大语言模型时代的科学产出

Keigo Kusumegi[1†], Xinyu Yang[1†], Paul Ginsparg[1], Mathijs de Vaan[2*], Toby Stuart[2*], Yian Yin[1*]

[1]信息科学系，康奈尔大学，美国纽约州伊萨卡[2]哈斯商学院，加州大学伯克利分校，美国加利福尼亚州伯克利

†这些作者对本研究的贡献相同 *通讯请联系：mdevaan@haas.berkeley.edu、tstuart@haas.berkeley.edu、yian.yin@cornell.edu。

大语言模型（LLM）正在迅速重塑科学研究。我们利用多个大规模数据集分析这些变化，数据集涵盖 210 万份预印本、2.8 万份同行评审报告，以及对科学文献的 2.46 亿次在线访问。我们发现：1）采用大语言模型起草手稿的科学家，论文产出大幅增加，增幅为 23.7% 至 89.3%，具体取决于科学领域和作者背景；2）大语言模型的使用颠覆了写作复杂度与论文质量之间的关系，导致大量语言复杂、但实质内容乏善可陈的手稿涌现；3）大语言模型的使用者访问并引用了更多样化的既有研究，包括书籍以及较新、引用较少的文献。这些发现凸显了科学产出正在经历一场惊人的转变，期刊、资助机构和终身教职评审委员会很可能需要改变评估科学成果的方式。

这是作者版本的作品。经美国科学促进会许可，此处仅供个人使用，不得再分发。最终版本于 2025年12月18日发表于*Science*，DOI：10.1126/science.adw3000。

最终版本链接　已发表版本：https://www.science.org/doi/10.1126/science.adw3000

科学事业与技术创新密切相关。例如，显微镜（*1*）、计算领域的进步（*2, 3*）以及新一代测序仪（*4*）都推动了研究前沿的发展。如今，生成式人工智能（Gen AI）在各个学术学科中的迅速普及（*5-8*），正在重塑科学产出。尽管人们对生成式人工智能在研究中的作用日益感到振奋（也日益担忧），但实证证据仍然零散，对于大语言模型（LLM）在各科学领域所产生影响的系统性认识仍然有限。

研究人员已在许多具体的科学研究场景中证明了人工智能的价值（*9-11*），例如蛋白质结构预测（*12*）和材料发现（*13*）。大语言模型的最新进展拓展了其在自然科学（*14-16*）和社会科学（*17-20*）众多任务中的应用。已有研究凸显了大语言模型在具体科学研究中的巨大潜力，同时也提出了一个尚无定论的问题：大语言模型对科学事业的宏观影响是什么?

为回答这一问题，我们从三个预印本仓库收集了大规模数据（2018年1月至2024年6月，详情见补充材料 S1.1-1.3）：（1）arXiv（120万篇预印本），涵盖数学、物理学、计算机科学、电气工程、定量生物学、统计学和经济学；（2）bioRxiv（22.1万篇预印本），涉及生物学和生命科学的众多子领域；（3）社会科学研究网络（SSRN，67.6万篇预印本），这是一个工作论文仓库，收录社会科学、法律和人文学科领域的手稿。这三个数据集各自都是所属领域内规模最大的。总体而言，它们为考察大语言模型对众多科学领域科研实践及科学生产力的影响，提供了前所未有的实证基础。

为识别科学论文创作过程中大语言模型的使用情况，我们开发了一种基于文本的人工智能检测算法（*5*），并将其应用于数据集中的所有摘要。我们使用2023年之前提交的论文摘要——即ChatGPT时代到来之前的摘要——来估算人类撰写文本的词元分布；随后，我们提示OpenAI的GPT-3.5turbo0125模型改写这些摘要，以生成大语言模型撰写文本的词元（单词）分布。我们将这两种分布进行比较，以识别ChatGPT 3.5发布后撰写、可能由大语言模型辅助完成的摘要。有关模型训练、验证、潜在局限性以及大语言模型检测替代方法的更多详情，请参见SM S2.1、S4和S5。

结果

大语言模型使用与科学生产力

我们预测，采用大语言模型的作者将提高生产力（*21-23*）。为将大语言模型的一般生产力效应与人工智能研究的快速增长区分开来，我们首先从样本中排除人工智能核心子学科的手稿（SM S1.1, S5.7）。随后，我们将作者首次采用大语言模型的时间界定为其首篇表现出大语言模型辅助统计特征的手稿（m_i）（α），即 $\alpha(m_j) > \tau$，，其中 τ 是检测阈值。检测到使用后，作者在之后所有月份的采用状态均由 0 变为 1。基于这一衡量方式，我们在作者层面固定效应事件模型中考察大语言模型采用者与

相似的未采用者（SM S2.2-2.4）在采用前后的手稿提交率变化（SM S3.1）。

图1显示，在三个预印本仓库中，大语言模型的采用都与研究人员科学产出的显著增加相关。arXiv、bioRxiv 和 SSRN 的效应量分别为36.2%、52.9%和59.8%，表明大语言模型使用与生产力的大幅提升相关。尽管根据用于识别大语言模型采用情况的检测方法和阈值不同，估计系数会有所变化，但敏感性分析（SM S5.3-5.6）表明，无论分析选择如何，这种正相关关系都十分稳健。

