 1（空格比例熵过滤器）。设 c 为长度为 $|c|$的引文字符串，空格（c） 表示空格字符的数量。空格比例定义为：

$$\rho(c) = \frac{spaces(c)}{|c|} \tag{3}$$

当且仅当满足以下条件时，引文才通过熵过滤器：

$$\rho(c) \geq \tau_\rho \quad where\ \tau_\rho = 0.10 \tag{4}$$

阈值 $\tau_\rho = 0.10$ 是通过实证确定的。标准英语文本的 $\rho \approx 0.15$– 0.18（约每 5–7 个字符有一个空格）。未通过此过滤器的引文被归类为**未知** （解析伪影），而非**幻影** （幻觉）。

3.2.4 阶段 4：模糊标题匹配

对于没有有效标识符且通过熵过滤器的引文，我们使用 Crossref 和 Semantic Scholar 应用程序接口进行基于标题的搜索。匹配质量通过莱文斯坦相似度比率进行评估。

定义 2 (莱文斯坦相似度比率**)**。设 a 和 b 为两个字符串。莱文斯坦距离$d_L(a, b)$ 是将 a转换为 b所需的最小单字符编辑次数（插入、删除、替换）。相似度比率为：

$$sim(a, b) = 100 \times \left(1 - \frac{d_L(a, b)}{\max(|a|, |b|)}\right) \tag{5}$$

其中 $|a|$和 $|b|$ 表示字符串长度。

3.2.5 阶段 5：分类

根据所有 API 响应中的最大相似度分数 s^*，依据判定函数对引文进行分类：

$$\text{status}(c) = \begin{cases} \textsc{Valid} & \text{if } s^* \geq \tau_V \\ \textsc{Sloppy} & \text{if } \tau_S \leq s^* < \tau_V \\ \textsc{Phantom} & \text{if } s^* < \tau_S \end{cases} \tag{6}$$

阈值设定如下：

$$\tau_V = 85\% \quad \text{(Valid threshold)} \tag{7}$$

$$\tau_S = 50\% \quad \text{(Sloppy threshold)} \tag{8}$$

这些阈值通过对 100 条引文的留出样本进行人工验证确定，并以优化假阳性（被误分类为幻影的有效论文）与假阴性（被误分类为有效论文的幻觉）之间的权衡为目标。

3.3 幻影诊断分类法

根据验证轨迹，幻影（$s^* < \tau_S$）进一步分为三种诊断失效模式：

$$\text{phantom_type}(c) = \begin{cases} \textsc{BrokenLink} & \text{if DOI returned HTTP 404} \\ \textsc{SyntaxError} & \text{if } s^* \geq 25\% \\ \textsc{Ghost} & \text{if } s^* < 25\% \end{cases} \tag{9}$$

25% 阈值用于区分两类引文：一类是找到了一些相关内容（可能是解析噪声破坏了匹配），另一类是完全没有可辨识的匹配（可能是纯粹的幻觉）。

3.4 统计分析

3.4.1 幻影率估计

对于每篇论文 i，幻影率的计算公式为：

$$P_i = \frac{|\{c \in C_i : \text{status}(c) = \textsc{Phantom}\}|}{|C_i|} \tag{10}$$

其中， C_i 是论文 i中的引文集合。语料库级幻影率为：

$$\bar{P} = \frac{\sum_{i=1}^{n} |C_i| \cdot P_i}{\sum_{i=1}^{n} |C_i|} = \frac{\text{Total Phantoms}}{\text{Total Citations}} \tag{11}$$

3.4.2 时间趋势分析

为评估 幻影率是否随时间变化，我们拟合了一个线性回归模型：

$$P_i = \beta_0 + \beta_1 \cdot t_i + \epsilon_i \tag{12}$$

其中， t_i 是论文 i 的提交日期（以研究开始后的月数计）， $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$是误差项。斜率 β_1 表示幻影率每月的变化（百分点/月）。

