部署约束会让大语言模型产生引用幻觉吗？跨四种模型和五种提示方案的实证研究

Chen Zhao
cz1296@nyu.edu 纽约大学 美国

Yuan Tang
yuantang@alumni.cmu.edu 卡内基梅隆大学 美国

Yitian Qian
qyt024@bu.edu
波士顿大学 美国

arXiv:2603.07287v1 [cs.IR] 7 Mar 2026

摘要

大语言模型日益广泛地用于撰写学术文本和支持软件工程（SE）证据综合，但它们经常会生成看似可信的书目参考文献，却存在引用幻觉。我们研究部署驱动的提示约束如何影响闭卷设置下引用的可验证性。我们使用 144 条论断（其中 24 条涉及软件工程与计算机科学），并采用确定性验证流程（Crossref + Semantic Scholar），评估两种专有模型（Claude Sonnet、GPT-4o）和两种开放权重模型（LLaMA 3.1–8B、Qwen 2.5–14B）在五种方案下的表现：基线、时间约束（发表年份窗口）、综述式广度、不披露政策，以及上述条件的组合。在生成的 17,443 条引文中，没有任何模型的引用级存在率超过 0.475；时间约束和组合条件导致的降幅最大，而输出仍符合格式要求（书目信息字段格式完整）。未解决结果占主导（36–61%）；对 100 条引文进行审计后发现，相当一部分未解决案例是伪造的。研究结果表明，在将大语言模型的输出用于软件工程文献综述或工具流程之前，应进行事后引用验证。

CCS 概念

• **通用与参考 → 实证研究**；• **计算方法 → 自然语言处理**。

关键词

大语言模型幻觉、引用幻觉、大语言模型、实证评估、面向软件工程的人工智能、系统性文献综述、自动化验证、大语言模型可靠性

1 引言

大语言模型（LLM）已展现出生成流畅学术文稿的强大能力，因此越来越多地应用于学术写作和证据综合。然而，它们仍会出现引用幻觉 [2, 19]：生成看似完整（包含作者、发表场所、数字对象标识符），却无法对应到可验证文献的书目参考文献。这些错误通常被归因于下一词元预测目标和分布偏移 [13]。然而，在真实的部署约束下——包括时间限制、综述式广度压力，以及禁止声称能够访问训练数据的不披露政策——引用可靠性的表现仍未得到充分理解。

这一问题与软件工程（SE）研究尤其相关，因为系统性文献综述（SLR）、映射研究和相关工作章节是核心方法论活动 [10]。随着大语言模型日益融入软件工程工作流程——从起草技术报告和架构决策记录到协助开展文献综述 [5, 7]——不可靠的引文可能会在软件工程领域依赖的证据综合流程中不断传播。因此，了解部署约束如何影响引文质量，是在软件工程实践中负责任地采用大语言模型辅助写作工具的前提。

我们通过一个自动化验证框架弥补这一空白：该框架解析每条引文，查询 Crossref 和 Semantic Scholar，并将其归为三个标签之一：已存在、未解决或伪造。我们在五种提示条件下评估了四种大语言模型，采用确定性解码，发现：即使模型仍能保持格式合规性，时间约束也与可验证性的最大幅度下降相关；综述式提示会扩大专有模型与开放权重模型之间的可验证性差距；未解决结果占引文的 36–61%，并在多数模型-条件单元中构成最大的类别。

我们的研究围绕三个研究问题展开：**RQ1** 部署约束（时间约束、综述式提示、不披露信息）如何影响引用可验证性?

RQ2 这些影响在专有模型与开放权重模型之间有何差异?

RQ3 将全部三项约束结合起来，与分别应用每项约束相比，效果如何?

我们的贡献包括：**(1)** 一个经过整理的数据集，包含涵盖六个学术领域的 144 条主张（其中 24 条属于软件工程与计算机科学）；**(2)** 一条配有经审计的三分类标签体系的自动化验证流水线（Cohen's κ=0.63 与人工标签对比）；**(3)** 对部署约束以及专有模型与开放权重模型之间的差异如何影响引用可验证性的实证分析。

2 相关工作

语言模型中的幻觉。大语言模型幻觉已在摘要生成和知识密集型生成领域得到广泛研究 [3, 8, 12, 18]。McKenna 等人 [13] 将许多失败归因于训练与部署之间的分布差异，而 Kalai 等人 [9]则认为，下一词元预测可能会促使模型生成看似合理的补全内容，而非表达校准不确定性。我们研究一种便于审查的表现形式：虚构的书目引文，其输出可通过外部学术索引进行核查。

引用幻觉与可验证生成。 先前研究审查了大语言模型生成的参考文献，但往往采用单一提示设置，并以二元标签区分真实与伪造。Walters 和 Wilder[19] 报告称，ChatGPT 生成的引文中有很高比例并不存在；Agrawal 等人 [2] 则表明，幻觉参考文献往往包含内部不一致的元数据。相关研究