生产力的跃升可能源于在多项研究任务中使用生成式人工智能，包括创意生成、文献发现、编程、数据收集或分析。但截至目前，大语言模型可能对写作产生了最大的影响（*24, 25*）。为了创作具有独特性的科学作品，研究人员必须提出有说服力的书面论证；将手稿中的论点、方法和结果与既有文献联系起来；详细阐述并置于语境中解读最重要的发现；并清晰说明可以从文本中学到什么。这些复杂的写作任务耗时甚久，尤其对于使用非母语进行交流的研究人员而言。因此，我们提出以下问题：大语言模型的采用对生产力的影响是否会因作者的母语熟练度而异？由于大多数高影响力研究都发表在英文期刊和会议论文集中，母语者在科学传播方面一直享有显著优势（*26-28*）。大语言模型可以缓解英语流利度方面的差异，从而应当不对称地降低不同科学家语言背景下的写作成本（*29*）。

为检验处理效应的异质性，我们根据作者姓名及其所属机构，估算作者为英语母语者的可能性（SM S2.5-2.6）。图2按研究人员的族裔和所在地区分别展示了系数。所有群体的效应在统计上仍然显著，但拥有亚洲姓名的学者从采用大语言模型中获得的生产力提升最大。在 bioRxiv 和 SSRN 上，拥有亚洲姓名且机构隶属关系位于亚洲的学者所受影响更为显著；但与在美国／英国／加拿大／澳大利亚机构任职的亚洲学者之间的差异，仅在 bioRxiv 上达到统计显著。估计显示，大语言模型的采用对这组学者科学产出的影响从 arXiv 上的最低 43.0%，到 bioRxiv 上的 89.3% 和 SSRN 上的 88.9% 不等。拥有高加索裔姓名、隶属于英语国家机构的研究人员，生产力提升幅度较为温和，但仍显著，介于 23.7%（arXiv）至 46.2%（SSRN）之间。

我们得出结论，即便使用上一代大语言模型——也就是我们数据中的手稿起草时学者能够使用的那些模型——也与生产力提升相关，尤其是对面临较高写作成本的研究人员而言。这些发现与已有研究相一致：大语言模型能够缓解技能差距的影响，在本例中，它们降低了使用第二语言写作的成本（*30*）。鉴于当前一代大语言模型的写作能力已大幅提升，且这些系统的可及性更加广泛，我们估算的生产力效应很可能大到足以推动科学产出的市场份额向非英语母语地区的学者转移。

大语言模型使用、科学写作与发表结果

大语言模型可能会在生产力效应之外重塑科学产出。高质量写作通常被视为科学价值的体现。语言清晰而复杂的论文往往被认为质量更高，引用也更频繁（*31*）。科学进展源于多年知识积累与打磨，因此，准确阐述科学发现的能力，可以作为科研团队工作是否用心的一个（很不完美的）替代指标。大语言模型几乎可以毫不费力地生成润色精良、专业规范且描述任何科学主题的文本，这引出了一个重要问题：大语言模型使用究竟能揭示还是掩盖底层研究的质量?

为评估这一问题，我们研究了写作复杂度与研究质量的关系，以及大语言模型的采用是否会改变写作复杂度在科学传播中的信号作用。我们采用弗莱施阅读易读性评分的相反数来衡量写作复杂度（SM S2.7）。这一指标通过平均句子长度和每词音节数的综合值来量化文本复杂度，分数越高表示文本越复杂。随后，我们将观察期结束时（2024年6月）是否发表于同行评审期刊或会议作为质量的替代指标，并据此为自2023年以来的所有预印本创建二元结果变量（SM S2.8、S5.9）。

将弗莱施评分的相反数与发表结果相关联时，我们发现了三种规律。首先，在全部三个档案库中，大语言模型辅助撰写的稿件，其写作复杂度评分显著高于以自然语言撰写的论文（$P < 0.001$，所有存储库，双尾t检验）（图3A-C）。这凸显了大语言模型生成复杂科学写作内容的非凡能力（*23*）。其次，在全部三个存储库中，非大语言模型辅助撰写的论文的写作复杂度与稿件质量呈正相关；稿件质量以发表于同行评审出版平台的概率作为近似指标（逻辑回归，图3E-G）。这些结果证实了先前研究所发现的写作复杂度与科学价值之间的正相关关系（*32*）。第三，也是至关重要的一点，我们发现，大语言模型辅助撰写的稿件中，写作复杂度与同行评审结果之间的关系出现了逆转。对于这些文档，写作复杂度提高与同行对其科学价值的评价降低相关（图3E-G）。

