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.

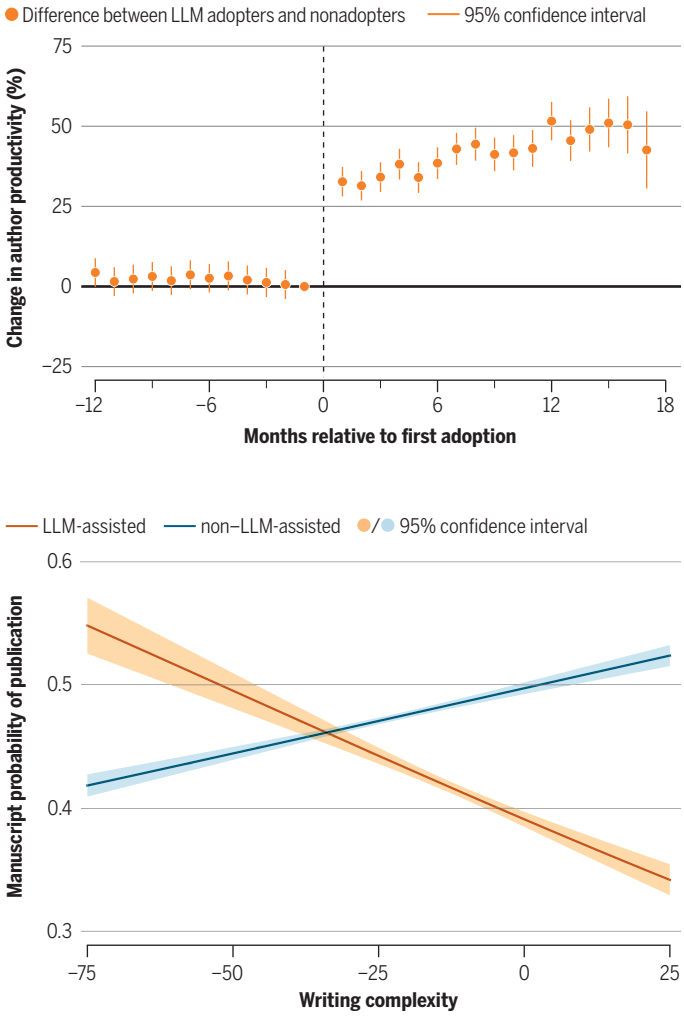
To identify the use of LLMs in the creation of scientific manuscripts, we applied a text-based AI detection algorithm (*5*) to all abstracts in our data. We used abstracts from papers submitted prior to 2023—before the ChatGPT era—to estimate the token (word) distribution of human-written text. We then 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.

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.

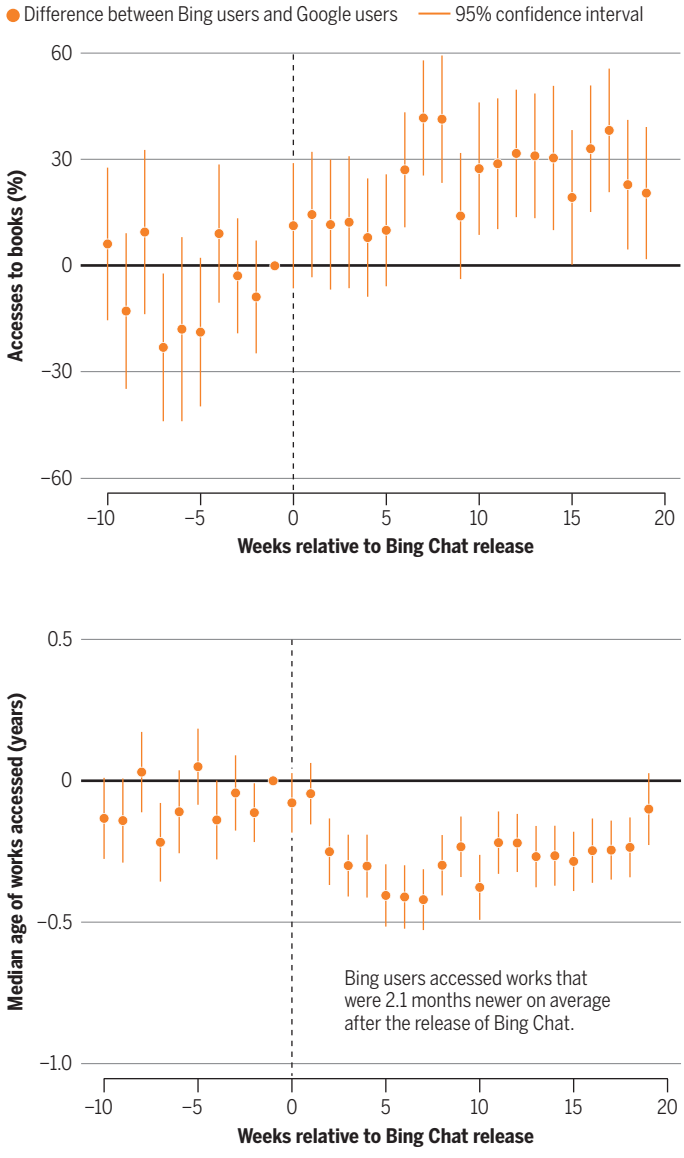
To evaluate these competing hypotheses, we leveraged a dataset capturing 246 million views and downloads of arXiv papers (see SM S1.5), each connected with a user ID, arXiv document ID, and referral source (Bing, Google, etc.). This dataset allowed us to explore changes in user-level reading behavior 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,

