arXiv:2603.17339v1 [cs.DL] 18 Mar 2026

CITECHECK：用于学术手稿自动化书目核验与修复的 MCP 服务器

预印本

Junhyeok Lee
ORCID: 0000-0001-7489-5829
evening0619@gmail.com

2026年3月19日

摘要

学术手稿中的参考文献列表经常包含各种错误，包括标识符错误、元数据不完整、作者归属有误，以及预印本版本与正式发表版本不匹配。这些问题手动修复起来十分繁琐；而在依赖大语言模型的工作流程中，这类问题也愈发突出，因为大语言模型可能会编造或篡改引文。我们介绍 *citecheck*，这是一个基于 TypeScript 的系统和 MCP 服务器，可在类论文项目文件夹中自动进行书目核验与修复。给定稿件文件或工作区后，citecheck 会选取最可能的论文文件，从 .bib、.tex、.md、.txt 或 .docx 中提取参考文献，通过 PubMed、Crossref、arXiv 和 Semantic Scholar 核验条目，并返回结构化的更正建议及替换安全诊断。当前代码仓库提供了一个可运行的研究原型，具备多轮检索、感知版本形态的匹配、策略门控的改写规划，并有 47 项通过的测试，覆盖修复行为、格式错误的载荷处理、传输故障及 MCP 接入。我们将 citecheck 定位为智能体学术编辑的基础设施，并将其视为一道实用防线，既可应对传统参考文献错误，也可防范由大语言模型引发的引文幻觉。

关键词 参考文献修复，书目核验，模型上下文协议，学术交流，人工智能智能体

1 引言

参考文献列表是学术写作中最难实现自动化的部分之一。手稿中常有条目不完整、缺少 DOI、作者格式不一致，以及预印本与其后续期刊版本或会议版本不匹配等问题。人工核验效率低下，因为作者必须逐条交叉比对多个外部注册库。

在使用大语言模型（LLM）的工作流程中，这一长期存在的问题变得更加紧迫。引文幻觉已有充分记录：模型可能捏造参考文献、混合不同论文的元数据，并生成看似合理但无效的标识符 [17, 12]。近期的大规模研究表明，这一问题已在已发表文献和预印本文献中显现，尤其是在发展迅速的人工智能领域会议中 [18, 2]。这些报告发布不久，其中一些仍是预印本，因此应谨慎看待其具体发生率，但问题的发展趋势十分明确。在大语言模型辅助写作出现之前，参考文献错误就已十分常见。先前的研究发现，公共卫生、社会工作和医学文献中的错误率相当高 [9, 16, 14]。变化之处在于实际应用场景：作者和人工智能智能体越来越需要一种工具，能够直接核验手稿中已有的参考书目，而不仅仅是格式化或存储经过整理的文献库条目。

我们介绍*citecheck*，一种原生支持 MCP 的书目核验与修复系统 [3]。citecheck 不要求用户预先准备干净的参考书目文件，而是接受项目路径，识别最像论文的源文件，提取参考文献，根据外部学术元数据服务对其进行验证，并返回修正后的输出以及

明确的诊断信息。该系统有意采取保守策略。它将参考文献修复既视为检索问题，也视为安全问题，并区分仅供审阅的发现与可安全覆盖的替换内容。
本工作的主要贡献如下：

- 一个 MCP 工作流，将工作区扫描、参考文献分析、重写规划和受控应用作为独立工具提供；
- 一个涵盖 PubMed、Crossref、arXiv 和 Semantic Scholar 的多来源、感知文献表现形式的核验流程 [15, 7, 4, 1]；
- 一个策略层，用于阻止不安全的替换，并提供整理工作清单，而不是悄然重写含义模糊的条目；
- 一篇面向 arXiv 的软件论文，介绍当前系统、其基于代码仓库的证据以及局限性。

2 相关工作与研究动机

Zotero、Mendeley 和 EndNote 等现有参考文献管理工具，主要用于整理和格式化用户自行维护的文献库 [6, 10, 5]。它们对于管理参考文献不可或缺，但不会独立核验手稿中的当前参考文献列表是否与权威外部记录相符。

其他工具则涵盖了这一问题的相关环节。AnyStyle 和 GROBID 等解析器可从原始引文字符串或文档中提取结构化元数据 [11, 13]。doi2bib 等单一用途工具可将单个标识符解析为格式化的 BibTeX 条目 [8]。Crossref、PubMed、arXiv 和 Semantic Scholar 等元数据 API 提供权威记录或用于信息补充的记录 [7, 15, 4, 1]。然而，这些组件单独都无法处理端到端任务：选择稿件文件、提取参考文献、与多个来源交叉验证、分类差异，并为代理工作流生成可安全替换的输出。
这一缺口对两类用户都很重要。研究人员需要在投稿前进行实用的最终核验。AI 辅助写作环境的开发者需要一个可由代理调用的程序化后端，而不是以图形用户界面为先的参考文献管理器。citecheck 面向这两类用户：它既可作为机器可调用的核验服务，也适用于常见的论文文件。