为评估这些发现的稳健性，我们考察了写作中其他词汇、句法和形态特征（SM S5.10）。我们使用词汇复杂度（每词音节数）和形态复杂度（现在分词从句的占比）重复了这些分析，结果与原发现一致：在大语言模型辅助撰写的论文中，写作复杂度提高与发表成功呈负相关；而在人类撰写的论文中则呈正相关。我们还发现，宣传性语言的使用也呈现相同模式，进一步证实了大语言模型的采用会从多个语言维度削弱传统质量信号（*33*）。

有许多因素会影响预印本的发表结果。虽然我们无法排除所有混杂因素，但在控制预印本的月份和研究领域后，结果依然一致（SM S3.2, S5.9）。作为稳健性检验，我们收集并分析了来自国际学习表征会议（ICLR-2024）的独立数据集。该会议是机器学习领域的顶尖会议，

机器学习（*34*）。无论最终是否被会议录用，ICLR-2024均提供该会议全部7,243篇投稿的28K份审稿报告（SM S1.4）。我们以专家给出的同行评审分数作为科学价值的替代衡量指标，图3D和3H以显著的一致性重现了关键发现。

两组——人类撰写的手稿与大语言模型辅助撰写的手稿——在语言复杂度分布上的质量评估存在鲜明反差，这证实了复杂的大语言模型生成语言往往会掩饰薄弱的科学贡献（*35*）。这些发现表明，一种传统启发式方法正在迅速失效。对于大语言模型辅助撰写的手稿，语言复杂度与科学价值之间的正相关不仅消失了，甚至发生了逆转。随着产出润色精美 prose 所需的努力降低，其作为作者掌握某一主题能力信号的效用也随之下降（*36*）。这给科学事业带来了风险，因为大量表面上令人信服、但科学上乏善可陈的研究可能充斥文献。如果这种情况发生，它将掩埋一些重要思想，并迫使科学界浪费宝贵时间，从一片无关紧要的研究泥淖中辨别真正的洞见。

对于同行评审者和期刊编辑而言，这代表着一个重大问题。随着科学传播的数量激增，写作特征作为一种（不完美地）筛选科学研究的捷径，正迅速变成缺乏信息价值的信号。随着传统启发式方法失效，编辑和评审者可能越来越依赖作者背景和机构隶属关系等身份标志来判断质量，这反而抵消了大语言模型对科学产出的民主化影响。一种潜在的应对方式，是利用同一项技术协助评估手稿。专业的“审稿代理”可以标记方法学不一致、核验论断，甚至评估新颖性。这种可扩展的方法究竟能否帮助编辑和评审者超越表层信号、专注于实质内容，还是会给科学流程引入新的、未预见的挑战，仍是一个关键的不确定因素。

大语言模型使用与引用行为

撰写科学论文还涉及将论点和研究发现置于现有文献的脉络中。由于大语言模型能够摄取并综合海量信息，它们可能会拓宽研究人员对既有研究的了解（*37, 38*）。或者，正如一些人所推测的那样，训练数据可能对高影响力作品有所偏重，进而使大语言模型扩大研究人员对易于检索的研究的接触（*39*）。因此，我们提出这样的问题：大语言模型如何影响既有文献的发现?

为了评估这些相互竞争的假设，我们利用一个独特的数据集，其中记录了 arXiv论文的 2.46亿次浏览和下载（SM S1.5），每次记录均关联一个用户 ID、arXiv文档 ID 和引荐来源（例如必应或Google）。借助该数据集，我们可以探究 Bing Chat（GPT-4）于 2023年2月发布后，用户层面的阅读行为变化。Bing Chat 是首个得到广泛采用的人工智能驱动的搜索工具（SM S3.4）。图4A-C 比较了这一外生变化前后必应用户访问的 arXiv文档。基于双重差分分析的估算结果显示，与由Google引流的访问相比，Bing Chat 发布后，用户访问的 arXiv文档种类更加多样。图4A比较了出版形式，显示

必应用户访问书籍的比例高出26.3%（$P < 0.001$，泊松回归），这可能反映了大语言模型能够从篇幅较长的文本中呈现内容。
获取图书的渠道增多，表明借助大语言模型开展的科学研究可能会利用更广泛的参考资料，但这并不能排除大语言模型只是强化了人们对科学经典的关注。我们探究了这种可能性，发现必应引荐访问也与较新的学术成果相关；所访问手稿的中位年代估计减少了0.18年（图4B）。此外，大语言模型用户并未增加对高被引作品的引用（图4C）。相反，我们发现，必应用户找到了既有引用次数较少的参考文献。

为考察搜索结果的这一变化是否转化为实际引用行为的改变，我们将 arXiv、bioRxiv 和 SSRN 上的预印本与两个大型引文数据库 OpenAlex 和 Semantic Scholar 关联起来。我们获得了对既有研究的1.016亿条引文（SM S1.6）。随后，我们采用图1中的事件研究方法，比较作者使用大语言模型前后的引用行为，并以未使用者作为对照组（SM S3.3）。我们的分析考察了被引参考文献的三个特征：（1）出版形式（对图书的引用）；（2）时间滞后（参考文献年代的中位数）；以及（3）被引作品影响力（被引文献的平均对数引文数）。