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).



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

Supplementary Materials for

Scientific production in the era of large language models

Keigo Kusumegi *et al.*

Corresponding authors: Mathijs de Vaan, mdevaan@haas.berkeley.edu; Toby Stuart, tstuart@haas.berkeley.edu; Yian Yin, yian.yin@cornell.edu

Science **390**, 1240 (2025)
DOI: 10.1126/science.adw3000

The PDF file includes:

Supplementary Text
Figs. S1 to S41
Tables S1 to S10
References

S1 Dataset description

S1.1 arXiv preprints (D_1)

We retrieved our first dataset D_1 from arXiv, one of the largest preprint repositories. Established in 1991, arXiv has hosted nearly 2.4 million preprints in the mathematical, physical, and computer sciences. We retrieved paper-level metadata from a weekly updated data dump hosted on Kaggle (https://www.kaggle.com/datasets/Cornell-University/arxiv), which includes information on arXiv ID, DOI, title, abstract, publication date (determined by the date at which the initial version was uploaded), field categories, and the list of authors. Our analysis includes 1.2M papers from January 2018 to June 2024.

To ensure our results are not heavily influenced by the unprecedented growth in research on Artificial Intelligence (AI), we excluded AI papers that are associated with at least one of the following categories:

Computer Vision and Pattern Recognition (cs.CV), Machine Learning (cs.LG), Artificial Intelligence (cs.AI), Information Retrieval (cs.IR), Computation and Language (cs.CL)

After these restrictions, we are left with 859K preprints.

S1.2 bioRxiv preprints (D_2)

Similar to D_1, we collected preprints in the biological sciences from bioRxiv (D_2). Established by Cold Spring Harbor Laboratory in 2013, bioRxiv is currently the largest preprint platform in life sciences. Submissions to bioRxiv cover a wide range of topics in biology, including genetics, neuroscience, ecology, and bioinformatics. We retrieved paper-level metadata through the bioRxiv API (https://api.biorxiv.org/), which includes information on DOI, title, abstract, publication date, field categories, and the list of authors. In our analyses, we focused on papers published between January 2018 and June 2024, covering 221K papers and 730K unique authors.

S1.3 SSRN working papers (D_3)

Our third dataset (D_3) is retrieved from the Social Science Research Network (SSRN). Founded in 1994, SSRN hosts working papers in the social sciences, law and the humanities. Since 2016, SSRN has expanded its scope through a partnership with Elsevier. Despite the emergence of other

recent platforms (e.g., PsycRxiv and others from the Open Science Framework), SSRN remains the largest preprint server by volume in the social sciences. A key challenge in collecting SSRN data stems from the lack of official data dumps or API services. To circumvent this impediment, we developed a multi-step strategy to create a sample of SSRN papers:

1. We started by constructing the universe of all SSRN papers using OpenAlex data. We identified all papers for which the DOI or landing page URL indicates an SSRN source. For each paper, we have information on DOI, title, field categories (assigned by OpenAlex algorithms), and the list of authors. These steps yield 1.3M papers hosted on SSRN.