则更广泛地评估带引文的生成：ALCE 基准测试端到端的带引文生成 [6], , Liu 等人 [11] 则审查生成式搜索引擎中的引文精确率和召回率。我们的研究在闭卷参考文献列表设置中系统地改变部署约束 （时间约束、综述式广度、不披露和组合），并采用三分类法（已存在、未解决、伪造），以明确区分伪造的参考文献与无法可靠核实的情况。

用于科学写作的大语言模型。采用结构化规划的检索增强流程可以减少文献综述生成中的幻觉 [1], , 但即使明确要求提供依据，幻觉参考文献仍很常见 [17]。这些问题在软件工程研究中常见的证据综合工作流程（例如系统综述） [10] 以及近期工作综述的新兴大语言模型辅助的软件工程工具 [5, 7]中尤为突出。我们的工作与之互补：我们不提出新的生成流水线，而是提供一个确定性验证框架，并用它量化在闭卷设置下，部署驱动的提示约束如何影响引文可靠性。

3 实验设计

任务定义。我们研究大语言模型生成的学术风格写作中的引文可靠性。每个模型都会收到一个问题式提示，并生成一段简洁的学术段落，随后附上包含固定书目信息字段（标题、作者、发表场所、年份、DOI，以及可用时的统一资源定位符）的结构化参考文献列表，以便进行确定性解析和自动化验证。提示要求恰好提供k 条引文（基线/时间约束/不披露为5条；综述/组合为8条），但我们不强制要求符合数量：输出中的所有参考文献都会经过验证，并记录实际数量。验证采用加权元数据相似度分数（公式1）；详情见第4节。

主张数据集。 我们的数据集包含 144 条问题式提示（例如“有哪些证据支持……？ ”），涵盖六个领域组：软件工程与计算机科学（24 条主张，涉及软件工程、人工智能与机器学习、数据与人机交互、安全和系统）、自然科学（24 条）、医学与健康（24 条）、社会科学（24 条）、人文学科（24 条）和跨学科（24 条）。每条记录都包含一个领域标签、一个可选的时间窗口，以及可选的种子锚点（仅含关键词的主题提示，用于引导模型，但不会指定具体论文）。若存在种子锚点，同一条主张在所有条件下使用的锚点均相同。144 条提示从一个包含 240 个条目的候选池中随机抽取（使用固定随机种子）；这些条目来源于公开可获取的学术资料，已预先筛选以确保适用性，并按领域分层。

模型。 我们评估了两个专有模型——Claude Sonnet（Anthropic）和 GPT-4o（OpenAI）——以及两个开放权重模型——Qwen 2.5–14B（阿里巴巴）和 LLaMA 3.1–8B（Meta） [4, 14, 15, 20].[1]由于专有模型可能规模更大，我们的设计无法完全区分模型规模与专有模型和开放权重模型之间的差异（第6节）。

提示条件。 每项论断均通过固定提示模板在五种条件下进行评估。基线 要求撰写一篇学术段落，且恰好包含 5 条引文。时间约束 设定严格的出版年份窗口（5 条引文）；全部 144 个窗口均以 2025 年为终点，中位跨度为五年，旨在针对近期文献，因为参数知识在此方面最薄弱。综述 要求撰写一篇相关工作综合，按 3–4 个方法类别组织，并包含 8 条引文，以体现证据综合写作中对广度的要求。综述/组合要求 8 条引文（其他条件为 5 条），因为 5 条引文的额度会导致每个类别不足两条参考文献，这不符合综述类写作的常见做法；由于我们的主要结果是引文级比例，不同额度改变的是观察数量，而非结果定义。此外，我们还报告按主张验证比例（图2），以呈现主张等权的视角。不披露 指示模型不得声称能够访问其记忆中的训练文档（5 条引文），以模拟商业写作助手中的部署策略。组合 结合了上述三项约束（8 条引文）。所有条件均采用确定性解码（温度为 0），且不使用检索增强。完整设计共进行 $144 \times 5 \times 4 = 2,880$ 次运行，生成 17,443 条独立引文。

4 验证流程

我们的流水线会使用两个学术数据库——Crossref 和 Semantic Scholar——核查每一条生成的引文，它们共同覆盖了大部分已索引文献。该流水线在所有模型和条件下均以相同方式运行。代码已公开提供。[2]

解析。 流水线会从每个模型的输出中提取引文列表，并将每条参考文献拆分为结构化字段：标题、作者、发表场所、年份和 DOI。与任务定义一致，我们会提取输出中出现的所有参考文献（不截断、不填充，也不重试）。我们使用基于规则的启发式和正则表达式；无法解析的参考文献会作为未解决项继续处理。

候选项检索。 对于每条已解析的引文，我们最多执行三次查询：(1) 如果模型提供了 DOI，我们会通过 Crossref 直接核查；(2) 按标题在 Semantic Scholar 中搜索（最多返回 k=5 个候选项）；(3) 按标题在 Crossref 中搜索（同样最多返回 k=5 个候选项）。所有检索到的记录都会转换为统一格式。