与图 4A-C 一致，我们发现大语言模型使用会改变作者的引用行为，似乎引导他们采用更多元的知识基础。我们发现，总体而言，大语言模型的采用者引用书籍的可能性高出 11.9%（图 4D），但在其中一个档案库 SSRN 中，这一影响在统计上并不显著。采用者引用的文献平均也新 0.379 年（图 4E），且没有证据表明这些文献的影响力更大（引用影响力低 2.34%，图 4F）。尽管这些影响的幅度因预印本存储库而异，但总体模式大体一致（图 4G–I）。

图 4 提供了一致证据，表明人工智能辅助会引导学者接触更广泛的知识。研究人员面临时间和注意力方面的限制，难以处理不断扩展的研究领域（*40, 41*）。大语言模型似乎能帮助研究人员克服发现相关文献的障碍。

这些发现表明，虽然大语言模型可能会掩盖作者付出的努力，但似乎能让知识发现的路径更加清晰。人们普遍担心，人工智能辅助搜索可能会强化既有的科学经典。然而，我们发现，大语言模型的采用产生了相反的效果。无论是人工智能辅助搜索行为，还是作者的引文模式，都显示出研究者正显著转向更加多元的知识基础，其中包括更多书籍，以及较新的、引用较少的学术成果。这种关注范围的扩大表明，大语言模型有助于研究者克服认知限制，而这些限制一直妨碍着他们涉猎不断扩展的科学文献。

局限性与未来方向

本研究探讨了大语言模型对科学产出的影响，但我们的研究结果受到若干局限因素的制约，这些局限也为未来研究提供了方向。

首先，我们未进行因果识别。研究现实环境中的大语言模型时，一个核心难题在于无法完美地衡量其使用情况。我们的人工智能检测方法并不完美，容易受到多项挑战的影响（SM S5.1-5.4）：该方法依赖摘要而非全文（SM S5.5）；无法明确识别团队中的哪位具体共同作者使用了大语言模型（SM S5.6）；而且，对于那些大幅编辑大语言模型辅助文本的作者，该方法几乎肯定无法检测到其使用情况。此外，生成式人工智能工具的非随机采用带来了自我选择偏差的可能性，而我们将重点放在已发表论文上，意味着“采用时间”可能与生产力存在内生性。我们的补充材料包含许多额外分析，以评估这些问题的影响范围。尽管我们的结果似乎具有稳健性，但未来研究仍应继续探索应对这些挑战的方法论策略，这一点十分重要。

其次，我们的发现只是对一项快速演进技术的阶段性观察。我们的分析基于更先进的推理模型和深度研究能力问世前生成的数据。随着模型不断改进，科学家也不断探索将其融入工作的新方式，这些技术未来产生的影响很可能远超我们在此强调的效果。这为未来研究指出了一个关键方向：持续追踪科学事业如何纳入一代又一代的人工智能模型。随着这些更强大的工具融入科学工作流，研究需要考察我们记录的影响是会增强、改变，还是甚至逆转。

未来研究有许多方向。其中一个主要方向是更细致地探讨大语言模型如何影响科学实践。长期以来，科学进步受到获取非正式资源和知识的限制——Diana Crane 将其称为“隐形学院”（*42*）。一种假设认为，大语言模型可以大规模替代这类非正式知识，提供从实验设计到了解某一领域隐性课程等各方面的指导，从而让科学竞争环境更加公平。未来研究的另一个有趣方向，是大语言模型突破学科边界的潜力。随着时间推移，各学术学科形成了深厚的知识库，而这些知识往往通过学科专属行话来传达。如果大语言模型能帮助外行克服这一障碍，孤岛化学科之间或许就能开展更富成效的交流。

结论

我们的研究结果表明，大语言模型已开始重塑科学产出。使用大语言模型加快了稿件产出，降低了非英语母语者的障碍，并拓宽了对既有文献的发现范围。这些变化可能使科学产出更加民主化。然而，科学质量的传统信号（如语言复杂度）正逐渐成为不可靠的价值判断指标，而科学工作的数量却在不断增长。这些变化预示着研究格局将持续演变：英语流利度的重要性会逐渐减弱，但健全的质量评估框架和深入的方法论审查将至关重要。随着人工智能系统不断进步，它们将挑战我们对研究质量、学术交流以及智力劳动本质的基本假设。一项紧迫的议程