2. A manual inspection of the sample above reveals a substantial recent surge in non-social science papers. This trend is likely due to SSRN's partnerships with journals outside the social sciences (e.g., The Lancet) and may introduce temporal bias in panel data. To address this, we query additional data from CrossRef, which is the largest DOI registration agency. CrossRef tags preprints directly submitted to SSRN with a container titled 'SSRN Journal', while leaving journal-referred preprints untagged. We restrict our sample to the former (394K papers, Jan 2018–June 2024) and use CrossRef's date-created field for its superior quality relative to OpenAlex.

3. Abstract information is missing for 12.1% of the papers in the filtered sample, which is essential for LLM detection. To this end, for any article that (i.) lacks an abstract, and (ii.) is authored by at least one researcher in our panel (S2.3), we supplement its missing abstract by querying the Dimensions API or scraping the corresponding article pages on SSRN.

S1.4 ICLR submissions and reviews (D_4)

Using the OpenReview API (https://api2.openreview.net), D_4 includes 7.4k papers submitted to the International Conference on Learning Representation (ICLR-2024), a leading conference in machine learning. To the best of our knowledge, ICLR is the only large-scale academic conference that makes public referee reports for all submissions, irrespective of their final acceptance status. The 2024 edition of ICLR marks the first one after the release of ChatGPT, with submissions due in Sept 2023.

After excluding papers that were desk rejected or that authors withdrew prior to the review process, our dataset includes 28K peer-review evaluations corresponding to 7.2K papers. Each review contains a six-category rating: strong reject (numerical score: 1), reject (numerical score: 3), marginally below the acceptance threshold (numerical score: 5), marginally above the acceptance threshold (numerical score: 6), accept (numerical score: 8), and strong accept (numerical score: 10).

S1.5 arXiv accesses (D_5)

To understand how LLM use influences how scholars discover prior literature, we leveraged a dataset of online accesses to arXiv preprints. Extracted from arXiv's internal web logs between Jan 2023 and Oct 2023, these data cover the full set of accesses to arXiv preprints across the world. In our analysis, we focused on the first 25 weeks of 2023 (Jan 1st - Jun 24th), corresponding to 598M records.

Of these 598M records, approximately 45% contain explicit referral information. We further excluded visits to RSS feeds and mailing lists—accounting for roughly 10% of the remaining sample—as these entries cannot be reliably linked to individual papers. By filtering out these records, our final analytical sample includes 246 million access records. For each of the remaining accesses (i.e. referred visits to individual arXiv papers), we obtained information on paper ID, access time, url referral, and hashed session number (based on cookie and IP address).

S1.6 Citation data linkage

We also collected data on references cited by recent preprints. As high-quality citation information is not available from sources $D_1 - D_3$, we use DOI to link each preprint repository with two state-of-the-art citation databases: Semantic Scholar and OpenAlex.

Manual inspections of random samples suggest that the two citation databases are complementary in their reference coverage, leading us to combine both for citation analysis (Table S1). S5.3 documents robustness checks using either data source, showing highly consistent results.

For each preprint, we retrieved a list of cited references. We used these to construct three variables:

- **Reference age**: the time lag between the citing preprint and cited reference.

- **Reference to books**: a dummy variable indicating whether the reference is labeled as a 'book' or 'book chapter'. This variable is exclusively based on OpenAlex data, as Semantic Scholar does not assign publication types for the vast majority of indexed papers.

  Note that the OpenAlex classification of publication formats may also be imperfect. Here we first manually checked a random sample of 100 works classified as book chapter in OpenAlex. By consulting publisher websites, we identified 4 instances that were indeed conference papers, suggesting a modest rate of misclassification. Yet at the same time, all these cases were published by Springer Nature, a major publisher that accounts for 29.3% of OpenAlex "book-chapter" works. To this end, we conducted a more targeted manual review, on another random sample of 50 "book chapter" works specifically published by Springer Nature. In this sample, we observe a 22% misclassification rate, in which conference papers are listed as book chapters.

