POLICY ARTICLE

ARTIFICIAL INTELLIGENCE

Scientific production in the era of large language models

With the production process rapidly evolving, science policy must consider how institutions could evolve

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

Despite growing excitement (and concern) about the fast adoption of generative artificial intelligence (Gen AI) across all academic disciplines, empirical evidence remains fragmented, and systematic understanding of the impact of large language models (LLMs) across scientific domains is limited. We analyzed large-scale data from three major preprint repositories to show that the use of LLMs accelerates manuscript output, reduces barriers for non-native English speakers, and diversifies the discovery of prior literatures. However, traditional signals of scientific quality such as language complexity are becoming unreliable indicators of merit, just as we are experiencing an upswing in the quantity of scientific work. As AI systems advance, they will challenge our fundamental assumptions about research quality, scholarly communication, and the nature of intellectual labor. Science policy-makers must consider how to evolve our scientific institutions to accommodate the rapidly changing scientific production process.

...LLM adoption is associated with a large increase in researchers' scientific output...

The scientific enterprise is intimately connected with technological innovation. The microscope, advances in computing, and next-generation sequencers, for example, shifted the frontier of research. Researchers have demonstrated the value of AI in many specific scientific contexts (*1*, *2*), such as protein-structure prediction and materials discovery. Recent advancements in LLMs have expanded their use across a wide range of tasks in the natural (*3*) and social sciences (*4*). This work highlights the incredible potential of LLMs across specific scientific undertakings, raising an open question: What is the macro-level impact of LLMs on the scientific enterprise?

To address this question, we collected large-scale data from three preprint repositories [spanning January 2018 to June 2024, see supplementary materials (SM) S1.1 to S1.3 for details]: arXiv (1.2 million preprints), which includes mathematics, physics, computer science, electrical engineering, quantitative biology, statistics, and economics; bioRxiv (221,000 preprints), which spans a wide range of subfields in biology and the life sciences; and Social Science Research Network (SSRN; 676,000 preprints), a working-paper repository that hosts manuscripts in the social sciences, law, and the humanities. Each of the three datasets represents the largest within its domain. Collectively, they offer an unprecedented empirical basis to examine some of the impacts of LLMs on scientific productivity practices across many scientific fields.

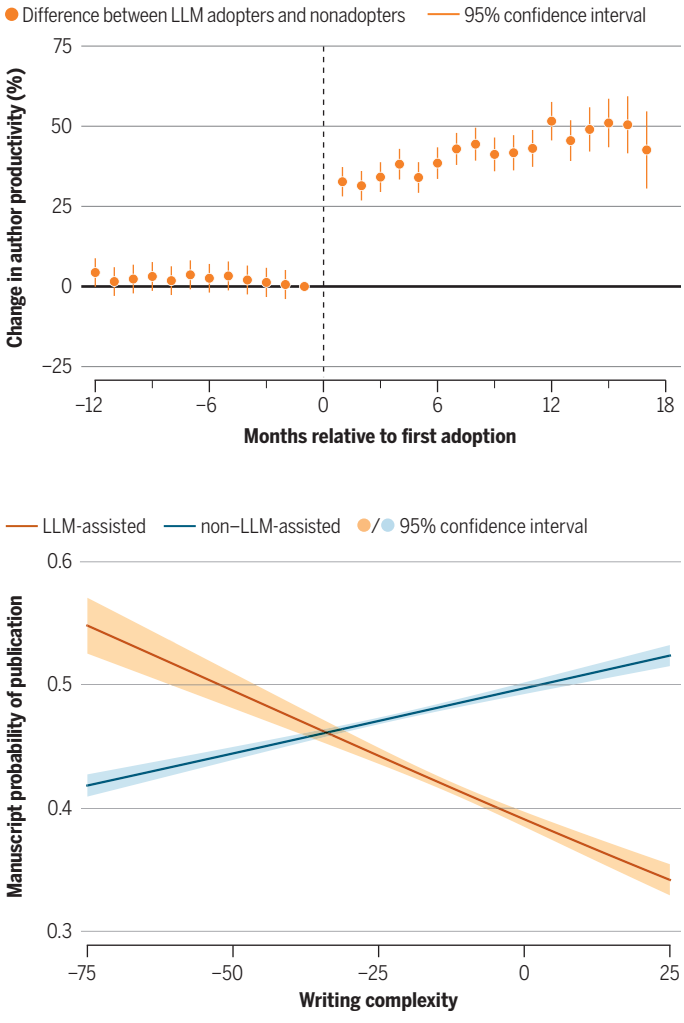
To identify the use of LLMs in the creation of scientific manuscripts, we applied a text-based AI detection algorithm (*5*) to all abstracts in our data. We used abstracts from papers submitted prior to 2023—before the ChatGPT era—to estimate the token (word) distribution of human-written text. We then prompted OpenAI's GPT-3.5turbo0125 model to rewrite these abstracts, generated the token distribution of LLM-written text, and compared the two. This allowed us to quantify differences in word distributions between LLM-assisted and human writing and identify probable LLM-assisted abstracts written after the release of ChatGPT. Further details on model training, validation, potential limitations, and alternative methods of LLM detection are provided (see SM S2.1, S4, and S5).