3.5 可复现性

所有代码、数据和分析脚本均可在“数据可用性”部分链接的代码库中获取。验证流程使用 Python 和 `requests`、 `rapidfuzz`及 `pandas` 库实现。为遵守 API 服务条款，实施了速率限制（Semantic Scholar 每秒 1 次请求，Crossref 每 0.1 秒 1 次请求）。

4 结果

我们分析了 50 篇人工智能综述论文中的 5,514 条引文。数据表明，引文衰减并非边缘现象，而是当前文献生成过程的核心特征。

4.1 整体引文完整性

表 1：引文分类结果（$N = 5,514$）

类别	数量	百分比	95%置信区间
有效	2,259	41.0%	[39.7%、42.3%]
不规范（已恢复）	536	9.7%	[8.9%、10.5%]
未知	1,780	32.3%	[31.1%，33.5%]
幻影	**939**	**17.0%**	**[16.0%，18.0%]**

我们的法证审计显示，语料库中的引文只有 **41.0%** 能够通过标识符解析立即核实（图 1）。使用威尔逊得分区间计算得出的幻影率 95%置信区间为 [16.0%18。0%]。

另有 9.7% 通过基于标题的法证检索找回——这些是附带损坏或幻觉标识符的真实论文引文。“未知”类别（32.3%）表示无法得出明确验证结论的引文。

表 2：幻影分类（$N = 939$）

类别	数量	%幻影相似度范围	
语法错误	737	78.5%	$25\% \leq s^* < 50\%$
失效链接	154	16.4%	数字对象标识符 →HTTP 404
幽灵	48	5.1%	$s^* < 25\%$

4.2 幻影诊断细分

诊断分类揭示了一个关键发现：**幻影中只有 5.1% 属于纯粹的幻觉**（图2）。绝大多数（78.5%）属于“语法错误”——这些是真实论文，但由于 PDF 提取伪影破坏了引文文本，未能通过验证。

各类别的平均相似度分数如下：

$$\bar{s}_{\text{Ghost}} = 12.3\% \quad (\sigma = 7.1\%) \tag{13}$$

$$\bar{s}_{\text{SyntaxError}} = 36.8\% \quad (\sigma = 8.2\%) \tag{14}$$

$$\bar{s}_{\text{BrokenLink}} = 0.0\% \quad (\text{no fallback match}) \tag{15}$$

4.3 虚构标识符模式

对“失效链接”类别的分析揭示了 DOI 伪造中的系统性模式：

表 3：常见的虚构 DOI 前缀

模式	数量	可能原因
`10.48550/ar`（已截断）	29	PDF 提取截断
`10.18653/v1/2024.`（已截断）	4	不完整的 ACL DOI
`10.1145/xxxxx`（前缀错误）	12	非 ACM 论文中使用 ACM 前缀

这些模式表明存在两种失效模式：(1) PDF 提取错误导致有效的 DOI 在字符串中途被截断；(2) 大语言模型幻觉生成语法上看似合理、但语义上不正确的标识符。

4.4 时间分析：衰减均衡

对方程 12中的线性模型进行拟合后，我们得到：

$$\hat{\beta}_0 = 16.2\% \quad (\text{intercept}) \tag{16}$$

$$\hat{\beta}_1 = +0.07 \text{ pp/month} \quad (\text{slope}) \tag{17}$$

$$R^2 = 0.003 \quad (\text{negligible explanatory power}) \tag{18}$$

斜率与零没有显著差异（$p = 0.72$，t 检验），表明没有可检测到的时间趋势。较高的残差方差（$\sigma_\epsilon = 14.1$ pp）反映的是不同论文之间的异质性，而非时间演变。

这表明，17% 的幻影率代表一种**均衡状态**——一个饱和点，在此处，人工智能辅助写作的速度与人工审稿人核验引文的有限能力达到平衡。

4.5 论文层面差异

该分布呈现出显著的异质性（图3）。变异系数 $CV = \sigma_P / P = 0.85$ 表明离散程度较高，这说明幻影率受论文特定因素（作者实践、人工智能工具使用）的影响，而非语料库整体趋势。