3 系统概述

citecheck 围绕项目文件夹而非孤立的引文字符串进行设计。其顶层工作流程如下：

1. 扫描工作区或检查单个文件路径；2. 选择最可能的手稿或参考书目文件；3. 提取参考文献部分并规范化条目；4. 查询外部元数据来源；5. 对候选匹配项进行聚类并确定首选表现形式；6. 执行替换安全策略检查；7. 输出结构化审查结果以及可选的重写补丁。

MCP 服务器通过六个工具公开此工作流程：scan_workspace、analyze_references、plan_reference_rewrite、apply_reference_rewrite、repair_paper 和 citecheck_version。将分析与写回分离是有意为之。在代理环境中，不应将二者混为一谈。

4 方法

4.1 文档选择与提取

工作区选择采用文件扩展名优先级、reference、paper、manuscript 和 draft 等文件名提示，以及内容启发式规则。系统会跳过隐藏目录或生成目录。对于纯文本手稿，citecheck 会检测“参考文献”“参考书目”和“引用作品”等章节标题。对于 .tex 文件，它会追踪由 \\bibliography 和 \\addbibresource 等命令引用的参考文献资源。对于 .docx 文件，当前实现会从文档 XML 中提取段落文本。

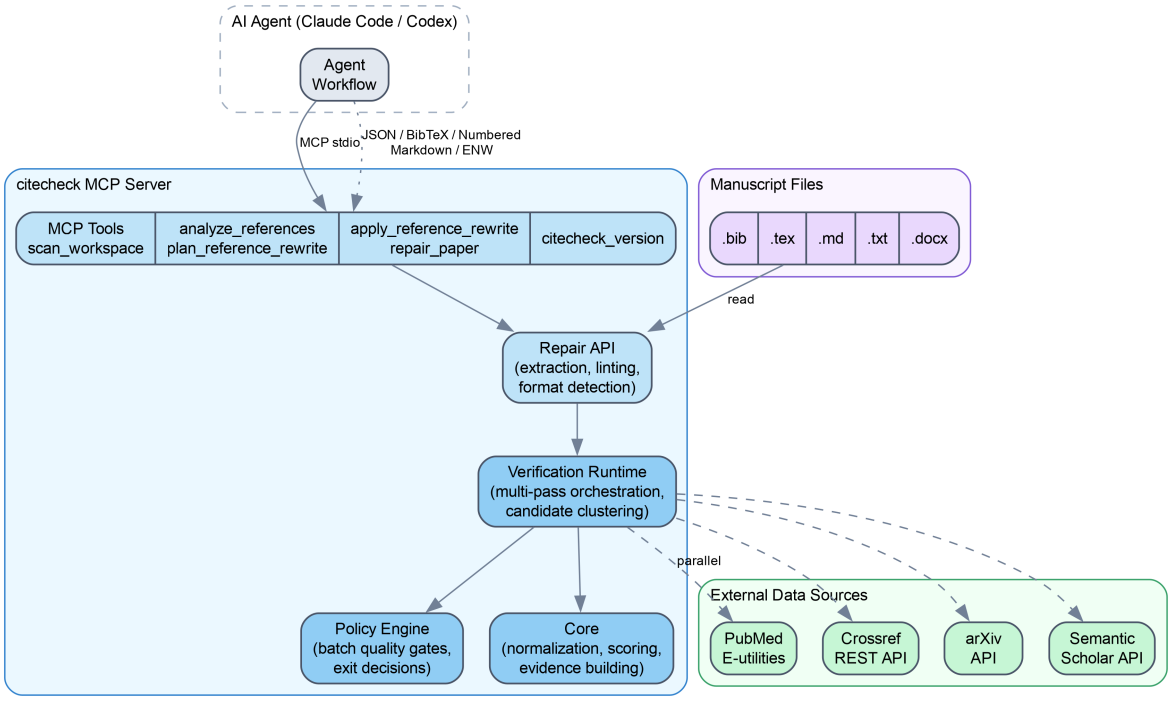


图 1：citecheck 的简化架构。MCP 服务器将任务委托给修复 API，运行时通过外部连接器协调多轮核验，而策略引擎则在写回前评估替换安全性。

表 1：当前实现支持的输入和输出。

类别	当前支持
输入文件	.bib、.tex、.md、.txt、.docx
模式	审阅、替换
输出格式	JSON、BibTeX、编号文本、Markdown、EndNote
写入模式	预览，伴随文件，替换
来源	PubMed ，Crossref ，arXiv ，Semantic Scholar