LLM USAGE AND SCIENTIFIC PRODUCTIVITY

We predicted that authors who adopted LLMs would experience increased productivity (*6*, *7*). To isolate the general productivity effects of LLMs from rapid growth in research on AI, we first exclude manuscripts in core AI subdisciplines (see SM S1.1 and S5.7) from our sample. We then identified an author's initial adoption of LLMs as marked by the first manuscript (m_i) that exhibited statistical signatures of LLM assistance (α), such that $\alpha(m_i) > \tau$, where τ is the detection threshold. An author's adoption status changes from 0 to 1 for any months occurring after the first detected use. Based on this measure, we examined the change in manuscript submission rates between LLM adopters and similar non-adopters (see SM S2.2 to S2.4) before and after adoption in author-level fixed-effects event models (see SM S3.1).

Productivity and publication

Between January 2022 and July 2024, the number of arXiv preprints published monthly once an author had adopted LLMs in their writing increased by 36.2% relative to nonadopters (top). Since 2023, for LLM-assisted manuscripts, a greater writing complexity of arXiv manuscripts is correlated with a lower probability of being published. The relationship is inverted for non–LLM-assisted manuscripts (bottom).



We show that LLM adoption is associated with a large increase in researchers' scientific output in all three preprint repositories. The estimated coefficients for arXiv (see the first figure, top), bioRxiv, and SSRN (see fig. S1) are 36.2, 52.9, and 59.8%, respectively, suggesting that LLM use is associated with sizable increases in productivity. Although estimated coefficients vary by the detection method and threshold used to identify LLM adoption, sensitivity analyses (see SM S5.3 to S5.6) demonstrate that a positive association is robust across analytical choices.

A productivity jump may stem from the use of Gen AI across multiple research tasks, including idea generation, literature discovery, coding, data collection, or analysis. But to date, LLMs likely have had the largest impact in writing. To create distinctive scientific works, researchers must present compelling written arguments; link a manuscript's arguments, methods, and results to prior literature; detail and contextualize the most important findings; and articulate what can be learned from the text. These complex writing tasks are time consuming, particularly for researchers who are communicating in a non-native language. We therefore ask whether the productivity impact of LLM adoption varies across authors' native-language proficiencies. Because most high-impact research is published in English-language journals and proceedings, native speakers have had a substantial advantage in scientific communication. LLMs can mitigate disparities in English fluency, which should asymmetrically reduce the cost of writing across scientists' linguistic backgrounds.

To test for heterogeneity in productivity changes, we approximated the likelihood that an author is a native English speaker based on names and the institutions with which they are affiliated (see SM S2.5 to S2.6). Coefficients were broken out by researchers' ethnicities and home geographies (see fig. S2). The effects remain statistically significant across all groups, but scholars with Asian names experienced the greatest productivity boost from LLM adoption. In bioRxiv and SSRN, effects were even more pronounced for scholars with Asian names and institutional affiliations in Asia, with bioRxiv showing a statistically significant additional productivity gain for Asian-named scholars in Asian institutions (relative to those in US, UK, Canadian, and Australian institutions). For Asian-named scholars affiliated with Asian institutions, the estimated LLM-related productivity gain ranged from a low of 43.0% in arXiv to 89.3% for bioRxiv and 88.9% for SSRN. Researchers with Caucasian names affiliated with institutions in English-speaking countries experienced more modest but still significant productivity gains of 23.7% (arXiv) to 46.2% (SSRN).

