arXiv:2603.19316v1 [cs.DL] 15 Mar 2026

医学文献中检测到的人工智能生成文本日益普遍：开放获取文章的纵向分析

NathanWolfrath[1,2], Simrin Patel[1], Madelyn Flitcroft[1], Anjishnu Banerjee[1],Melek Somai[1],Bradley H. Crotty[2],Anai N Kothari[1,2]

[1]威斯康星医学院 Bud and Sue Selig 外科数据科学中心[2]威斯康星医学院 Inception Health Labs[3]威斯康星医学院生物统计学系

1 引言

生成式人工智能（AI）工具改变了各个领域文本的生成方式，其中包括学术写作。商业和开放权重大语言模型（LLM）的迅速普及，引发了人们对学术内容真实性和来源，以及人工智能在出版流程中应发挥何种作用的新疑问。这些工具的广泛普及促使许多学术出版商发布强制披露要求及其他相关政策，以指导人们在写作中使用人工智能。然而，对于人工智能生成内容实际出现在同行评审医学文献中的程度，人们知之甚少。

随着大语言模型的兴起，学术界和商业机构也相继开发了各种用于检测人工智能生成文本的工具和方法。这些方法的复杂程度差异很大，涵盖从纯粹的统计分析到深度学习和基于变换器的方法[10]。由于生成式人工智能领域发展迅速，且新的大语言模型不断发布，这些工具的准确性仍是持续研究的课题。然而，人们已越来越多地致力于以标准化方式严格评估检测器的性能[2, 11]。

在这项简短研究中，我们旨在使用商业检测软件，评估高质量开放获取医学文献中AI生成写作的普遍程度及趋势。我们考察了从2022年到2025年初这一分析时段内的绝对比例和变化趋势。此外，我们还将作者披露的AI使用情况与检测到的使用比例进行了比较。

2 方法

通过网页抓取，我们从期刊网站获取了JAMA Network Open自2022年1月至2025年3月发表的所有原创研究、研究简报和特邀评论文章的正文纯文本。期刊已批准将其用于此项研究。之所以选择JAMA Network Open，是因为其文章数量较多、涵盖多个具有标准化分类的领域、可获取开放获取的完整稿件，并且影响因子较高 [1]。

使用 Python 进行网络爬取。检索后，使用 Beautiful Soup 解析每个网页，从文章元数据中提取稿件正文文本以及发表月份、出版格式和领域。通过程序移除了文内参考文献编号。使用基于正则表达式的匹配，检查每篇文章的致谢部分（JAMA Network 网站上披露人工智能使用情况的部分），以确定是否披露了在写作过程中使用语言模型。搜索词包括 ChatGPT、GPT、Claude、Gemini、Llama、Mistral、Mixtral、Sonnet、ChatGLM 和 Grok，并采用词边界匹配。对匹配结果进行了人工审核，以确认其有效性。对于提及使用大语言模型编写分析代码、生成图表或用于稿件文本起草或编辑以外其他用途的致谢内容，不纳入分析。

使用商业人工智能检测器 Originality.AI [8], 检测人工智能生成文本。该工具此前已在学术研究中接受评估，且相较于其他检测器表现稳定良好 [5, 6, 3]。对每篇文章进行分析，以计算稿件中人工智能生成文本占比超过 10% 的概率（该阈值由模型制造商定义）。按照 Originality.AI 的建议，将概率以 0.5 为阈值转换为二元标签。随后按发表月份汇总所得数据，用于时间序列分析；并按出版类型和领域汇总，以进行进一步比较。使用卡方检验确定不同出版类型和领域的分布是否存在显著差异。使用 Mann-Kendall检验对所分析时段内的单调趋势进行检测。

3 结果

共检索并分析了7251篇文章，其中195篇（2.7%）在整个分析时段内被归类为含有人工智能生成文本。随着时间推移，每月发表的含有人工智能生成文本的文章比例总体呈上升趋势，但期间有所波动。该比例从分析首月（2022年1月）的0.0%升至末月（2025年3月）的11.3%。这一时期呈现显著的上升趋势（P $<$0.001）。