  The significant proportion of miscategorized works identified in this exercise, especially within Springer Nature publications, prompted us to conduct a more systematic effort to accurately identify and exclude conference papers. To achieve this at scale, we obtained access to the Springer Nature meta-API, which provides a more reliable publication type provided by publishers ("contentType" returned by API). The content type can be: 'Chapter conference Paper', 'Chapter', 'Chapter Reference WorkEntry', or 'Chapter Protocol'. There were 171,326 book-chapter OpenAlex works published from Springer Nature across 3 repositories, and we successfully collect and detect the content type for 166,328 works (97%) using this approach. In our analysis we have marked these items as non-books.

- **Scientific impact of reference**: number of citations received by each reference. To account for the skewed distribution of citations, we compute $\log(\#\text{ citations} + 1)$.

S2 Methods

S2.1 Detecting LLM usage

We build on the distributional LLM quantification framework developed in (*S29, S30*) to identify the use of an LLM in scientific writing. The framework estimates the proportion of AI-modified sentences (α) in a document by modeling token occurrence probabilities in human-written ($P_T(X)$) and AI-modified ($Q_T(X)$) texts. These probabilities are parameterized as $P_T(X) = \prod_{t\in T} p_t^{1\{t\in X\}}(1-p_t)^{1\{t\notin X\}}$ and $Q_T(X) = \prod_{t\in T} q_t^{1\{t\in X\}}(1-q_t)^{1\{t\notin X\}}$, where p_t and q_t are estimated using collections of known human-written and AI-modified documents as $\hat{p}_t = \frac{1}{n_P}\sum_{j=1}^{n_P} 1\{t \in X_j^P\}$ and $\hat{q}_t = \frac{1}{n_Q}\sum_{j=1}^{n_Q} 1\{t \in X_j^Q\}$, respectively.

For each of the three preprint datasets (arXiv, bioRxiv, and SSRN), we constructed ground-truth data of human-written and AI-modified documents by randomly selecting 2,000 papers each month from January 2022 to October 2022. Specifically, we (1) used the original abstracts to estimate the token distribution of human-written text, and (2) prompted GPT-3.5 turbo-0125 to rewrite these abstracts and then used the resulting text to estimate the token distribution of LLM-written text. For the observed documents $\{X_i\}_{i=1}^N \sim \mathcal{D}_\alpha$, the framework infers α by maximizing the log-likelihood of the observed mixture distribution $\hat{D}_{\alpha,T}(X) = (1-\alpha)\hat{P}_T(X) + \alpha\hat{Q}_T(X)$, where $\alpha_T^{MLE} = \arg\max_{\alpha\in[0,1]} \sum_{i=1}^N \log((1-\alpha)\hat{P}_T(X_i) + \alpha\hat{Q}_T(X_i))$. This approach is applied using sentences as data points.

Fig. S5 visualizes the distribution of estimated α for preprints published after the initial release of ChatGPT. In each dataset, $P(\alpha)$ follows a bimodal distribution, reflecting a mixture of (i.) human-written text ($\alpha \approx 0$) and (ii.) potentially LLM-written text ($\alpha >> 0$).

To distinguish between the two groups, we define $\alpha_0 = 0.1$ (for each of the three preprint platforms) and 0.15 (for ICLR submissions) as the thresholds, corresponding approximately to the mean minus one standard deviation of subset (ii.). A paper is classified as LLM-assisted if it (i.) is posted after Dec 2022; and (ii.) registers an $\alpha > \alpha_0$.

At present, there is no widely accepted gold standard to determine the optimal threshold, as the α distributions cannot be easily approximated by standard parametric forms. To better justify the threshold choice, here we consider a comparative approach and examine how the distribution of α shifted before and after the availability of LLMs (comparing 2022 and 2023+ submissions). In Fig.

S6, we plot the relative density ratio, $P_{2023+}(\alpha)/P_{2022}(\alpha)$, which crosses the baseline value of 1 around $\alpha \approx 0.1$. This suggests a sharp increase in the prevalence of abstracts with $\alpha > 0.1$ (and a corresponding decline below), which is a good reason to select $\alpha_0 = 0.1$ as a meaningful dividing line for identifying LLM-assisted writing. S5.3 presents sensitivity analyses based on alternative values of α_0, suggesting that our results are robust across a wide range of thresholds.

S2.2 Author name disambiguation

Our analysis relies on author name disambiguation, a central task in the science of science (*S31*).