We conclude that even the use of previous-generation LLMs—those available to scholars at the time the manuscripts in our dataset were drafted—are associated with productivity gains, particularly for researchers facing higher costs of writing. These findings concur with work showing that LLMs mitigate the impact of skill disparities, in this case, by reducing the cost of writing in a second language (*8*). Given considerable advances in the writing ability of present-generation LLMs and the more widespread availability of these systems, the productivity effects that we estimated are likely substantial enough to imply a shift in the market share of scientific production toward scholars in non-native English–speaking geographies.

LLM USE, SCIENTIFIC WRITING, AND PUBLICATION OUTCOMES

LLMs are likely to reshape science production beyond productivity effects. High-quality writing is often construed as a signal of scientific merit (*9*). Papers with clear but complex language are perceived to be stronger and are cited more frequently. Because scientific advances are the product of years of knowledge refinement, the ability to precisely articulate scientific discoveries is a (very imperfect) proxy for the care taken during a scientific team's work. The fact that LLMs can almost effortlessly produce polished, professional text describing any scientific topic raises an important question: Does LLM use reveal or conceal the quality of the underlying research?

To assess this question, we investigated how writing complexity relates to research quality and whether LLM adoption changes the signaling power of writing complexity in scientific communication. We gauged writing complexity with the additive inverse of the Flesch Reading Ease score (see SM S2.7). This measure quantifies text complexity as a composite of average sentence length and syllables per word, with higher scores indicating more complex text. As a proxy for quality, we then created a binary outcome defined as publication in a peer-reviewed journal or conference by the end of our observation window (June 2024) for all preprints since 2023 (see SM S2.8 and S5.9).

When we correlated the additive inverse of the Flesch score with publication outcomes, three patterns emerged. First, writing-complexity scores in LLM-assisted manuscripts were significantly higher compared with papers written in natural language in all three archives ($P < 0.001$, all repositories, two-tailed t test) (see fig. S3, A to C). This underscores the remarkable capability of LLMs to produce complex scientific writing (*7*). Second, in non–LLM-assisted papers

across all three repositories, writing complexity was positively associated with manuscript quality, as approximated by the probability of publication in a peer-reviewed venue (logistic regressions; see the first figure, bottom, and fig. S3). These results confirm prior research that showed a positive association between writing complexity and scientific merit. Third, and critically, we found a reversal in the relationship between writing complexity and peer-review outcomes for LLM-assisted manuscripts. For these documents, increases in writing complexity were associated with lower peer assessments of scientific merit (see the first figure, bottom; fig. S3; and SM S3.2).