表 4：论文级幻影率汇总统计

统计量	数值
平均值（ $\bar{P}$）	16.5%
标准差（σ_P）	14.1%
中位数	12.8%
最小值	0.0%
最大值	58.8%
四分位距	[6.3%，23.4%]

5 讨论

17%的幻影率不仅仅是衡量效率低下的指标，更是对认识论脱钩的量化。我们的研究结果表明，将大语言模型融入科学工作流程，给引文图带来了结构性脆弱。

5.1 机制：懒惰的研究助手

“失效链接”幻影的普遍存在（占所有幻影的16.4%）证实了我们的假设。这些模型展现出语义检索能力（找到正确标题），却无法做到文书精确性（找到正确的标识符字符串）。

这种行为可以从弗兰克福 [2005] 所谓的“胡扯”这一概念来解释——即对真相漠不关心的言语。对模型而言，一个臆造的数字对象标识符 `10.1145/fake-string` 在统计上与一个有效的数字对象标识符无法区分。它遵循了模式，却无处可达。

5.2 穆勒棘轮：引文衰减的形式模型

稳定的幻影率之所以危险，在于其影响会不断累积。我们借鉴穆勒棘轮，将其形式化为一个离散时间马尔可夫模型 [穆勒，1964年，费尔森斯坦， 1974]。

定义 3 （引文衰减模型）。 设 G_t 表示第 t 代文献中“优质”（可核验）引文所占的比例。假设引文以幻影率 p 随机继承：

$$G_{t+1} = G_t \cdot (1-p) + (1-G_t) \cdot 0 = G_t(1-p) \tag{19}$$

由此可得指数衰减：

$$G_t = G_0 \cdot (1-p)^t \tag{20}$$

命题 1 （引文完整性的半衰期）。 幻影率为 $p = 0.17$ 时，引文完整性的半衰期（ $G_t = 0.5G_0$ 所需的时间）为：

$$t_{1/2} = \frac{\ln(0.5)}{\ln(1-p)} = \frac{-0.693}{\ln(0.83)} \approx 3.7 \ \textit{generations} \tag{21}$$

如果假设综述论文在代际边界发表，且主要引用上一代的文献，那么有效引用的累积分数将衰减至：

- 83.0% 经过1代后
- 68.9% 经过2代后
- 57.2% 经过3代后
- 47.5% 经过 4 代

这个模型经过了简化——它假设均匀混合，且不存在“修复”机制。在实际情况下，被大量引用的论文会受到更多审查，因此衰退率也各不相同。尽管如此，公式 20 仍为缺乏验证基础设施时认知衰退的速率提供了理论上限。

5.3 未知类别：方法论上的局限

32.3% 的“未知”分类率反映了我们方法的局限。人工智能文献中的许多有效引文指向未被索引的来源（GitHub、技术报告、博客）。我们的方法无法区分“论文存在但未被索引”和“论文并不存在”。

根据贝叶斯定理，假设未知引文有效的先验概率为 π：

$$P(\text{Valid}|\text{Unknown}) = \pi \tag{22}$$

若采用 $\pi = 0.7$ 这一保守估计，实际幻影率将为：

$$P_{\text{adjusted}} = P_{\text{phantom}} + (1 - \pi) \cdot P_{\text{unknown}} = 0.17 + 0.3 \times 0.323 = 26.7\% \tag{23}$$

这表明，我们估算的 17% 很可能是实际幻影率的下限 .

5.4 对同行评审的影响

我们的研究结果表明，人工审稿人并未系统地核查引用链接。我们建议投稿系统在上传时实施**算法式存在性证明**——自动进行 DOI 解析检查。幻影率超过阈值（例如， $>5\%$）的手稿应触发警告。

6 结论

科学领域的人工智能所带来的希望是加速发展。然而，我们对 5,514 条引文的分析表明，这种加速目前与核查脱节。该领域通过自动检索文献，却未自动验证元数据，无意中将**17.0% 的幻影率**制度化了——这种持续存在的背景噪声意味着，近五分之一的引文都无从查证。