For arXiv, we disambiguate individual authors solely based on their names, using first name, last name, and name suffix (if available) as parsed from the official data dump, treating each unique combination of these three fields as a distinct author profile. To focus on individual careers, we exclude entries representing named collaborations (e.g., the LIGO Scientific Collaboration) from the authorship fields.

While arXiv publication can be linked to the works in OpenAlex, we find that OpenAlex's algorithm struggles with disambiguating authors in arXiv mainly due to the large number of authors for whom we have first name initials only: 24.5% of the author records in arXiv include author first name initials only (compared with only 1.7% in bioRxiv and 1.5% in SSRN). As a robustness check, we apply the same technique on OpenAlex ids and limit the analysis to authors with full names, finding that our results can be qualitatively replicated on this relatively higher-quality subset. Indeed, Fig. S7A confirms increased scientific productivity following first-detected LLM use in this subsample, although the effect size appears attenuated – likely two different full-name authors may still be mis-merged into the same OpenAlex author id.

For bioRxiv, we first source author name information from OpenAlex – as the official metadata APIs do not provide author first names. We then used a name-based approach similar to arXiv to disambiguate authors. Our main results remain robust when using author-id-based (assigned by OpenAlex) identifications. Fig. S7B visualizes our estimates for bioRxiv based on OpenAlex author id, which appears highly consistent with Fig. 2B.

For SSRN, since our samples are sourced directly from OpenAlex, we relied on OpenAlex disambiguated author id when constructing the panel data.

S2.3 Constructing panel data of individual careers

To examine the impact of LLMs use on research productivity, we track the publishing dynamics of individual researchers, covering 505,350 researchers (302,474 for arXiv; 183,255 for bioRxiv; and 19,621 for SSRN).

Given the initial release of ChatGPT in December 2022, we focus on researchers with at least 4 works published between 2018 and 2021. For each author, we track the number of preprints they posted each month ('monthly productivity') during a 30-month period (Jan 2022 - June 2024).

We also compiled quarterly panel data for the same set of authors by aggregating reference-level metrics (S1.6) to author × quarters. For the reference-level analyses presented in Fig. 4 (main text), note that the panel is not necessarily balanced, as some author-quarter pairs are not associated with any paper.

S2.4 Defining treatment and control groups

We examine the effect of LLM adoption using a stacked Difference-in-Difference approach. One key assumption is that once an author adopts an LLM, they continue to use it in subsequent work. Accordingly, we define treatment time as the author's first month of LLM adoption, identified by their earliest LLM-assisted publication. The control group comprises never-treated authors–those with no LLM-assisted publications as of June 2024. Each author in this group is assigned a unique event time, randomly drawn between January 2023 and June 2024.

S2.5 Name-based race and ethnicity inference

For each author in the data, we infer their race and ethnicity using name-based probabilistic models.

An author's race is estimated using the Python software ethnicolr2 (*S32*), which is a pytorch-implemented model built on the US census data and Florida voting registration data. Leveraging long short-term memory (LSTM), the model performs at 0.85 accuracy (*S32*) and has been applied in other works (*S33, S34*). This software predicts five types of races: Non-Hispanic Whites, Non-Hispanic Blacks, Asians, Hispanics, and Others. In our dataset, Asians are the largest sample, followed by Non-Hispanic Whites and Non-Hispanic Blacks.

One potential issue with this model is the uneven distribution of race in training data (*S35*) – often

biased toward the non-Hispanic white population. This imbalance may cause over-representation of white people and worse accuracy for smaller sample groups (*S36*). Indeed, misclassifications are most likely to default to "Non-Hispanic White," (NHW) the largest group in the training data (*S36*). Hence our sample of authors labeled as NHW may inadvertently include authors from other racial or ethnic backgrounds. Assuming that individuals who are truly NHW generally have the highest English writing proficiency (and thus benefit least from LLM adoption) this misclassification would most likely over-estimate LLM adoption's productivity effect for them, hence leading to a conservative estimation of racial heterogeneity in LLM-related productivity gains.

For ethnicity, another name-based classifier model, ethnicseer (*S37*), is applied to predict 12 ethnicity labels including English, Chinese, and Japanese. The model performs with 85% accuracy for the test dataset (*S37*). This model is particularly effective for one of our key populations of interest, showing F1 scores over 90% for East Asian names (e.g., Chinese, Japanese, Korean). Conversely, its performance appears weaker when distinguishing among European nationalities (e.g., English, French, German). Therefore, the "English" group may also be noisy, as it likely includes a mix of native and non-native English speakers–again leading to an underestimation of the true productivity differences between linguistic subpopulations.