评分。 每个候选项都会根据以下加权组合与解析后的引用进行评分：

$$s = 0.60 \cdot t + 0.20 \cdot a + 0.15 \cdot y + 0.05 \cdot v \tag{1}$$

其中， t 为模糊标题相似度（词元集比率）， a 为作者姓氏重叠度， y 为年份一致性（完全一致时为 1.0，相差一年时为 0.5，否则为 0），v 为刊物相似度（部分比率）。标题的权重最高，因为在实际应用中，它是区分度最高的字段。

标注。 每条引文均会获得三个标签之一：已存在 （得分 ≥ 0.85）、未解决 （$0.60 \leq$ 得分 < 0.85）或伪造 （得分 < 0.60 或未找到候选项）。对于时间条件，超出所述发表年份窗口的引文还会被额外标记。我们将 0。85 设为匹配阈值：得分高于此阈值时，即使存在轻微的元数据漂移（例如发表场所缩写、缺少中间名首字母），候选项也始终对应于正确的论文。0。60 这一界限将较弱且很可能为虚假的匹配与部分看似合理的匹配区分开来；低于 0。60 时，候选项通常仅共享通用的

[1]具体的模型标识符列于复现包中。

[2]请参阅复现包：https://github.com/Zerichen/Citation-Hallucination。

表1：在包含100条引文的分层样本中，流水线标签与人工标签的对比。总体一致性：75%；Cohen's κ=0.63。

	人工标签		
流水线标签	存在。	未解决。	伪造。
已存在	31	1	0
未解决	4	15	16
伪造	2	2	29

标记。未解决 有意作为一个分诊桶：其中包括：（i）由于元数据不完整或不一致，流水线无法完全确认的真实论文；以及（ii）与无关记录存在部分重叠的伪造引文。因此，我们将其单独报告，而不是强行作出二元决策。

聚合与指标。在模型—条件单元格中，对所有引文的已存在、伪造和未解决比例（总和为 1）取平均，并报告 95% 聚类自助法置信区间（对 144 条主张进行 1,000 次重采样），以考虑主张内相关性。对于关键成对比较，我们报告差值(Δ)的自助法置信区间，用于比较存在率 [16]；当置信区间不包含零时，差异具有统计学意义。除了引文加权率，我们还报告逐条主张的核验比例（图2），f_i =#已存在$_i$/#总解析数$_i$，其中每条主张的权重相同，不受引文数量影响。

流水线验证。 我们从 Google Scholar 和 DBLP 中分层抽取了 100 条引文进行人工核验（表1）。总体一致性为 75%，Cohen's κ（流水线与人工核验之间）为 0.63。已存在 引文的精确率为 0.97，伪造引文的精确率为 0.88，但未解决引文的精确率仅为 0.43。占主导地位的错误是将未解决 → 伪造 （16 例），因此我们报告的捏造率可能偏保守；据此，我们将未解决 视为高风险，并在第 5 节报告重新分类敏感性分析。

5 结果

表2报告了引文加权验证指标；图1以可视化方式呈现结果分布；图2总结了按主张验证比例（每项主张权重相等，不受引文数量影响）。图3按领域细分了存在率；表3给出了关键成对比较中存在率差值（Δ）的自助法置信区间。

没有任何模型核验通过其多数引文。 在任何条件下，没有任何模型的存在率超过 0.50（峰值为 Claude Sonnet 在综述条件下的 0.475）。按照最严格解释——仅当每一条 引文均通过核验时，才将某项主张视为已核验——即便是表现最佳的模型，也只有极少数主张符合条件。不同约束会导致不同的失效特征，而非仅仅造成一致性的性能下降。

逐项主张分布揭示了部分核验现象。 图2展示了每项主张的核验比例。即使总体比例较低，部分主张仍会获得部分已核验的引文集合：Claude Sonnet 在基线条件下的逐项主张比例中位数为 0.40（四分位距 0.20–0.60），而对于开放权重模型，在大多数条件下其中位数为零。总体比例较低并不意味着每项主张都完全未经核验，但许多主张根本没有获得任何已验证引文。

专有模型表现更好，但仍不理想（*RQ2*）。在基线条件下，Claude Sonnet 和 GPT-4o 的存在率分别为 0.381 和 0.235，而 LLaMA 3.1–8B 和 Qwen 2.5–14B 的存在率仅为 0.068 和 0.090。专有模型与开放权重模型之间的差距很大且具有明确的统计显著性（$\Delta = +0.229$，95% 置信区间为 $[0.191, 0.266]$；表3），不能由引文数量解释（所有模型每条论断均为 ≈ 5），并且在完整的逐论断分布中得到证实（图2）。这一差异在压力条件下依然存在，而且通常更大，这可能反映了训练数据覆盖范围和模型规模的差异，但我们无法通过当前的研究设计将这些因素分离开来。