对于科学政策制定者而言，将是推动科学机构适应快速演变的科学生产过程。

参考文献与注释

1. N. Rasmussen，《图像控制：电子显微镜与美国生物学的变革，1940—1960年》。（斯坦福大学出版社，1999年）。2. W. W. Ding、S. G. Levin、P. E. Stephan、A. E. Winkler，信息技术对学术科学家生产力与合作模式的影响。《管理科学》**56**，1439-1461（2010年）。3. J. L. Furman、F. Teodoridis，自动化、研究技术与研究人员的发展轨迹：来自计算机科学和电气工程领域的证据。《组织科学》**31**，330-354（2020年）。4. D. C. Koboldt、K. M. Steinberg、D. E. Larson、R. K. Wilson、E. R. Mardis，下一代测序革命及其对基因组学的影响。《细胞》**155**，27-38（2013年）。5. W. Liang等，量化科学论文中大语言模型的使用情况。《自然·人类行为》，（2025年）。6. H. Bao、M. Sun、M. Teplitskiy，有志者事竟成：在禁止使用 ChatGPT 的国家，人们反而更多地将其用于科学研究。*arXiv*预印本 *arXiv:2406.11583*，（2024年）。7. D. Kobak、R. González-Márquez、E.-Á. Horvát、J. Lause，通过过度词汇深入探究大语言模型辅助写作在生物医学出版物中的应用。《科学进展》**11**，eadt3813（2025年）。8. W. Zhu、L. W. Cong，研究写作中的大语言模型采用差异与异质性趋同路径。康奈尔大学*SC*约翰逊商学院研究论文（即将发表），（2024年）。9. H. Wang等，人工智能时代的科学发现。《自然》**620**，47-60（2023年）。10. Q. Hao、F. Xu、Y. Li、J. Evans，人工智能扩大科学家的影响力，却使科学的关注重点收窄。*arXiv*预印本 *arXiv:2412.07727*，（2024年）。11. J. Gao、D. Wang，量化人工智能在科学研究中的应用及潜在益处。《自然·人类行为》，1-12（2024年）。12. E. Callaway，人工智能蛋白质折叠革命的下一步是什么？《自然》**604**，234-238（2022年）。13. A. Merchant等，扩展深度学习以推动材料发现。《自然》**624**，80-85（2023年）。14. K. Swanson、W. Wu、N. L. Bulaong、J. E. Pak、J. Zou，人工智能智能体虚拟实验室设计出新型 SARS-CoV-2 纳米抗体。《自然》，（2025年）。15. B. Romera-Paredes等，利用大语言模型进行程序搜索，推动数学发现。《自然》**625**，468-475（2024年）。16. C. Si、D. Yang、T. Hashimoto，大语言模型能否生成新颖研究构想？一项针对100+ 自然语言处理研究人员的大规模人类研究。*arXiv*预印本 *arXiv:2409.04109*，（2024年）。17. J. Kim、B. Lee，人工智能增强型调查：借助大语言模型和调查进行意见预测。*arXiv*预印本 *arXiv:2305.09620*，（2023年）。18. M. Binz等，用于预测和捕捉人类认知的基础模型。《自然》，1-8（2025年）。19. S. Lai、J. Kim、N. Kunievsky、Y. Potter、J. Evans，基于价值观的人工智能改善人类决策，却降低信任。*arXiv*预印本 *arXiv:2508.09297*，（2025年）。

20. C. Xie 等，大语言模型智能体能否模拟人类信任行为？《神经信息处理系统进展》37，15674-15729（2024年）。21. E. Zhou、D. Lee，生成式人工智能、人类创造力与艺术。PNAS Nexus 3，pgae052（2024年）。22. S. Peng、E. Kalliamvakou、P. Cihon、M. Demirer，人工智能对开发者生产力的影响：来自 GitHub Copilot 的证据。arXiv 预印本 arXiv:2302.06590，（2023年）。23. S. Noy、W. Zhang，生成式人工智能生产力效应的实验性证据。《科学》381，187-192（2023年）。24. C. Bazerman，《塑造书面知识：科学中实验性论文的体裁与活动》。（威斯康星大学出版社，1988年）。25. D. N. McCloskey，《经济学修辞学》。《经济文献杂志》21，481-517（1983年）。26. D. I. Hanauer、K. Englander，量化以第二语言撰写研究论文的知识负担：来自墨西哥科学家的数据。《书面交流》28，403-416（2011年）。27. V. Ramírez-Castañeda，英语在科学领域占主导地位所造成的科学论文撰写与发表劣势：哥伦比亚生物科学研究人员的案例。《公共科学图书馆·综合》15，e0238372（2020年）。28. D. D. Belcher，在仅使用英语的研究世界中寻求认可。《第二语言写作杂志》16，1-22（2007年）。29. V. Berdejo-Espinola、T. Amano，人工智能工具有助于促进科学领域的公平。《科学》379，991 991 2023-（）。30. E. Brynjolfsson、D. Li、L. Raymond，工作中的生成式人工智能。《经济学季刊》140，889-942（2025年）。31. H. Hu、D. Wang、S. Deng，科学文献摘要写作风格与引文分析。《在线信息评论》45，1290-1305（2021年）。32. J. Feld、C. Lines、L. Ross，写作至关重要。《经济行为与组织杂志》217，378-397（2024年）。33. H. Peng、H. S. Qiu、H. B. Fosse、B. Uzzi，宣传性语言与创新理念在科学领域的采用。《美国国家科学院院刊》121，e2320066121（2024年）。34. W. Liang 等，大语言模型能否为研究论文提供有用的反馈？一项大规模实证分析。NEJM AI 1，AIoa2400196（2024年）。35. E. A. M. Van Dis、J. Bollen、W. Zuidema、R. Van Rooij、C. L. Bockting，ChatGPT：研究领域的五项优先事项。《自然》614，224-226（2023年）。36. Z. Wojtowicz、S. DeDeo，削弱思维证明：人工智能如何通过降低思考难度加大合作难度。《人工智能促进协会人工智能会议》39，1592-1600（2025年）。37. Y. Tian、Y. Liu、Y. Bu、J. Liu，谁会获得推荐？调查大语言模型论文推荐中的性别、种族和国家差异。arXiv 预印本 arXiv:2501.00367，（2024年）。38. B. C. Lee、J. Chung，ChatGPT 对创造力影响的实证调查。《自然·人类行为》8，1906-1914（2024年）。39. A. Algaba 等，大语言模型反映人类引文模式，但引文偏差更为明显。arXiv 预印本 arXiv:2405.15739，（2024年）。40. B. F. Jones，知识负担与“文艺复兴全才之死”：创新是否越来越难？《经济研究评论》76，283-317（2009年）。