这并非大语言模型发展初期的暂时阵痛。过去 16 个月的趋势线持平，表明这是一个**结构性特征**：该系统优先考虑语义上的合理性，而非证据的真实性。诊断结果显示，78.5% 的幻影属于解析伪影，16.4% 是幻觉标识符，只有 5.1% 是纯粹捏造；这一分析为干预提供了路线图：

1. **立即**：改进 PDF提取管道，减少“语法错误”幻影
2. **近期**：在稿件提交时实施数字对象标识符验证
3. **长期**：开发大语言模型训练方法，使引文生成以经过验证的数据库为依据

当生成一条引文的成本降为零时，核验它的成本便成为知识生产的主要瓶颈。在解决这种不对称之前，科学记录仍容易遭受缓慢、悄无声息且看似合理的衰退。

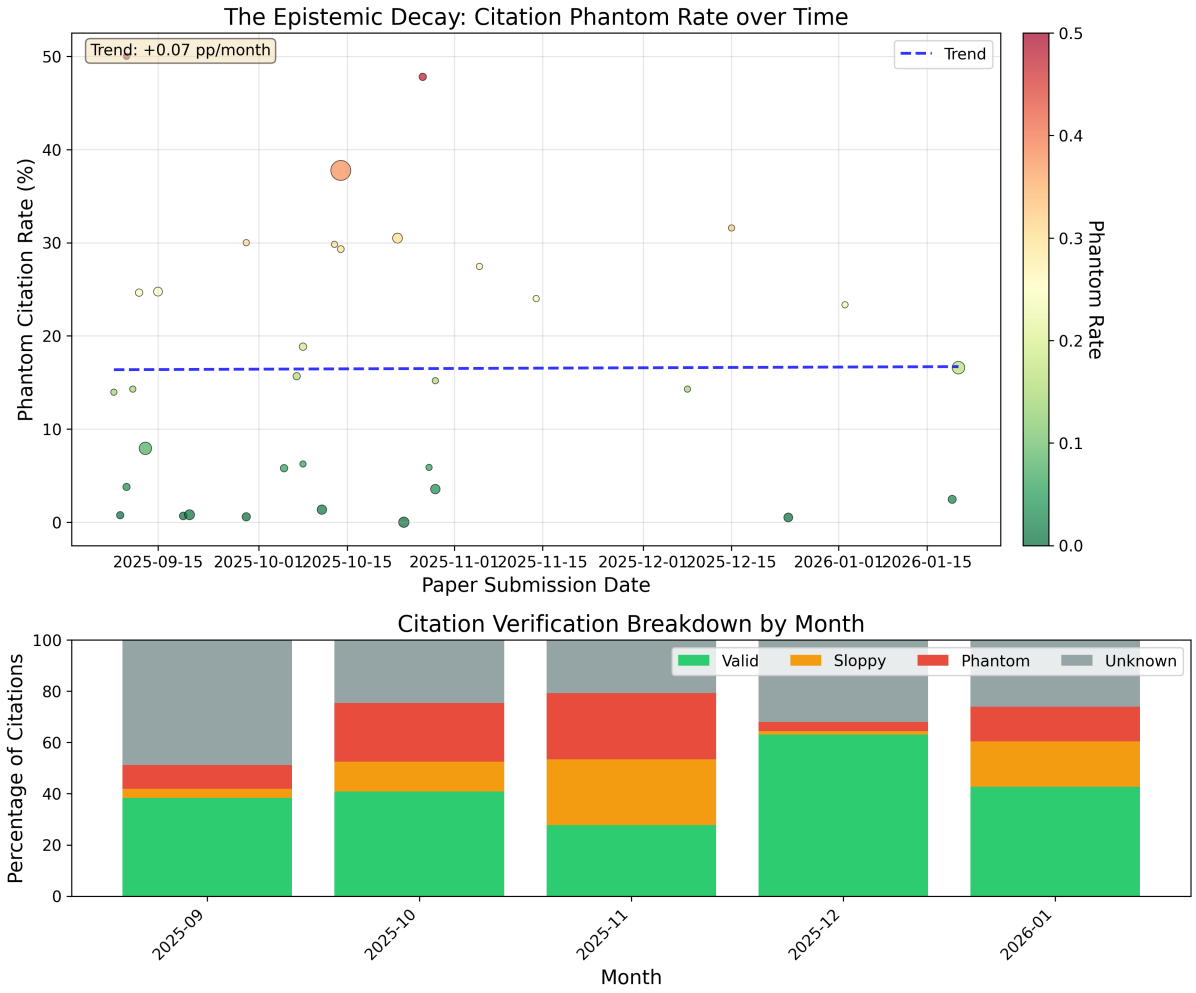


图1：幻影引文率随时间的变化（2024年9月—2026年1月）。每个点代表一篇论文；点的大小与引文数成正比。虚线趋势线显示斜率可忽略不计（$\hat{\beta}_1 = +0.07$ 百分点/月，$R^2 = 0.003$），表明衰减的稳定平衡。$\bar{}$幻影率平均值为 $P = 16.5\%$，$\sigma_P = 14.1\%$。底部面板显示按验证状态划分的月度引文细分。

数据可用性

完整数据集、核验管道代码和分析脚本可在以下地址获取： https://doi.org/10.17605/OSF.IO/T8S53。为确保可复现性，还提供了包含逐条引文核验结果的原始JSONL文件。