时间约束导致最显著的下降（*RQ1*）。 时间约束对引文质量的降低幅度超过任何其他单一条件。Claude Sonnet 从 0.381 降至 0.119（$\Delta = -0.261$，95 4% 置信区间 $[-0.317, -0.207]$）；GPT-o 的降幅相近（$\Delta = -0.216$，置信区间 $[-0.266, -0.168]$）；开放权重模型原本就接近零，进一步下降（Qwen 为 $\Delta = -0.076$，LLaMA 为 -0.057；表3）。关键在于，直接的时间违规极为罕见（0.001–0.026）：模型遵守 时间约束，却无法在该约束范围内生成可验证参考文献；这是仅检查格式层面合规性时完全无法发现的失败模式。

调查式提示与更大的专有模型–开放权重差距相关（*RQ1*、*RQ2*）。在调查条件下，专有模型与开放权重模型之间的差距增至观测到的最大值（$\Delta = +0.310$，95 3%置信区间 $[0.274, 0.349]$；表）。模型相对于所要求的8条引文略有不足（平均引文数7.28–8.00；表2），但该模式仍然存在。Claude Sonnet有所改善——存在率从0.381升至 0.475（$\Delta = +0.094$，置信区间 $[0.028, 0.162]$），捏造率相当（0.157→0.161）——表明其依托了充足的真实参考文献储备。GPT-4o在调查条件下的下降在统计上不显著（$\Delta = -0.032$，置信区间 $[-0.092, 0.029]$）。开放权重模型呈现相反趋势：Qwen 2.5–14B显著下降（$\Delta = -0.070$，置信区间 $[-0.101, -0.041]$），捏造率达到0.547，为本研究中的最高值。

不披露指令重新分配错误，而非消除错误（*RQ1*）。不披露提示的影响更为微妙：存在率仅小幅下降，GPT-4 Δ = o（，CI ）和 −0.060 $[-0.119, -0.001]$ $\Delta = -0.023$ LLaMA （，CI $[-0.045, -0.001]$）出现边际下降，而 Claude Sonnet 和 Qwen 则没有显著变化（表3）。变化之处在于错误的分布：部分已验证引文转入未解决类别（例如，Claude Sonnet 的存在率为 0.381→0.349，未解决率为 0.462→0.487）。在不披露条件下，DOI 完整性下降（LLaMA 的降幅最为明显，为 −11.4 个百分点[3]）；由于 DOI 是最强的验证信号，其被抑制后，会使引文从已存在转入未解决类别——错误也由「明显错误」转变为「难以判断」。

组合约束产生了最差结果（*RQ3*）。组合条件导致的性能下降最为严重（图 1）。四种模型中有三种的存在率接近于零；只有 Claude Sonnet 保留了非零的 0 106 $\Delta = -0.275$ $[-0.329, -0.220]$ 4 $\Delta = -0.230$ （，CI）。GPT-o 的下降程度相近（，CI $[-0.274,)$；两种开放权 −0.185] 3 重模型均彻底崩溃（表）。对于专有模型，组合约束造成的下降幅度超过了单一约束造成的最大降幅；开放权重模型则呈现地板效应。关键在于，

[3]各模型和条件下的 DOI 完整性比率详见复现包。

表 2：引文级验证指标及其 95% 自助法置信区间（每个单元格包含 N=144 条主张）。条件：基线 = 基线，温度约束 =时间约束，生存 =综述，非歧视 =不披露。时间违规 =时间违规率。

模型	条件	N	已存在 ↑	伪造 ↓	未解决	T. 违规。	平均引用数
Claude Sonnet	Base	144	.381 [.335,.426]	.157 [.126,.190]	.462 [.382,.549]	.000	5.00
GPT-4o	Base	144	.235 [.189,.281]	.281 [.239,.327]	.484 [.403,.569]	.000	4.97
LLaMA 3.1–8B	Base	144	.068 [.049,.090]	.369 [.327,.411]	.563 [.479,.646]	.000	4.98
Qwen 2.5–14B	Base	144	.090 [.065,.118]	.442 [.398,.486]	.468 [.382,.549]	.000	4.98
Claude Sonnet	Temp	144	.119 [.089,.151]	.347 [.306,.392]	.533 [.451,.618]	.015	5.00
GPT-4o	Temp	144	.019 [.008,.033]	.451 [.412,.493]	.529 [.444,.611]	.001	5.00
LLaMA 3.1–8B	Temp	144	.011 [.004,.019]	.394 [.351,.440]	.595 [.514,.674]	.026	5.01
Qwen 2.5–14B	Temp	144	.014 [.004,.025]	.455 [.408,.499]	.531 [.444,.611]	.015	4.99
Claude Sonnet	Surv	144	.475 [.425,.523]	.161 [.133,.189]	.364 [.285,.444]	.000	8.00
GPT-4o	Surv	144	.203 [.165,.246]	.302 [.262,.345]	.495 [.410,.576]	.000	7.59
LLaMA 3.1–8B	Surv	144	.038 [.025,.053]	.436 [.392,.475]	.526 [.444,.611]	.000	7.95
Qwen 2.5–14B	Surv	144	.020 [.009,.032]	.547 [.509,.584]	.433 [.347,.514]	.000	7.28
Claude Sonnet	非歧视	144	.349 [.302,.397]	.165 [.129,.201]	.487 [.403,.569]	.000	4.90
GPT-4o	非歧视	144	.175 [.142,.210]	.317 [.271,.360]	.508 [.424,.590]	.000	5.00
LLaMA 3.1–8B	非歧视	144	.045 [.029,.066]	.398 [.357,.435]	.557 [.472,.639]	.000	4.97
Qwen 2.5–14B	非歧视	144	.078 [.054,.104]	.410 [.368,.453]	.512 [.431,.597]	.000	4.99
Claude Sonnet	组合	144	.106 [.078,.142]	.359 [.318,.399]	.536 [.451,.618]	.027	7.61
GPT-4o	组合	144	.005 [.000,.012]	.452 [.418,.489]	.543 [.465,.625]	.000	7.38
LLaMA 3.1–8B	组合	144	.008 [.003,.014]	.386 [.343,.423]	.606 [.521,.681]	.000	7.99
Qwen 2.5–14B	组合	144	.001 [.000,.003]	.507 [.468,.543]	.492 [.410,.576]	.000	7.56