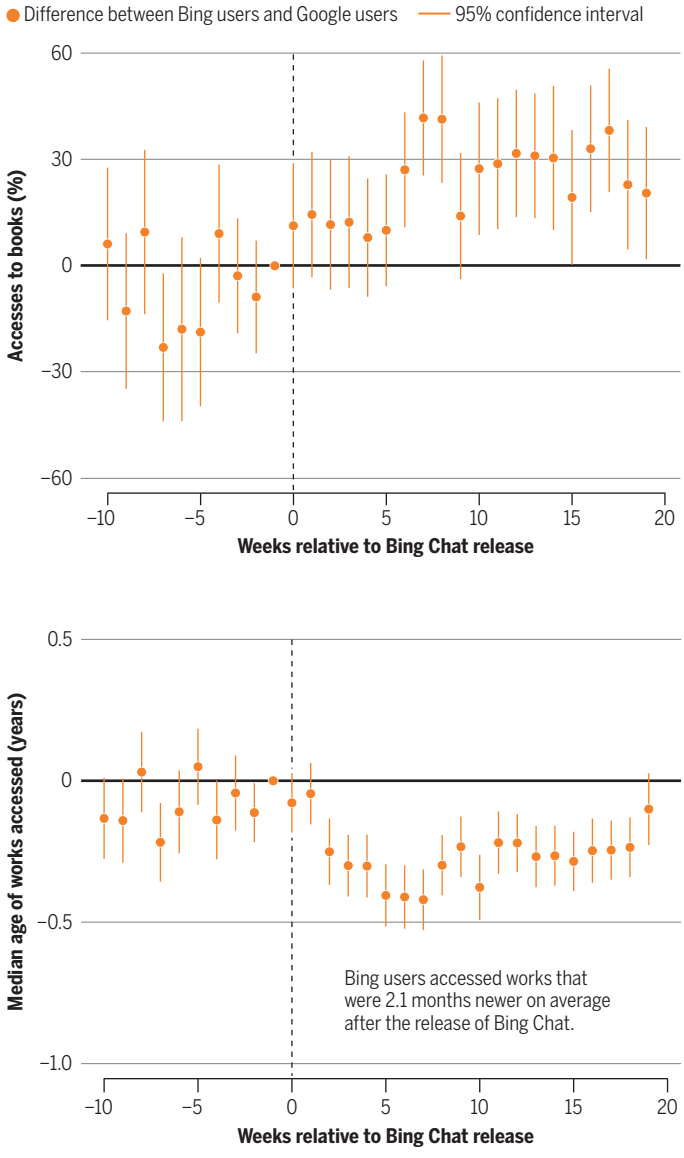
To assess the robustness of these findings, we examined additional features of writing (see SM S5.10). We replicated the findings using lexical complexity (syllables per word) and morphological complexity (fraction of present participial clauses). Both showed the same reversal pattern, in which increased writing complexity correlates negatively with publication success in LLM-assisted papers but positively in human-written papers. We also found the same pattern for the use of promotional language, measured using a standard lexicon (*10*), further confirming that LLM adoption erodes traditional quality signals across multiple linguistic dimensions.

Myriad factors influence the publication outcomes of preprints. We cannot rule out all confounding factors, but the results remain consistent after controlling for preprint month and field of study (see SM S3.2 and S5.9). As a robustness check, we collected and analyzed an independent dataset from the 2024 International Conference on Learning Representation (ICLR-2024), a leading conference in machine learning. ICLR-2024 provides access to 28,000 referee reports for the full set of 7243 submissions to the conference, regardless of their final acceptance status (see SM S1.4). Using the peer-review score assigned by experts as an alternative measure of scientific merit, the key findings were replicated with remarkable consistency (see fig. S3, D and H).

The sharp contrast in quality assessments across the distribution of language complexity in the two groups—human-written and LLM-assisted manuscripts—confirms that complex LLM-generated language often disguises weak scientific contributions. These findings demonstrate the rapid erosion of a traditional heuristic. For LLM-assisted manuscripts, the positive correlation between linguistic complexity and scientific merit not only disappears, it inverts. As the effort required to produce polished prose declines, so, too, does its utility as a signal of an author's command of a topic (*11*). This creates a risk for the scientific enterprise, as a deluge of superficially convincing but scientifically underwhelming research could saturate the literature. If this occurs, it will cause the community to waste valuable time separating genuine insights from a morass of unimportant and potentially misleading work.

Accesses to prior works

Bing Chat, powered by GPT-4, was released in February 2023. Relative to accesses to arXiv manuscripts redirected from Google, users of Bing accessed more books (top) and more recent works (bottom).



LLM USE AND ENGAGEMENT WITH PRIOR LITERATURE

Writing scientific papers also involves embedding claims and findings within existing literature. Because LLMs have the capacity to ingest and synthesize vast quantities of information, LLMs may broaden researchers' exposure to prior work (*12*). Or, as some have speculated, training data may overrepresent high-impact works, leading LLMs to amplify exposure to easily discoverable research (*13*). We therefore asked how LLMs affect the discovery of prior literature.

To evaluate these competing hypotheses, we leveraged a dataset capturing 246 million views and downloads of arXiv papers (see SM S1.5), each connected with a user ID, arXiv document ID, and referral source (Bing, Google, etc.). This dataset allowed us to explore changes in user-level reading behavior after the February 2023 release of Bing Chat (powered by GPT-4), the first widely adopted LLM-integrated search engine (see SM S3.4). We compared arXiv documents accessed by Bing users before and after this exogenous shift (see the second figure). Our estimations based on a differences-in-differences analysis showed that, compared with accesses redirected by Google, Bing users discovered a more diverse set of arXiv documents after the introduction of Bing Chat. We compared publication formats, showing that Bing users access books at a 26.3% higher rate ($P < 0.001$, Poisson regression), presumably reflecting an LLM's ability to surface content embedded in lengthy texts (see the second figure, top).