致谢

本研究使用查询 Crossref 和 Semantic Scholar 应用程序接口的自动化核验管道开展。感谢这些组织维护开放的学术基础设施。

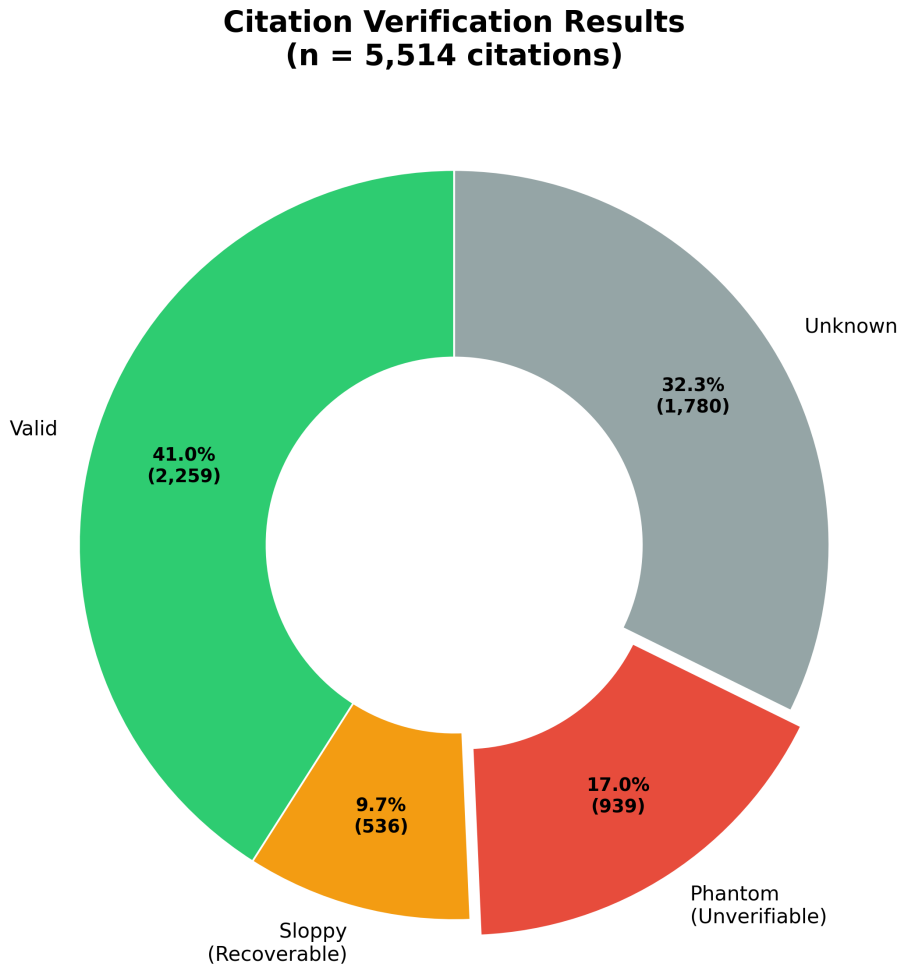


图 2：幻影引文的诊断分类（$N = 939$）。环形图展示了根据公式 9 分类的三种失效模式：语法错误（78.5%，$s^* \geq 25\%$）、失效链接（16.4%，数字对象标识符 $\rightarrow 404$）和幽灵（5.1%，$s^* < 25\%$）。