To address potential systematic biases from either name-based model and ensure the robustness of our findings, we incorporated an author's institutional country as an independent proxy for race and ethnicity, as detailed in our discussion of Figure 2. This triangulation of name-based inference with geographic data provides a crucial cross-validation of our classifications. While we concede that each method has its own imperfections, the consistent patterns observed across these different estimation techniques give us confidence in the validity of our heterogeneity analysis. The proportion of both race and ethnicity within our dataset are shown in Table S2.

S2.6 Identifying research affiliation

In addition to name-based inference methods, we estimate whether a researcher is likely a native English speaker based on their research affiliation.

In arXiv data, not all authors are required to report their research affiliations. To address this, we use the affiliation and country of the submitter (analogous to corresponding authors in journal articles) as a proxy. For each preprint, the submitter's registration data provides their country (or

region), which we extend to approximate the country of other coauthors. This approach assumes that if a researcher's publication profile is predominantly associated with submissions from a specific country, the researcher is also likely to reside in that country. In contrast, both bioRxiv and SSRN data provide high-quality author-level research affiliations. These raw affiliation strings have been processed and structured by OpenAlex, allowing us to retrieve standardized country codes for the research affiliation of each author.

Note that we assign each author to one institution/country only. If a researcher is affiliated with different research institutions (or even countries) across papers, we classify authors as affiliated with (1) **East Asian countries** if at least 80% of their affiliations are in China, Korea, or Japan; or (2) **English-speaking institutions** if at least 80% of their affiliations are in the United States, United Kingdom, Canada, or Australia.

S2.7 Quantifying writing complexity

We computed the Flesch Reading Ease Score for each document (*S38*). This frequently used measure of writing complexity is a joint function of a text's syntactic complexity (sentence length) and lexical complexity (word length) (*S39*):

$$\text{Writing complexity} = 1.015\frac{\#\text{ total words}}{\#\text{ total sentences}} + 84.6\frac{\#\text{ total syllables}}{\#\text{ total words}} - 206.835$$

Following existing practice, the measure is only calculated for abstracts with at least 100 words.

S2.8 Tracing publication outcomes of preprints

We traced the publication outcomes of preprints from arXiv, bioRxiv, and SSRN using OpenAlex.

For arXiv and bioRxiv papers, we linked them to their corresponding records in OpenAlex using their DOI.

For SSRN papers, we used a title matching approach to establish links by searching OpenAlex for records with titles that match the paper titles in lower case. OpenAlex provides details about the source location and publication status, enabling us to determine whether a preprint was later accepted or published in a conference or journal.

Our main analysis (Fig. 3F) is based on an exact lower-case title match between SSRN and OpenAlex papers. Yet in some cases, the preprint title may change during peer review for publication,

leading to potential false negatives of this linkage. To this end, we randomly sampled 50 SSRN preprints that lacked any journal / conference publication matches, and manually checked the publication outcomes for these manuscripts via Google Scholar and other websites. We found that 5 out of the 50 non-matched preprints were published. Two of these appeared in venues not indexed by OpenAlex, and three were published with slight title modifications that our exact-match step failed to capture. Hence the error attributable specifically to title changes is 3/50 ($\approx$ 6%). This exercise yields two key insights: First, the false-negative rate from exact matching appears modest, which is reassuring. Second, for these error cases, the discrepancies between preprint and published titles are generally minor (Table S3), suggesting that slightly relaxing the matching criterion would recover most missed matches.

Guided by this, we developed a fuzzy-matching procedure for unmatched preprints. As pairwise comparisons between these preprints and all OpenAlex papers is not computationally feasible, we instead use Elasticsearch, an open search and analytics engine available for all OpenAlex titles. For each SSRN preprint, we feed in the title and retrieve the top candidate matches. To minimize false positives, we retained only pairs with an edit distance (between lower cased query and candidate title) < 6. When applied to the full dataset, this procedure recovered 5% additional matches, which is broadly consistent with our manual validation exercise. Fig. S8 compares two versions of our results, suggesting that while the inclusion of these fuzzy matches raises the overall publication rate, it does not alter our key finding on the decoupling between writing complexity and publication outcomes.