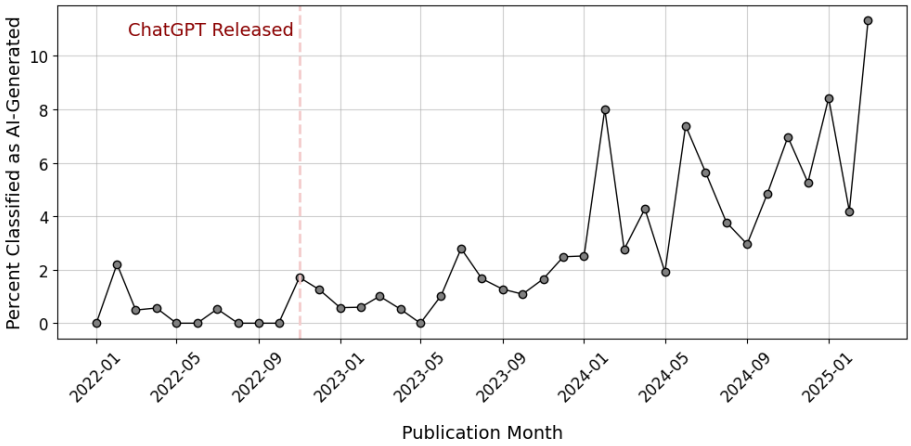


图1：自2022年1月以来，每月被归类为受人工智能影响的稿件比例。ChatGPT由OpenAI于2022年11月30日发布。

不同出版类型中检测到的大语言模型文本比例存在显著差异（$P<0.001$）；特邀评论中被归类为含有人工智能生成文本的文章比例最高（6.7%），其次是原创研究（2.2%）和研究简报（1.4%）。不同领域之间的差异也具有统计学意义（$P=0.04$）；全球健康（8.6%）、遗传学与基因组学（5.7%）和医学教育（5.6%）中，被归类为含有人工智能生成文本的出版物比例最高，而有几个领域的所有文章均被归类为完全由人工撰写（图2）。

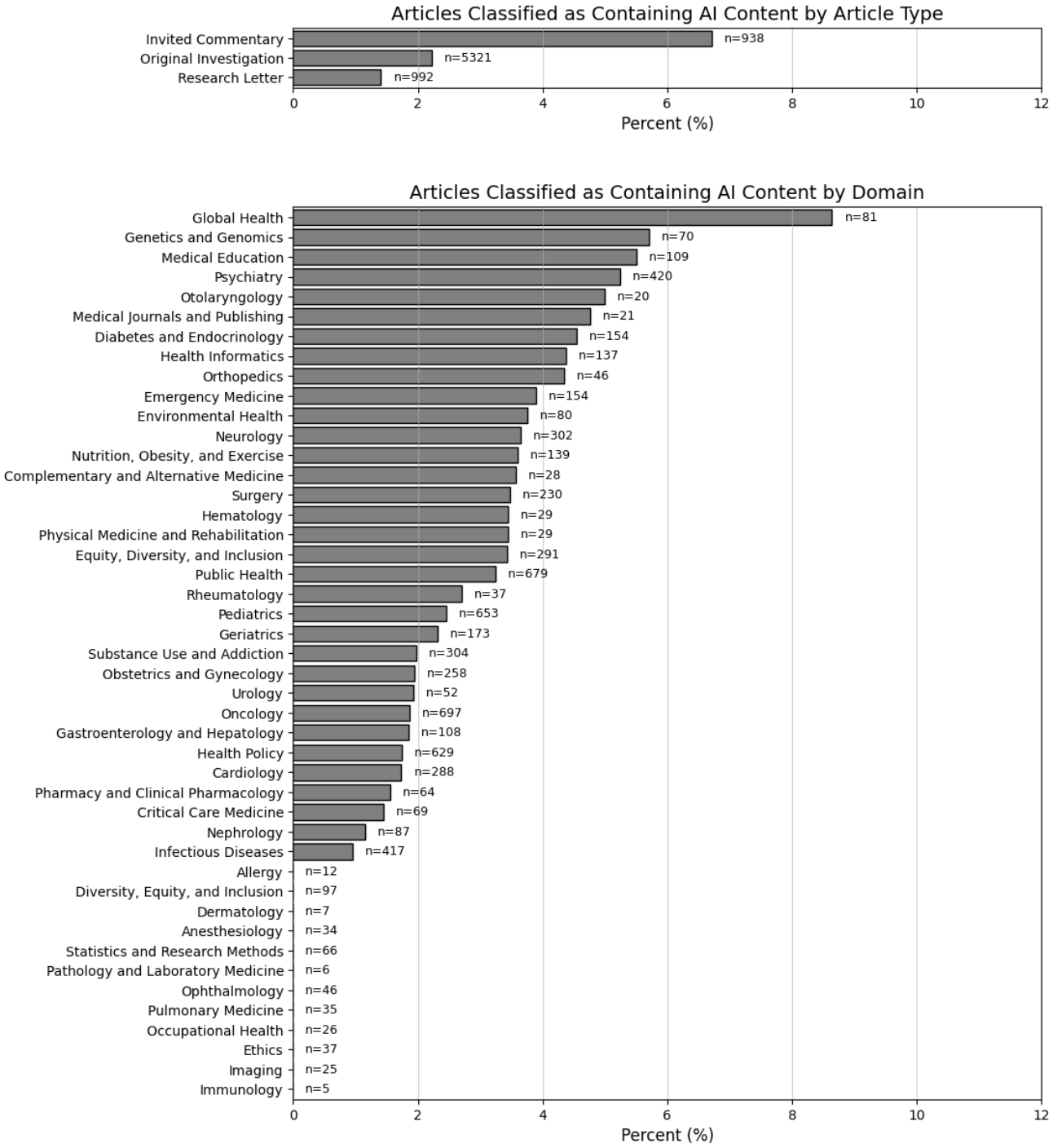


图2：自2022年1月以来，按出版类型和领域分层的受人工智能影响稿件占比。N表示各领域分析的文章总数。4

在所有文章中，有15篇（0.2%）披露在写作过程中使用了大语言模型，其中6篇（40.0%）被归类为人工智能生成文本占比超过10%。在检测出人工智能生成文本的195篇文章中，披露在写作过程中使用人工智能的文章占3.1%。