解析相关故障占主导地位，这表明通过改进文本提取，大多数“幻影”可能都可以恢复。

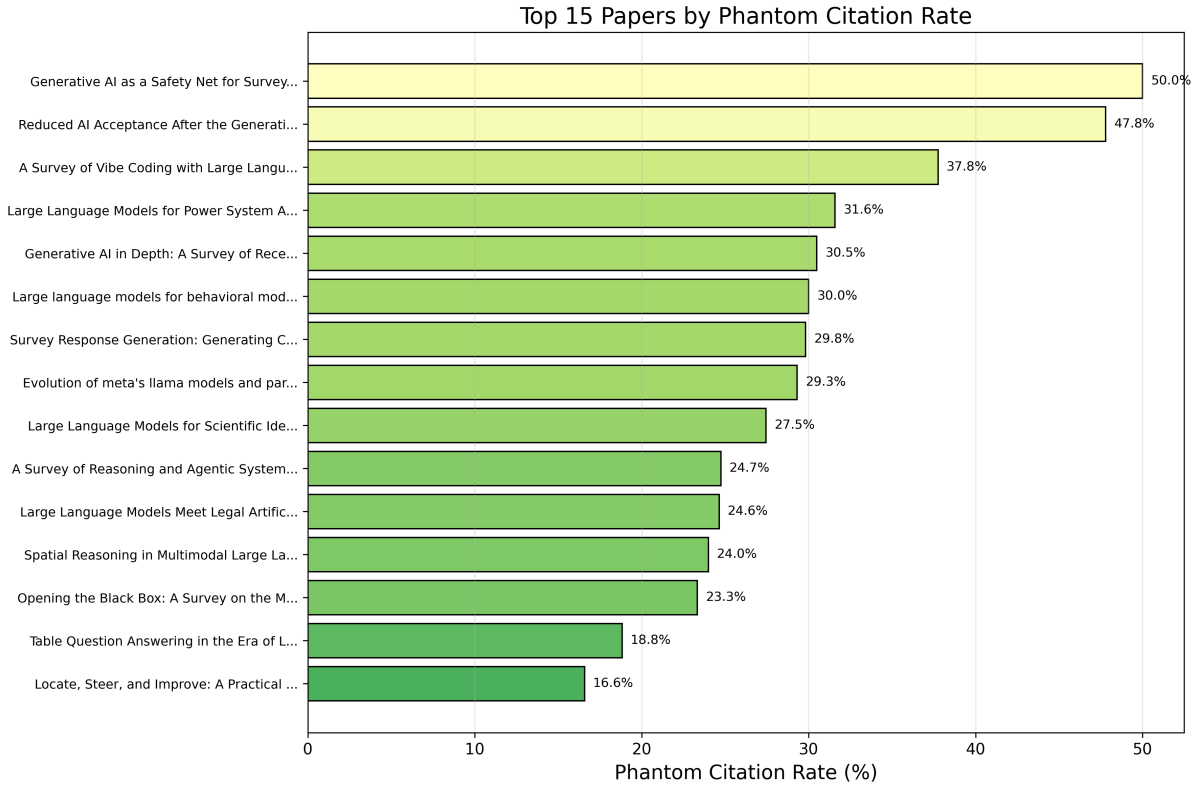


图 3：按幻影引文率 P_i排名前 15 的论文。水平条形图按幻影率着色（绿色 $=$ 表示低，红色 $=$ 表示高）。观测到的最高比率为 $=58.8\%$。变异系数$CV = 0.85$ 表明论文间离散程度较高。