4.2 多源验证

没有任何单一书目注册库涵盖所有相关文献。PubMed 是生物医学记录的权威来源。Cross-ref 提供广泛的、以数字对象标识符为中心的覆盖范围。arXiv 提供预印本元数据。Semantic Scholar 则提供额外的丰富信息，包括引文相关元数据。citecheck 会并行查询已启用的来源，因此单个薄弱来源不一定会导致结果变为未解决。

4.3 多轮检索与匹配

提取出的条目会先规范化为内部参考文献输入类型，然后最多经过三轮检索。第一轮使用最直接的查询。若证据仍然薄弱，后续轮次会重新组织搜索。检索到的候选项会去重、聚类，并依据标题、作者、发表 venue、年份和标识符等信号与规范化后的查询进行比对。随后，系统会识别标题不匹配、标识符冲突、年份不一致或证据不足等问题。

表 2：撰写本文时，代码仓库中可见的自动化测试覆盖情况。

测试组	数量
修复流水线测试	17
连接器与 HTTP 行为测试	8
运行时、工厂、策略与核心测试	10
MCP 集成测试	3
评估与回归测试夹具测试	9
总计	47

4.4 感知呈现形式的解析

同一项底层作品可能以期刊文章、会议论文或预印本的形式呈现。citecheck 会明确对此加以建模，并在存在关联呈现形式时应用优先级策略“期刊 > 会议 >arXiv”。重要的是，呈现形式之间的差异不会在未作说明的情况下被悄然改写。如果证据支持将条目置于仅供审核状态，而非安全地进行替换，该条目就会继续标记为待整理。

4.5 策略门控的重写规划

自动化参考文献替换风险较高，因为看似合理但实际错误的匹配可能会悄无声息地损坏手稿。因此，citecheck 将分析与修改分开。策略引擎会检查未_检查、未解决、已验证和需要_审阅条目的批次级比例，以及传输、身份验证、速率限制和载荷结构故障等故障类别。名为默认、严格和宽松的预设会将这些信号转化为退出决策。

这种设计将安全元数据置于核心位置。即使结果有助于审阅，也可能因重复键、不安全的引文键重写或表现形式冲突而无法进行原位替换。因此，重写规划会先生成补丁预览，然后才尝试写入伴随文件或进行原位写入。

5 实现

当前代码仓库是一个 TypeScript 工作区，应用程序和库代码合计约 6.8k 行。MCP 标准输入输出服务器封装了修复 API，并直接向 Codex 或 Claude Code 等客户端提供工具接口。运行时工厂同时支持实时 HTTP 连接器和基于测试夹具的重放，因此无需每项测试都访问外部 API，也能进行确定性的回归测试。

在 API 层，citecheck 返回的不只是修正后的参考书目文本，还包括逐条状态、置信度信息、证据追踪记录、参考书目 lint 检查结果、整理工作清单、重复键检测、来源健康度摘要、故障摘要、替换状态以及明确的键映射风险。这种结构化输出是智能体集成的核心，因为它使调用方能够区分“当前可用的最佳建议”和“可安全自动应用的建议”。

6 当前证据

该代码仓库目前尚未提供涵盖数千条参考文献的大型基准测试，我们也不声称已经做到这一点。相反，我们报告的是当前代码库能够直接支持的内容。

测试套件涵盖了对 BibTeX、Markdown、LaTeX 和 DOCX 风格输入进行端到端修复的测试；替换规划和伴随文件重写行为；重复键拦截；表现形式冲突；格式错误的来源载荷分类；传输和速率限制故障；以及 MCP 工具辅助函数行为。代码仓库还包含小规模的基于测试夹具的评估检查，其中包括一项离线生物医学 DOI 恢复回归测试。这些测试夹具有助于防止回归，但不应被误认为是广泛的实证基准测试。

在内部使用中，citecheck 已应用于实际稿件准备，作为一种实用的验证辅助工具。我们将这一经验视为其效用的定性证据，而非正式评估结果。目前，该代码仓库更充分地展现了工程完整性和面向安全的行为，而非基准测试规模。

7 讨论与局限

citecheck 位于一个特定的系统领域。它既不是完整的参考文献管理器，也不是通用的文献检索引擎。它的作用是衔接杂乱的手稿材料与智能体编辑工作流所需的结构化输出。这一定位解释了两项核心设计选择：第一，参考文献修复被视为文件与工作区问题，而非单纯的字符串问题；第二，安全诊断与候选项检索同等受到重视。