41. A. Zeng 等，《科学的科学：复杂系统视角》。《物理学报告》 **714**，1-73（2017）。
42. D. Crane，《隐形学院》；科学共同体中的知识扩散。（1972）。

致谢： 感谢 Mor Naaman、Will Cong、Wu Zhu，以及复杂性科学中心（维也纳）、加州大学洛杉矶分校普莱斯中心、哈斯 MRL 研讨会和哥伦比亚大学 MAD 会议的研讨会参与者提供有益讨论，并感谢 Alex Cui 提供 GPTZero API 的学术访问权限。本研究得到美国国家科学基金会资助，资助编号为 2311521、2404035 和 2412389。

补充材料

补充文本 图 S1 至
S37 表 S1 至 S10

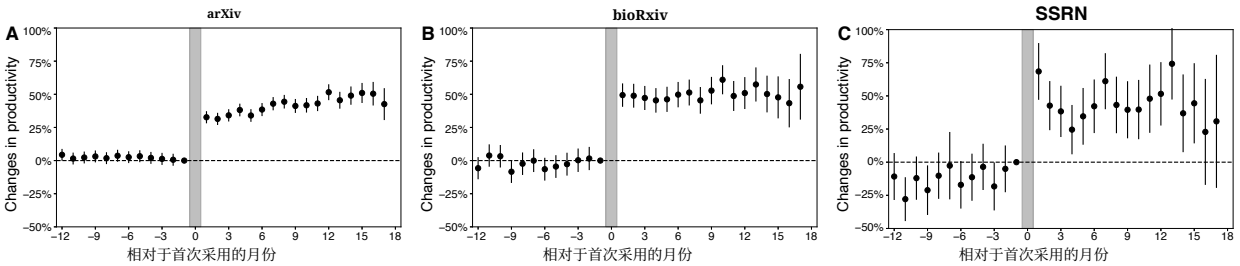


图1。大语言模型使用与科学生产力。我们追踪了2022年1月至2024年7月期间168,553名作者的生产力动态（以每月发表的预印本数量衡量），这些作者分布于 arXiv（109,965名作者）、bioRxiv（43,218名作者）和 SSRN（15,370名作者）。对于每位作者，我们使用基于文本的检测器分析其预印本，以确定他们是否以及何时在科学写作中“采用”了大语言模型。（**A-C**）通过堆叠式双重差分回归，我们估计大语言模型对个人生产力的影响。在研究者层面对比首次采用前后的情况后，我们保守地观察到生产力显著提高：相较于未采用者，arXiv、bioRxiv 和 SSRN 的生产力分别提升了36.2%、52.9%和59.8%。