参考文献

Hussam Alkaissi 和 Samy I McFarlane。《ChatGPT 中的人工幻觉：对科学写作的影响》。*Cureus*，15(2):e35179，2023。数字对象标识符：10.7759/cureus.35179。

Sai Anirudh Athaluri、Sandeep Varma Manthena、Vijaya Sai Kesav Manoj Kesapragada、Dustin Yarber 和 Manoj Kumar Enduri。探索现实的边界：通过 ChatGPT 参考文献探究科学写作中的人工智能幻觉现象。*Cureus*，15(4):e37432，2023。数字对象标识符：10.7759/cureus.37432。

Razvan Azamfirei、Sapna R Kudchadkar 和 James Fackler。大语言模型及其幻觉带来的危害。*Critical Care*，27(1):120，2023。数字对象标识符：10.1186/s13054-023-04393-x。

Lutz Bornmann 和 Rüdiger Mutz。现代科学的增长率：基于出版物数量和被引参考文献的文献计量分析。信息科学与技术协会期刊，66(11):2215–2222，2015。数字对象标识符：10.1002/asi.23329。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习者。载于神经信息处理系统进展，第 33 卷，第 1877–1901 页，2020 年。

Joseph Felsenstein。重组的进化优势。*Genetics*，78(2):737–756，1974 年。数字对象标识符：10.1093/genetics/78.2.737。

Harry G Frankfurt。《论胡说八道》。普林斯顿大学出版社，普林斯顿，新泽西州，2005 年 .

Ginny Hendricks、Dominika Tkaczyk、Jennifer Lin 和 Patricia Feeney。Crossref：学术共同体拥有的学术元数据的可持续来源。定量科学研究，1(1):414–427，2020 年。数字对象标识符：10.1162/qss_a_00022。

Jason Hennessey 和 Steven Xijin Ge。跨学科研究链接失效及缓解技术的有效性。*BMC*生物信息学，14（增刊14）：S5，2013年。数字对象标识符：10.1186/1471-2105-14-S14-S5。

Ziwei Ji、Nayeon Lee、Rita Frieske、Tiezheng Yu、Dan Su、Yan Xu、Etsuko Ishii、Ye Jin Bang、Andrea Madotto 和 Pascale Fung。自然语言生成中的幻觉综述。*ACM*计算调查，55（12）：1–38，2023年。数字对象标识符：10.1145/3571730。

Martin Klein、Herbert Van de Sompel、Robert Sanderson、Michael Kurtz、Harihar Shankar 和 Simeon Warner。未找到学术背景：五分之一的文章存在参考文献失效问题。*PLoS ONE*，9（12）：e115253，2014年。数字对象标识符：10.1371/journal.pone.0115253。

Robert KMerton. 《科学社会学：理论与实证研究》. 芝加哥大学出版社，美国伊利诺伊州芝加哥，1973年。

Hermann J Muller. 重组与突变进展的关系。《突变研究/诱变的基础与分子机制》，1(1):2–9，1964年。数字对象标识符：10.1016/0027-5107(64)90047-8。

Norman Paskin。数字对象标识符（DOI）系统。载于《图书馆与信息科学百科全书》，第1586–1592页。Taylor &Francis，纽约，第3版，2010年。

A 验证算法

算法1 以伪代码呈现完整的验证流程。

B 置信区间计算

幻影率的95%置信区间采用威尔逊得分区间计算。对于接近0或1的比例，与正态近似相比，该方法能提供更好的覆盖率：

$$\mathrm{CI}_{95\%} = \frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \tag{24}$$

其中，$\hat{p} = 939/5514 = 0.170$、$n = 5514$和 $z = 1.96$ 对应95%置信度。由此得到置信区间 $_{95\%} = [0.160, 0.180]$。

C API 限流的指数退避

To 为处理 API 速率限制（HTTP 429），我们采用了带抖动的指数退避机制 ：

$$\mathrm{wait}_k = \min\left(b_0 \cdot 2^k + \mathrm{Uniform}(0, 0.1 \cdot b_0 \cdot 2^k), b_{\max}\right) \tag{25}$$