该系统也存在明显的局限。评估仍处于小规模阶段。版本形态策略基于启发式规则，并未实现完整的图谱覆盖。Semantic Scholar 为可选项，且在运行时工厂中默认禁用。提取功能适用于多种类论文格式，但并未尝试实现完整的文档版面理解。我们也尚未报告用户研究或与替代工具进行的大规模比较基准测试。

有前景的后续工作包括扩大基于测试夹具的评估，构建一个针对格式错误参考文献的真实基准测试，测量真实 LaTeX 项目中的引文键重写准确率，并研究仅供审阅与安全替换运行模式在下游作者工作流中的权衡。

8 结论

我们介绍了 citecheck，这是一个用于学术手稿和类论文项目文件夹中自动化书目核验与修复的 MCP 服务器。该系统结合了感知工作区的提取、多源元数据检索、感知版本形态的匹配，以及策略门控的改写规划。当前代码仓库已提供具备结构化输出和自动化测试的功能原型，同时也明确说明了其局限。我们认为，citecheck 可作为可靠学术编辑系统的基础设施，并能有效防范传统参考文献错误以及新出现的 LLM 驱动的引文错误。

参考文献

[1] Allen 人工智能研究所。Semantic Scholar 学术图谱 API，2024年。网址：https://api.semanticscholar. org。访问日期：2026年3月18日。[2] Samar Ansari。《精英同行评审中的复合欺骗：NeurIPS 2025 伪造引文的失效模式分类》，arXiv 预印本 arXiv:2602.05930，2026年。[3] Anthropic。模型上下文协议规范，2024年。网址：https://modelcontextprotocol.io。版本：2025年3月26日。[4] arXiv。arxiv API，2026年。网址：https://info.arxiv.org/help/api/。访问日期：2026年3月18日。[5] 科睿唯安。EndNote，1988年。网址：https://endnote.com。[6] 数字学术公司。Zotero，2006年。网址：https://www.zotero.org。版本：7。[7] Crossref。Crossref REST API，2024年。网址：https://api.crossref.org。访问日期：2026年3月18日。[8] doi2bib 贡献者。doi2bib：将 DOI 转换为 BibTeX 条目，2014年。网址：https://www.doi2bib.org。[9] Pat Eichorn 和 Alfred Yankauer。作者会核查参考文献吗？对三种公共卫生期刊参考文献准确性的调查。《美国公共卫生杂志》，77(8):1011–1012，1987年。数字对象标识符：10.2105/ajph.77.8.1011。[10] 爱思唯尔。Mendeley 参考文献管理器，2008年。网址：https://www.mendeley.com。[11] Sylvester Keil。AnyStyle：快速智能的引文解析，2024年。网址：https://anystyle.io。[12] Jake Linardon、Hannah K. Jarman、Zoe McClure、Cleo Anderson、Claudia Liu 和 Mariel Messer。主题熟悉度和提示具体程度对使用大语言模型进行心理健康研究时引文捏造的影响：一项实验研究。《JMIR 心理健康》，12:e80371，2025年。数字对象标识符：10.2196/80371。[13] Patrice Lopez。GROBID：结合自动书目数据识别与术语提取，用于学术出版物。在《数字图书馆理论与实践国际会议》论文集，第473–474页。施普林格，2009年。[14] Scott A. Mogull。医学研究论文中所引“事实”的准确性：对研究方法的审阅及引文错误率的重新计算。《公共科学图书馆·综合》，12(9):e0184727，2017年。数字对象标识符：10.1371/journal.pone.0184727。[15] 美国国家生物技术信息中心。Entrez 编程工具（E-utilities），2024年。网址：https: // www.ncbi.nlm.nih.gov/books/NBK25501/。访问日期：2026年3月18日。[16] Christina A. Spivey 和 Scott E. Wilks。社会工作期刊参考文献列表的准确性。《社会工作实践研究》，14(4):281–286，2004年。数字对象标识符：10.1177/1049731503262131。

[17] William H. Walters 和 Esther I. Wilder。ChatGPT 生成的书目引文中的捏造与错误。*ScientificReports*，13:14045，2023年。数字对象标识符：10.1038/s41598-023-41032-5。[18] Zuyao Xu、Yuqi Qiu、Lu Sun、FaSheng Miao、Fubin Wu、Xinyi Wang、Xiang Li、Haozhe Lu、ZhengZe Zhang、Yuxin Hu 等。《GhostCite：大语言模型时代引文有效性的大规模分析》。*arXiv* 预印本 *arXiv:2602.06718*，2026年。