4 讨论

本分析显示，2022年11月ChatGPT发布后，检测为人工智能生成的内容呈上升趋势，其中以特邀评论中的占比最高。在各个领域中，全球健康领域被判定为人工智能生成的内容最多。据推测，主要不以英语为母语的作者可能需要大语言模型提供写作辅助，不过，以第二语言写作可能会影响检测器的表现 [7]。本研究未根据作者语言开展单独分析，因为若不采用机构所在地等不准确的替代指标，就无法可靠地确定作者的母语。

截至2025年3月，11.2%的出版物被归类为含有人工智能生成文本。尽管JAMA Network Open制定了披露人工智能使用情况的政策，但已发表的稿件中只有少数对此作出说明。随着生成模型和检测模型不断发展，仍有必要持续评估两者的性能与可靠性，以便为指导和规范学术写作中人工智能使用的政策提供依据。

重要的是，人工智能检测工具存在缺陷，因此必须谨慎解读结果。例如，在生成式人工智能工具普及之前，被标记为含有人工智能生成内容的文章，每月占分析稿件的0–2.2%。这些很可能是假阳性，不过，在当前大语言模型出现之前就已存在的工具（如文章改写工具、语法纠正工具或翻译软件）可能会影响这些分类。

在提及使用大语言模型的论文中，40%被检测器判定为阳性。大语言模型的使用程度尚不清楚，可能不足以达到10%的阈值，也可能属于假阴性。尽管这些研究披露了大语言模型的使用情况，但其中由大语言模型生成内容的实际比例尚不明确。鉴于这一点以及披露使用情况的样本量较小，我们没有尝试根据这一亚组对检测工具进行进一步校准。

最后，检测器的性能可能因多种因素而存在显著差异，包括不同的语言模型、模型参数、上下文、具体用途（例如编辑或改写，而非从头生成）以及规避检测的策略 [4, 9]。因此，这些结果不能被视为对使用 AI 的出版物或具体文章比例的确定性描述，而应视为基于当前最佳可用技术得出的估算。

5 致谢

2025年4月至5月期间，研究使用 GPT-4o 协助编写分析和图表生成代码。输出经人类专家审查验证，作者对其准确性负责。本研究使用的所有数据（文本稿件）均为开放获取；我们感谢 *JAMA Network Open* 批准此项用途。作者未获得这项工作的资助，也与 OpenAI 或 Originality.AI 没有财务关系或其他关联。

参考文献

[1] 美国医学会。*JAMA Network*。https://jamanetwork. com/。访问日期：2025年5月。JAMA Network / 美国医学会，2025。[2] Liam Dugan 等。“Raid：用于稳健评估机器生成文本检测器的共享基准”。载于：计算语言学协会第 *62* 届年会论文集（第 *1* 卷：长篇论文）。2024 年，第 12463–12492 页。[3] Madelyn A Flitcroft 等。“人工智能内容检测器对人类和人工智能生成科学写作的检测性能”。载于：外科肿瘤学年鉴 31.10（2024 年），第 6387–6393 页。[4] Kathleen C Fraser、Hillary Dawkins 和 Svetlana Kiritchenko。“检测人工智能生成文本：影响当前方法可检测性的因素”。载于：人工智能研究杂志 82（2025 年），第 2233–2278 页。[5] Jining Han、Yuying Yang 和 Geping Liu。“教师评估的是学生写的作业，还是人工智能写的作业？生成式人工智能内容检测研究的快速文献综述”。载于：欧洲教育杂志 60.4（2025 年），e70240。[6] Frederick M Howard 等。“2021 至 2023 年肿瘤学科学摘要中人工智能内容检测数量增长的特征分析”。载于：*JCO* 临床癌症信息学 8（2024 年），e2400077。[7] Weixin Liang 等。“GPT 检测器对非英语母语写作者存在偏见”。载于：*Patterns* 4.7（2023 年）。[8] Originality.AI。*Originality.ai*。https://originality.ai/。访问日期：2025年5月。Originality.AI，2025 年。[9] Vinu Sankar Sadasivan 等。“能否可靠地检测人工智能生成文本？”载于：*arXiv* 预印本 *arXiv:2303.11156*（2023 年）。[10] Junchao Wu 等。“大语言模型生成文本检测综述：必要性、方法与未来方向”。载于：计算语言学51.1（2025 年 3 月），第 275–338 页。 DOI: 10.1162/coli_a_00549。 URL: https: //aclanthology.org/2025.cl-1.8/。

[11] Junchao Wu 等人：《Detectrl：在真实场景中对大语言模型生成文本检测进行基准测试》。载于：神经信息处理系统进展 37 (2024)，第 100369–100401 页。