Increased exposure to books suggests that LLM-aided science may draw on a broader range of reference materials, but it does not rule out that LLMs simply reinforce attention to scientific canons. We investigated this possibility and found that Bing-referred visits were also linked to more recent scholarship; the median age of manuscripts accessed decreased by an estimated 0.18 years (see the second figure, bottom). Consistent with this shift toward the discovery of younger work, LLM users did not increase the number of times they accessed well-cited works. Instead, we found that Bing users uncovered references with fewer existing citations (see fig. S4C).

To examine whether this shift in search results translated to a change in actual citation behavior, we linked preprints in arXiv,

GRAPHIC: V. PENNEY/*SCIENCE*

bioRxiv, and SSRN to two large-scale citation databases: OpenAlex and Semantic Scholar. We obtained 101.6 million citations to prior works (see SM S1.6). We then used the event study methodology (see the first figure) to compare authors' citation behavior before and after they adopted LLMs, relative to a control group of nonusers (see SM S3.3). Our analysis explored three characteristics of cited references: publication format (citations to books), time lag (median reference age), and impact of cited work (mean log citations of referenced documents).

We found that LLM use alters authors' citation behavior, seemingly steering them toward a more diverse knowledge base (see fig. S4). LLM adopters overall were 11.9% more likely to cite books (see fig. S4D), but the effect is not statistically significant in one of the archives, SSRN. Adopters also cited documents that are on average 0.379 years younger (see fig. S4E) and have accumulated fewer citations (2.34% lower citation impact; see fig. S4F). Although the magnitude of these effects varied by preprint repository, the overall pattern appears broadly consistent (see fig. S4, G to I).

We present consistent evidence that AI assistance directs scholars to a broader body of knowledge (see the second figure and fig. S4). Researchers face time and attention constraints that limit their ability to process the expanding universe of research (*14*). LLMs appear to help researchers overcome obstacles in discovering pertinent literature.

These findings suggest that although LLMs may obscure signals of authorial effort, they broaden the path to knowledge discovery. A common concern has been that an AI-assisted search might reinforce the existing scientific canons. We found, however, that LLM adoption has had the opposite effect. Both AI-assisted search behavior and author citation patterns show a substantial shift toward a more diverse knowledge base, one that includes more books as well as younger and less-cited scholarship. This broadening of attention suggests that LLMs help researchers overcome cognitive constraints that have limited their ability to engage with the ever-expanding universe of scientific literature.

LIMITATIONS, IMPLICATIONS, AND FUTURE DIRECTIONS

In this study, we explored the impact of LLMs on scientific production, but our findings are subject to several limitations that offer avenues for future research (see SM S4 and S5). First, interpreting the estimated effects as causal requires assumptions that are difficult to satisfy, given data limitations that are inherent in studying LLMs "in the wild." Our AI detection method is imperfect and susceptible to several challenges (see SM S5.1 to S5.4): It relies on abstracts rather than full text (see SM S5.5), it cannot definitively identify which specific co-author on a team used an LLM (SM S5.6), and it almost certainly fails to detect use by authors who heavily edit LLM-assisted text. Furthermore, the nonrandom adoption of Gen AI tools creates the potential for self-selection bias, and our focus on posted preprints means the "adoption time" may be endogenous to productivity. The supplementary materials contain many additional analyses to evaluate the scope of these issues, and although our results appear robust, it is important for future work to continue to identify methodological strategies to address these challenges.

Second, our findings represent a snapshot of a rapidly evolving technology. Our analysis is based on data generated before the arrival of more advanced reasoning models and deep-research capabilities. As models improve and scientists discover new ways to integrate them into their work, the future impact of these technologies will likely dwarf the effects that we have highlighted here. This presents a crucial direction for future research: to continuously track how the scientific enterprise incorporates successive generations of AI models. Studies will need to examine whether the effects we have documented are amplified, altered, or even reversed as these more powerful tools are integrated into the scientific workflow.