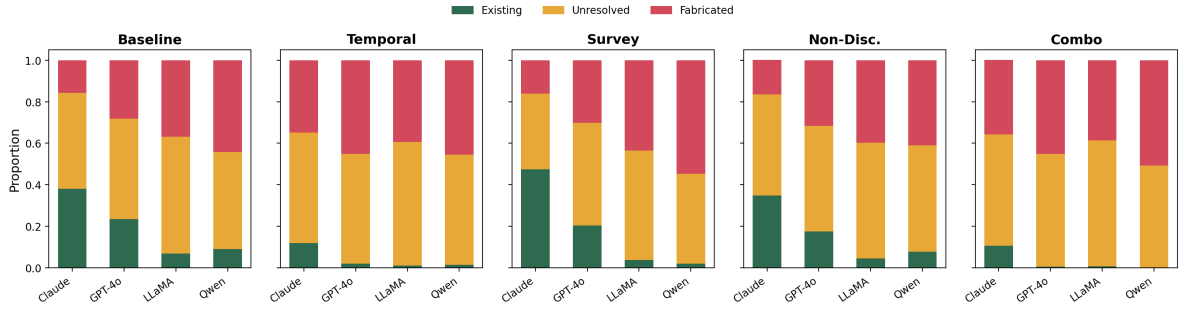

图1：所有五种条件下各模型的引用级结果分布（已存在、未解决、伪造），以堆叠比例图展示，比例总和为一。“非歧视”表示不披露条件。

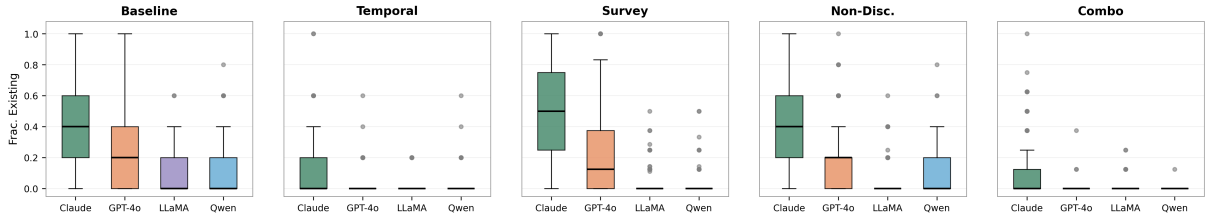

图2：按模型和条件划分的每项主张的核验比例（箱体显示四分位距及中位数；须线遵循 1.5×IQR 约定）。“非歧视”表示不披露条件。

引文数量上升（每条论断 7.4–8.0 条）：即便可验证性不断下降，模型仍会继续生成参考文献。

很大比例的引文仍无法通过自动方式判定。几乎所有单元格中，未解决结果占引文的 36–61%。人工验证表明，抽样的未解决引文中近一半是伪造的（35 条中有 16 条；表1），因此不应将此类别理解为“几乎正确”。这验证了三分类的合理性：若合并为二元方案，就会掩盖大量确实无法确定的引文。

重新分类的敏感性：未解决。在审计中，35条“未解决”引文中只有15条确实未解决；16条是伪造的，4条已存在（表1）。如果按照这些比率按比例重新分配每个单元格中“未解决”部分的占比，捏造率将升至0.33–0.75，存在率将升至0.06–0.52，而所有约束效应的方向以及专有模型与开放权重模型之间的差距均保持不变。这表明，我们的主要结论是稳健的，但绝对捏造率可能被低估了。

按领域分层分析凸显了软件工程的相关性。 图3按领域组和模型细分了存在率，并汇总了全部五种条件；领域层面的差异应作描述性解读。软件工程与计算机科学组（24条主张，2,926条引文）的存在率为0.132，与跨领域平均值（0.120）相当。社会科学的存在率最高（0.187），