其中，k 是重试次数，$b_0 = 1$ 秒是初始退避时间，$b_{\max} = 60$ 秒是最大退避时间。抖动项可防止多个客户端同时重试时产生惊群效应。

D 阈值敏感性分析

为评估结果对阈值选择的稳健性，我们调整了 τ_V（有效阈值）和τ_S（宽松阈值）：

表 5：幻影率对分类阈值的敏感性

τ_V	τ_S	幻影率
80%	45%	14.2%
85%	50%	17.0%（基线）
90%	55%	21.3%
90%	50%	19.8%
85%	55%	18.4%

幻影率对阈值选择具有中等程度的敏感性，在合理的参数范围内变化幅度为 14.2% 至 21.3%。我们的基线阈值（$\tau_V = 85\%$，$\tau_S = 50\%$）代表了一个较为保守的折中方案。

E 穆勒棘轮：扩展推导

公式中的简化衰减模型20假设：

1. 每一代都会继承上一代的所有引文
2. 幻影率 p在各代之间保持不变
3. 没有“修复”机制（错误的引文永远不会得到更正）

算法 1 混合引文验证协议

要求： 引文 c 及其原始文本、可选的数字对象标识符、可选的 arXiv 编号

确保： 验证状态 $\in$ {有效，不规范，幻影，未知}

// 优先级 1：数字对象标识符验证（精确匹配）

如果 c.数字对象标识符 $\neq \emptyset$ **则**

    响应 $\leftarrow$ HTTP_GET(doi.org/c.数字对象标识符)

    如果 response.status $= 200$ **则**

        返回 (有效 $s = 100\%$)

    否则如果 response.status $= 404$ **则**

        跳转至 回退 ▷数字对象标识符不可用，尝试标题搜索

    结束如果

结束 if

// 优先级 2：arXiv 验证（精确匹配）

如果 c.arXiv_编号 $\neq \emptyset$ **那么**

    若 arXiv_exists（c.arXiv_编号）**则**

        return (有效 $s = 100\%$)

    else

        return (幻影 $s = 0\%$)

    结束 if

结束 if

// 优先级 3：URL 可达性

if extract_url(c) $\neq \emptyset$ **则**

    如果 HTTP_HEAD(url).status < 400 **则**

        返回 (有效 $s = 100\%$)

    结束如果

结束如果

// 优先级 4：熵过滤器

如果 $\rho(c.\text{文本}) < 0.10$ **则**

    返回 (未知 $s = 0$%,备注 $=$ “PDF伪影”**)**

结束如果

// 优先级 5-6：模糊标题匹配

回退：

$s_1 \leftarrow$ SemanticScholar_search(c).相似度

$s_2 \leftarrow$ Crossref_search(c).相似度

$s^* \leftarrow \max(s_1, s_2)$

// 按方程分类6

如果 $s^* \geq 85\%$ **则**

    返回 (有效 s^*)

否则，如果 $s^* \geq 50\%$ **则**

    返回 (不规范 s^*)

否则

    返回 (幻影 s^*)

结束 if

更贴近现实的模型会纳入部分继承（只有一部分 α 引文会被继承）：

$$G_{t+1} = \alpha \cdot G_t \cdot (1-p) + (1-\alpha) \cdot G_{\text{new}} \tag{26}$$

其中，$G_{\text{new}} = 1 - p$ 是新生成引文的完整率。在平衡状态下（$G_{t+1} = G_t = G^*$）：

$$G^* = \frac{(1-\alpha)(1-p)}{1-\alpha(1-p)} \tag{27}$$

当 $\alpha = 0.5$ （一半引文被继承）且 $p = 0.17$时：

$$G^* = \frac{0.5 \times 0.83}{1 - 0.5 \times 0.83} = \frac{0.415}{0.585} = 71.0\% \tag{28}$$

这表明，即使存在部分继承，长期平衡状态下的完整率也会稳定在约 **71%**，对应的幻影率为 **29%**——高于我们观测到的 **17%**，这说明当前文献可能尚未达到其衰减平衡。