There are many directions for future research. A primary avenue is more nuanced explorations of how LLMs are affecting scientific practice. Advancement in science has long been constrained by access to informal resources and knowledge. One hypothesis is that LLMs provide a scalable substitute for this informal knowledge, offering guidance on everything from experimental design to navigating a field's hidden curriculum, thereby leveling the scientific playing field. Another interesting avenue for future research is the potential for LLMs to transcend disciplinary boundaries. Over time, academic disciplines have developed deep knowledge bases that are often communicated through discipline-specific jargon. If LLMs help outsiders to overcome this hurdle, siloed disciplines may more productively engage with one another.

Our findings show that LLMs have begun to reshape scientific production. These changes portend an evolving research landscape in which the value of English fluency will recede but the importance of robust quality-assessment frameworks and deep methodological scrutiny is paramount. For peer reviewers and journal editors, and the community, more broadly, who create, consume, and apply this work, this represents a major issue. As a shortcut to (imperfectly) screen scientific research, writing characteristics are fast becoming uninformative signals, just as the quantity of scientific communication surges. As traditional heuristics break down, editors and reviewers may increasingly rely on status markers such as author pedigree and institutional affiliation as signals of quality, ironically counteracting the democratizing effects of LLMs on scientific production. One potential response is to leverage the same technology to assist in evaluating manuscripts. Specialized "reviewer agents" could flag methodological inconsistencies, verify claims, and even assess novelty. Whether this scalable approach will help editors and reviewers focus on substance over surface-level signals or introduce new and unforeseen challenges to the scientific process is a critical uncertainty. □

REFERENCES AND NOTES

1. J. Gao, D. Wang, *Nat. Hum. Behav.* **8**, 2281 (2024).
2. Q. Hao, F. Xu, Y. Li, J. Evans, arXiv:2412.07727 (2024).
3. K. Swanson, W. Wu, N. L. Bulaong, J. E. Pak, J. Zou, *Nature* **646**, 716 (2025).
4. M. Binz *et al*., *Nature* **644**, 1002 (2025).
5. W. Liang *et al*., *Nat. Hum. Behav.* 10.1038/s41562-025-02273-8 (2025).
6. E. Zhou, D. Lee, *Proc. Natl. Acad. Sci. U.S.A. Nexus* **3**, pgae052 (2024).
7. S. Noy, W. Zhang, *Science* **381**, 187 (2023).
8. E. Brynjolfsson, D. Li, L. Raymond, *Q. J. Econ.* **140**, 889 (2025).
9. C. Bazerman, *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science* (Univ. Wisconsin Press, 1988).
10. H. Peng, H. S. Qiu, H. B. Fosse, B. Uzzi, *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2320066121 (2024).
11. Z. Wojtowicz, S. DeDeo, *Proc. Conf. AAAI Artif. Intell.* **39**, 1592 (2025).
12. Y. Tian, Y. Liu, Y. Bu, J. Liu, arXiv:2501.00367 (2024).
13. A. Algaba *et al*., arXiv:2405.15739 (2024).
14. B. F. Jones, *Rev. Econ. Stud.* **76**, 283 (2009).
15. K. Kusumegi *et al*., Replication materials for "Scientific production in the era of large language models." Figshare (2025); https://doi.org/10.6084/m9.figshare.30359437.

ACKNOWLEDGMENTS

K.K. and X.Y. contributed equally to this work. The authors thank M. Naaman, W. Cong, W. Zhu, J. Mateos-Garcia, and seminar participants at the Complexity Science Hub (Vienna); the University of California, Los Angeles, Price Center; the Haas Macro Research Lunch seminar; and the Columbia Management, Analytics, and Data conference for helpful discussions. We also thank A. Cui for providing academic access to the GPTZero API. This work is supported by the National Science Foundation under grant nos. 2311521, 2404035, and 2412389. All data and code are available at Figshare (*15*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adw3000

10.1126/science.adw3000

[1]Department of Information Science, Cornell University, Ithaca, NY, USA. [2]Haas School of Business, University of California, Berkeley, Berkeley, CA, USA.
Email: mdevaan@haas.berkeley.edu; tstuart@haas.berkeley.edu; yian.yin@cornell.edu