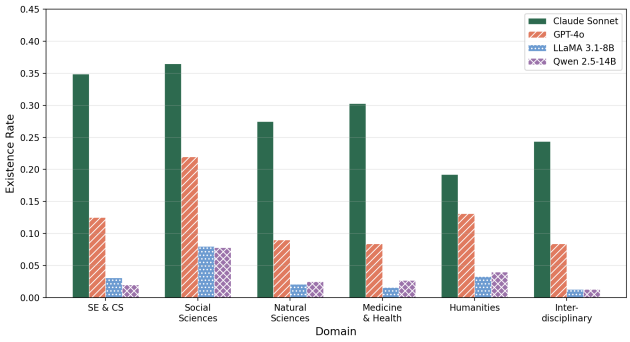

图3：按领域划分的存在率（每组24项主张，汇总各条件）。

表3：关键成对比较中存在率差值（Δ）的自助法95%置信区间。正值 Δ 表示前一项更高。置信区间不包含零的结果以粗体标示；（ns）= 表示置信区间包含零。

比较	Δ	95%置信区间
专有 *vs.* 开放权重（存在率）		
专有 vs. 开放权重（基线）	**+.229**	[.191, .266]
专有 vs. 开放权重（温度）	**+.057**	[.037, .078]
专有 vs. 开放权重（生存）	**+.310**	[.274, .349]
专有 vs. 开放权重（N-D）	**+.200**	[.164, .234]
专有 vs. 开放权重（组合）	**+.051**	[.035, .069]
约束与基线（*Claude Sonnet*）		
温度 −基线 −.261	[−.317, −.207]	**生存 −基线 +.094** [.028, .162]
N-D − 基线（*ns*）	−.032	[−.095, .031]
组合 −基线	**−.275**	[−.329, −.220]
约束与基线（*GPT-4o*）		
温度 −基线	**−.216**	[−.266, −.168]
生存 − 基线（不显著）	−.032	[−.092, .029]
N-D −基线	**−.060**	[−.119, −.001]
组合 −基线	**−.230**	[−.274, −.185]
约束与基线对比（*LLaMA 3.1–8B*）		
温度 −基线 −.057	[−.078, −.038]	**生存 −基线 −.030** [−.048, −.013]
N-D −基线	**−.023**	[−.045, −.001]
组合 −基线	**−.060**	[−.083, −.040]
约束 *vs.* 基线（*Qwen 2.5–14B*）		
温度 −基线	**−.076**	[−.106, −.047]
生存 −基线	**−.070**	[−.101, −.041]
N-D − 基线（*ns*）	−.012	[−.050, .025]
组合 −基线	**−.089**	[−.117, −.063]

这可能反映出其数据库覆盖范围更广。专有模型与开放权重模型之间的差距在每个领域组中都重现：Claude Sonnet在软件工程与计算机科学组中的存在率达到0.349，而两种开放权重模型均低于0.10。这些结果表明，本研究记录的幻觉模式直接适用于软件工程研究人员所产出和使用的文献类型，不过，鉴于样本量有限，子领域之间的比较应谨慎解读（第6节）。

6 讨论

引用幻觉是一种系统性失败，其表现形式取决于施加于模型的约束。我们发现了三个主题。首先是有合规而无实质：在时间条件下，每个模型都能生成格式规范、符合所要求年份范围的书目条目，但存在率却急剧下降——GPT-4o 从 0.235 降至 0.019，而两种开放权重模型的存在率实际上都降为零。因此，格式合规性掩盖了几乎完全丧失的可验证性。其次是未解决的引文问题：在几乎所有单元格中，36–61% 的引文无法通过自动方式确认或证伪，而我们的审计显示，其中约有一半是伪造的（表1）。任何二元的真实或伪造评估都会掩盖这一规模庞大且风险很高的类别。第三，专有模型与开放权重模型之间的差距 (Δ 最高达 +0.310；表3)在所有条件下都持续存在，这表明，与任何单一的提示约束相比，专有模型与开放权重模型之间的区别可能是引文质量更强的预测因素。

6.1 定性错误模式

人工检查伪造和未解决的引文后，发现了四种反复出现的失败模式，而格式层面的检查无法识别这些问题：

- **发表 venue 洗白。** 一个看似合理的发表场所名称（例如，*IEEETSE*、*ICSE*）与一个在该发表场所的索引中——或任何索引中——都不存在的标题相匹配。
- **作者拼凑。** 将目标领域中听起来真实的姓氏重新组合成作者名单，但这些名单并不对应任何实际论文。
- **标识符捏造与遗漏。** 在不披露条件下，DOI字段经常被省略或替换为“n/a”；在其他条件下，模型偶尔会生成语法上看似合理但无效的DOI。
- **标题漂移。** 生成的标题与真实论文标题近似改写——熟悉得足以让人似曾相识，却又差异太大，无法与相似度阈值高于 0.85 的任何索引记录匹配。