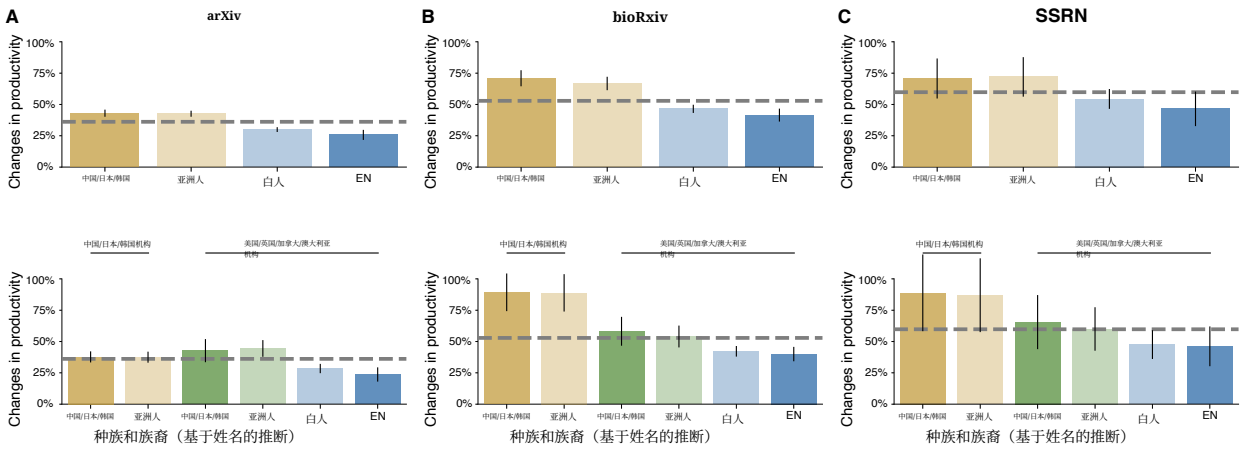


图2. 不同种族、族裔和所在地区之间的异质性。(A-C) 对于拥有东亚姓名的作者，生产力提升尤为显著，增幅分别为 43.0%（arXiv）、70.9%（bioRxiv）和 70.1%（SSRN）；对于拥有高加索裔姓名的作者，提升幅度也很可观，分别为 25.7%（arXiv）、41.5%（bioRxiv）和 46.5%（SSRN）。对于 亚洲姓名 且所属机构位于亚洲的作者，效果最为显著，生产力分别提升 37.7%（arXiv）、89.3%（bioRxiv）和 88.9%（SSRN）。相比之下，拥有高加索裔姓名且所属机构位于英语国家的作者，生产力提升幅度分别为 23.7%（arXiv）、40.0%（bioRxiv）和 46.2%（SSRN）。灰色虚线表示所有作者的平均效应。

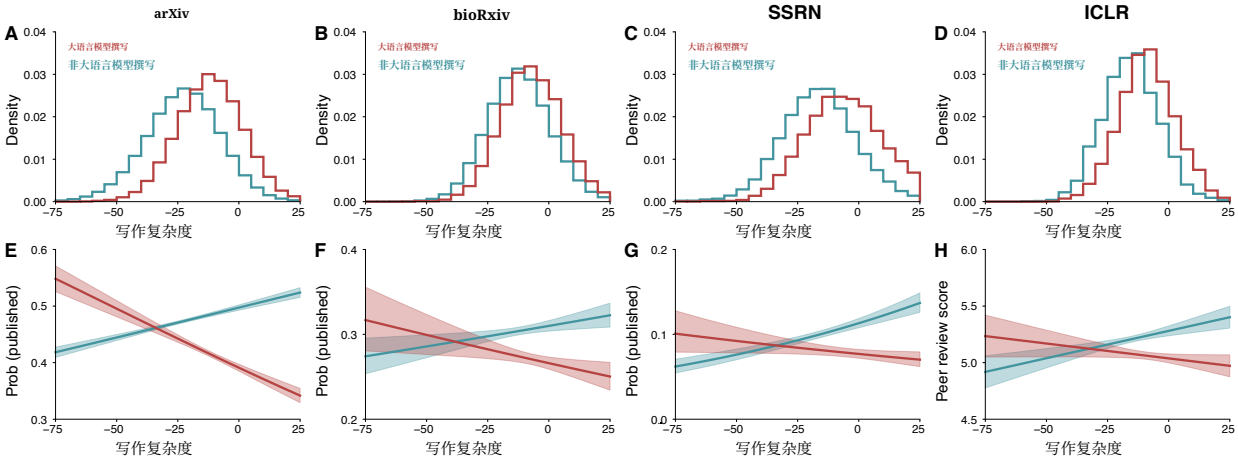


图 3. 大语言模型使用、科学写作与发表结果。 针对自 2023年 以来撰写的 264,125 篇手稿（arXiv 上 177,880 篇、bioRxiv 上 58,958 篇、SSRN 上 31,959 篇，以及提交至 ICLR 2024 的 7,243 篇）。我们将写作复杂度定义为弗莱施阅读易读性评分的相反数；该指标是文本平均句长和词长的联合函数。我们还通过以下指标衡量论文质量：（i）预印本（数据截止日期之前）经同行评审发表的概率；（ii）ICLR投稿的同行评审评分。**(A-D)** 大语言模型辅助撰写的手稿（红色）和非大语言模型辅助稿件（蓝色）的写作复杂度分布。在所有四个数据集中，大语言模型辅助撰写的手稿都表现出显著更高的写作复杂度。**(E-H)**写作复杂度与论文质量之间的关系。对于非大语言模型辅助稿件，写作复杂度与感知稿件质量的衡量指标呈正相关。与此形成鲜明对比的是，对于大语言模型辅助撰写的手稿，写作复杂度越高，稿件质量越低。预测结果分别基于逻辑回归（**E-G**中的发表概率）和普通最小二乘回归（**H**中的同行评审评分）。