这些模式解释了为什么格式合规性与低可验证性会并存：每个字段单独来看似乎 都正确，但组合起来却无法对应到真实出版物。

6.2 实践启示

面向工具和流水线开发者。 仅满足约束并非可靠的质量指标。展示 LLM 生成参考文献的系统应针对多个数据库进行事后验证，并将未解决 视为高风险情况。

面向使用 *LLM* 写作辅助工具的研究人员。 应将生成的引文视为候选项：要求提供持久标识符（DOI/arXiv），交叉核对元数据（Crossref/Semantic Scholar，以及用于计算机科学的 DBLP/OpenAlex），并在纳入前手动处理剩余的未解决项。

面向软件工程社区。 在我们的样本中，软件工程和计算机科学领域的引文总体上与其他领域大致相当。由于综述式提示——最接近系统文献综述撰写的方式 [10]——与更大的专有模型和开放权重模型之间的差距相关，因此依赖开放权重模型的软件工程团队应格外谨慎。

6.3 效度威胁

内部效度。 我们的标签边界（0.85/0.60）和评分权重是统一应用的设计选择；采用其他设置可能会使引文在不同类别之间移动。我们通过由作者开展的人工审计（75%的一致率；κ=0.63）来缓解这一问题。我们在温度为零的条件下进行测试，每种条件使用一个提示模板，每条论断使用一种固定表述；因此，结果反映的是约束类型，而非释义鲁棒性，并且同一论断采用其他措辞时，引文结果可能会有所不同。

外部效度。 我们的144条英语声明对于细粒度比较（例如单个子领域）而言可能统计功效不足，研究结果也可能无法推广到非英语领域或遵循学科特定引用规范的领域。每个领域组包含24条论断，因此按领域分层的分析（图3）仅具有描述性；更细粒度的拆分（例如软件工程子领域中的测试与需求工程）需要大幅增加论断数量。我们在单个时间截面上评估了四种模型；随着开放权重模型扩展规模或引入检索能力，专有模型与开放权重模型之间的差距可能会缩小，而更新版本也可能改变绝对比例。因此，应将结果理解为对结构性模式（约束效应、专有模型与开放权重模型之间的差距）的刻画，而不应视为针对任何模型版本的固定基准。

构念效度。 我们的流水线依赖 Crossref 和 Semantic Scholar；这两者都未收录所有内容，因此某些“伪造”标签可能是误报（例如，未被这两个数据库收录的预印本、研讨会论文或地区性发表场所的文章）。加入 DBLP/OpenAlex 可能会缩小未解决 类别的范围。我们也不会检查已验证的引文是否支持 该论断；引文与论断之间的对应关系是下一步需要关注的重要问题。

6.4 未来工作

有三个方向可以在上述局限之外拓展这项工作。首先，系统地改变声明措辞（检验释义的稳健性），有助于区分提示敏感性与约束效应，并弄清引文输出在多大程度上取决于表层措辞。其次，在相同的约束机制下，将闭卷生成与检索增强设置进行比较，可以分辨幻觉有多少源于缺乏依据 grounding，又有多少源于策略压力。第三，将验证流程嵌入为实时的生成后过滤器——例如，用于手稿草稿的 IDE插件或持续集成检查——可以检验其在软件工程写作工作流程中的实际效用。

7 结论

在四种模型和五种条件下，部署约束始终会加剧引用幻觉，但其表现方式存在显著差异。时间约束会在保持格式合规性的同时，显著降低可验证性；调查式提示会扩大专有模型与开放权重模型之间的差距；不披露指令会使错误从“显而易见的错误”转变为“难以判断的错误”；即使可验证性不断下降，各模型仍会大量生成引用。这些模式表明，仅靠提示工程不太可能解决引用幻觉问题；要实现可靠生成，需要采用检索增强架构、内置验证机制，或同时采用二者。对于软件工程研究人员而言，实际启示十分明确：任何大语言模型生成的参考文献列表都应被视为草稿，在纳入综述、技术报告或工具流程之前，必须通过学术数据库进行独立验证。更广泛地说，我们的基准测试和流水线为持续追踪引用可靠性是否会随着新模型和检索增强架构的出现而提升，提供了可复用的基础。

数据可用性。 复现包（主张数据集、提示模板、流程代码、人工验证标注和原始结果）可在 https://github.com/Zerichen/Citation-Hallucination 获取。

参考文献

[1] Shubham Agarwal、Gaurav Sahu、Abhay Puri、Issam H. Laradji、Krishnamurthy DJ Dvijotham、Jason Stanley、Laurent Charlin 和 Christopher Pal。2024。LitLLM：科学文献综述工具包。arXiv:2402.01788 [cs.CL] doi:10.48550/arXiv.2402.01788[2] Ayush Agrawal、Mirac Suzgun、Lester Mackey 和 Adam Kalai。2024。语言模型知道自己何时产生参考文献幻觉吗？载于 计算语言学协会年会论文集：*EACL 2024*。计算语言学协会，马耳他圣朱利安，912–928。doi:10.18653/v1/2024.findings- eacl.62[3] Aisha Alansari 和 Hamzah Luqman。2025。大语言模型幻觉：全面综述。arXiv: 2510 06265 [ ] 10 48550 2510 06265 . cs.CL [4] doi: . /arXiv. 。2024 Anthropic。。Claude 文档。https://docs.anthropic.com/claude/docs[5] Angela Fan、Beliz Gokkaya、Mark Harman、Mitya Lyubarskiy、Shubho Sengupta、Shin Yoo 和 Jie M. Zhang。2023。用于软件工程的大语言模型：综述与开放问题。载于 *2023 IEEE/ACM* 国际软件工程会议：软件工程的未来（*ICSE-FoSE*）。31–53。doi:10.1109/ICSE-FoSE59343.2023.00008[6] Tianyu Gao、Howard Yen、Jiatong Yu 和 Danqi Chen。2023。让大语言模型生成带引文的文本。载于 *2023* 年自然语言处理实证方法会议论文集。计算语言学协会，新加坡，6465–6488。doi:10.18653/v1/2023.emnlp- main.398[7] Xinyi Hou、Yanjie Zhao、Yue Liu、Zhou Yang、Kailong Wang、Li Li、Xiapu Luo、David Lo、John Grundy 和 Haoyu Wang。2024。用于软件工程的大语言模型：系统性文献综述。*ACM* 软件工程与方法学汇刊 33，8（2024），1–79。doi:10.1145/3695988[8] Lei Huang、Weijiang Yu、Weitao Ma、Weihong Zhong、Zhangyin Feng、Haotian Wang、Qianglong Chen、Weihua Peng、Xiaocheng Feng、Bing Qin 和 Ting Liu。2025。大语言模型中的幻觉综述：原则、分类体系、挑战与开放问题。*ACM* 信息系统汇刊 43, 2 2025 10 1145 3703155 （ ）。doi：。/ Adam Tauman Kalai、[9] Ofir Nachum、Santosh S. Vempala 和 Edwin Zhang。2025。语言模型为何产生幻觉。arXiv:2509.04664 [cs.CL]doi:10.48550/arXiv。2509.04664[10] Barbara Kitchenham 和 Stuart Charters。2007。软件工程系统性文献综述指南。技术报告 EBSE-2007-01。基尔大学和杜伦大学。https://ebse.webspace.durham.ac.uk/resources/guidelines-for-performing-systematic-literature-reviews-in-software-engineering/ 版本 2.3。[11] Nelson F. Liu、Tianyi Zhang 和 Percy Liang。2023。评估生成式搜索引擎中的可验证性。载于 计算语言学协会年会论文集：*EMNLP 2023*。计算语言学协会，新加坡，7001 7025 10 18653 1 2023 467 – 。doi：。/v / 。[12] findings-emnlp。Joshua Maynez、Shashi Narayan、Bernd Bohnet 和 Ryan McDonald。2020。论抽象式摘要的忠实性与事实准确性。载于 计算语言学协会第 *58* 届年会论文集。计算语言学协会，1906–1919。doi:10.18653/v1/2020.acl-main.173 [13] Nick McKenna、Tianyi Li、Liang Cheng、Mohammad Hosseini、Mark Johnson 和 Mark Steedman。2023。大语言模型在推理任务中产生的幻觉来源。载于 计算语言学协会年会论文集：*EMNLP 2023*。计算语言学协会，新加坡，2758–2774。doi: 10 18653 1 2023 182 。/v / 。findings-emnlp。Meta AI。。[14] 推出 Llama 。：迄今为止能力最强的模型。https: // 2024 3 1 ai.meta.com/blog/meta-llama-3-1/ [15] OpenAI。2024。GPT-4o 系统卡。https://openai.com/index/gpt-4o-system- card/[16] Nathaniel Schenker 和 Jane F. Gentleman。2001。通过考察置信区间的重叠来判断差异的显著性。《美国统计学家》55, 3 2001 182 186 10 1198 000313001317097960 （ ），– 。[17] doi：。/ Xuemei Tang、Xufeng Duan 和 Zhenguang G. Cai。。用于自动化文献综述的大语言模型：对参考文献生成、摘要撰写和综述撰写的评估。载于 *2025* 年自然语言处理实证方法会议论文集。计算语言学协会，中国苏州，2025 1602 1617 10 18653 1 2025 83 – 。doi：。/ [18] v / 。emnlp-main。S.M Towhidul Islam Tonmoy、S M Mehedi Zaman、Vinija Jain、Anku Rani、Vipula Rawte、Aman Chadha 和 Amitava Das。2024。大语言模型幻觉缓解技术全面综述。arXiv: 2401 01313 [ ] 10 48550 2401 01313 。cs.CL doi：。/arXiv。。[19] William H. Walters 和 Esther Isabelle Wilder。2023。。ChatGPT 生成的书目引文中的捏造与错误。《科学报告》13，1（2023），14045。doi:10.1038/s41598-023-41032-5 [20] An Yang 等。2025。Qwen2.5 技术报告。arXiv:2412.15115 [cs.CL] doi:10.48550/arXiv.2412.15115