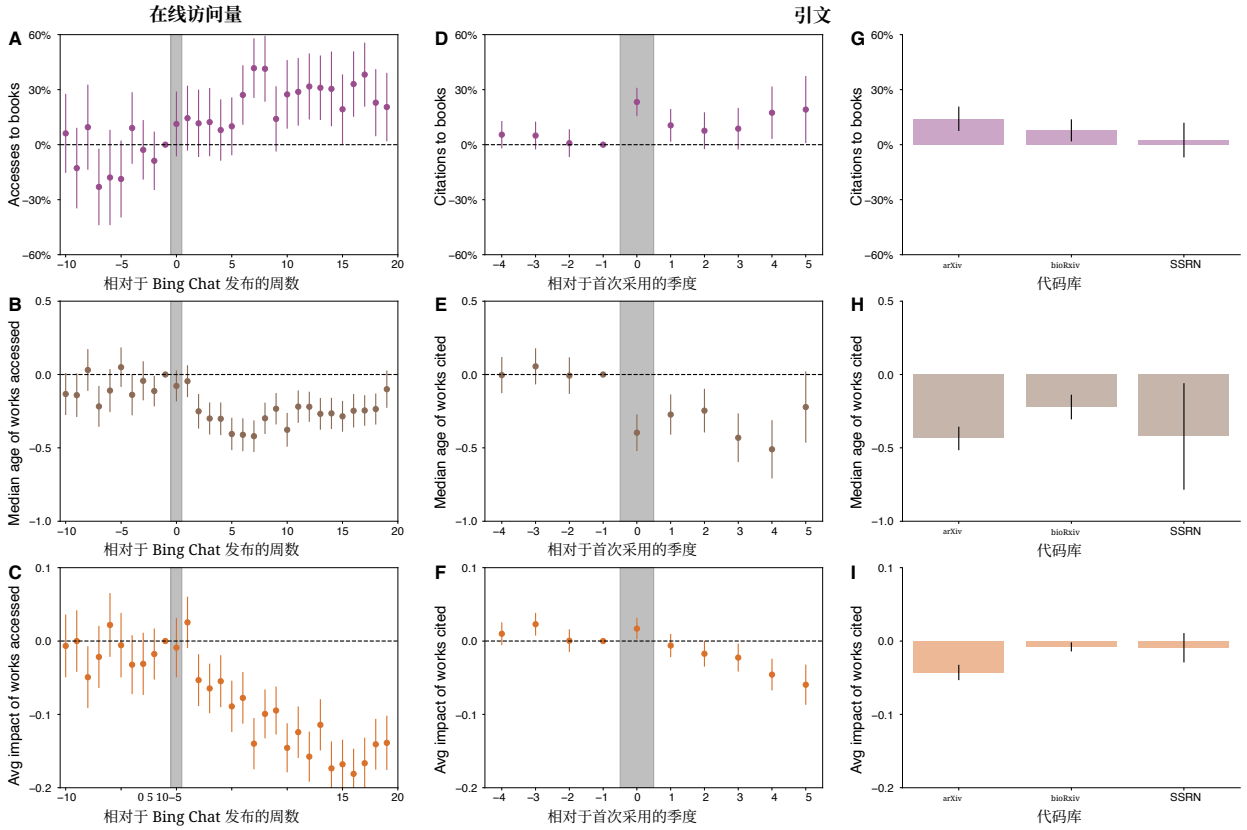


图4. 大语言模型使用与对既有研究的引用。（A-C）我们考察了2023年2月Bing Chat（由GPT-4驱动）发布后，用户访问arXiv手稿的变化。比较从Google和必应跳转而来的在线访问量后，我们发现，事件发生后，用户访问**(A)** 更多 书籍，**（B）**年代较近的研究，以及**（C）**被引次数较少的研究。**（D-I）**我们采用图1所示的双重差分策略，考察作者首次采用大语言模型后，其引用和引文模式的变化。由于数据稀疏，所有结果均按季度测量。**（D）**采用大语言模型的作者在采用后引用的书籍增加了11.9%，这体现了大语言模型处理和整合更多样化知识来源的先进能力。估计结果来自泊松回归。**（E）**采用大语言模型的作者在采用后引用的参考文献年代更近，中位参考文献年代减少了0.379年。**（F）**尽管有人担心大语言模型可能会加剧对公认科学成果的依赖，但我们发现，对高影响力研究的引用并未增加；其中，“参考文献影响力”是指所有被引参考文献的log(#citations+1)平均值。**（G）**在arXiv、bioRxiv和SSRN上使用大语言模型的作者，在采用后引用的书籍分别增加了14.1%（$P < 0.001$）、7.81%（$P = 0.011$）和2.50%（$P = 0.605$）。**（H）**在arXiv、bioRxiv和SSRN上使用大语言模型的作者，引用的参考文献年代分别近了0.436年（$P < 0.001$）、0.222年（$P < 0.001$）和0.422年（$P = 0.023$）。**（I）**在arXiv、bioRxiv和SSRN上使用大语言模型的作者，引用的高被引参考文献分别减少了4.29%（$P < 0.001$）、0.800%（$P = 0.012$）和0.916%（$P = 0.369$）。