S3 Regression models

S3.1 The impact of LLM use on scientific productivity

We estimate the productivity changes using a stacked difference-in-difference approach (see S2.4 for details). The variables are defined as follows:

Dependent variable y_{it}, defined as the number of preprints published by author i in month t.

Independent variables We define Treated_i to be a dummy variable that indicates whether author i is in the treatment group. To estimate dynamic effects, we introduce D^k, a relative event-time dummy variable for $k \neq -1$. Specifically, $D^k = 1$ if the time difference between the current

period and the event time equals k. We also consider a 2×2 DiD design, where the dummy variable $\text{Post}_{it} = 1$ if the current period is after the event time. Since treatment group members are by definition more productive in their treatment month (i.e., they upload a preprint that month), we exclude the month of treatment in the 2×2 estimations.

Control variables To control for variations in productivity across different authors and time, we introduce α_i, fixed effect terms for author i, and λ_t, fixed effect terms for time period t.

Since monthly productivity y_{it} is a discrete count outcome, we estimate the dynamic coefficients of interest γ_k using a Poisson regression:

$$y_{it} \sim \text{Poisson}\left(\alpha_i + \lambda_t + \sum_{k\neq-1} \gamma_k D^k \text{Treated}_i + \sum_{k\neq-1} \beta_k D^k + \epsilon_{it}\right)$$

We also estimate the overall effect γ using a similar approach:

$$y_{it} \sim \text{Poisson}\left(\alpha_i + \lambda_t + \gamma \text{Post}_{it} \times \text{Treated}_i + \sum_{k\neq-1} \beta_k D^k + \epsilon_{it}\right)$$

Besides the Poisson regression estimation, we also run a Diff-in-Diff model with OLS regression, which suggests productivity increases of 0.83, 0.66, and 0.55 additional papers per year following LLM adoption (with the caveat that the dependent variable is not normally distributed). For comparison, authors in our sample published an average of 2.38 (arXiv), 1.65 (bioRxiv), and 1.83 (SSRN) papers in 2022.

S3.2 Writing complexity and publication outcomes

To understand the relationship between writing complexity and publication outcomes, we first fit regression models for both LLM-written and non-LLM-written text (Fig. 3 in main text), followed by more comprehensive models presented by Table S4. The variables are defined as follows:

Dependent variable For a given paper i, publication outcome is measured by (i.) r_{ij} (for ICLR submissions), the rating score given by reviewer j, or (ii.) p_i (for preprints), a dummy variable indicating whether the paper has been published in a peer-reviewed venue prior to the time of censoring.

Independent variables The key variable of interest is WC_i, defined as the writing complexity score of paper i (S2.7). To examine the role of LLM use, we also define LLM_i, which takes the value of 1 if paper i is classified as LLM-assisted.

Control variables The publication outcome for preprints is right-censored by July 2024. To account for differences across scientific fields and preprint age, we introduce a full suite of field fixed effects F_{fi} (dummies that equal 1 if paper i is in field f), and "birth cohort" fixed effects M_{ti} (dummies that equal 1 if paper i is posted (as a preprint) in month t).

In Fig. S3, we separately estimate the same regression models for LLM-written and non-LLM written papers. For ICLR submissions (which receive non-binary reviewer ratings), we run OLS models of the form:

$$r_{ij} \sim \beta_c \mathrm{WC}_i + \epsilon_{ij}$$

For preprints, we estimate logistic regression models of the form:

$$p_i \sim \mathrm{Logit}\left(\beta_c \mathrm{WC}_i + \epsilon_i\right)$$

In S5.2, we also performed an additional analysis on the full dataset by including an interaction term in the regression model.

S3.3 Citation analysis

We analyze within-author changes in referencing behavior following authors' adoption of LLMs , using a statistical framework similar to S3.1. Specifically, we study changes in referencing behavior following a first use of an LLM using a stacked differences-in-differences approach (see S2.4 for details).

Dependent variables We focus on three dimensions of cited references: (i.) a_{it}, defined as the median age of references cited by author i in quarter t; (ii.) b_{it}, defined as the number of book / book chapters references cited by author i in quarter t; and (iii.) c_{it}, defined as the mean (logged) count of citations received by the references cited by author i in quarter t as of July 2024.

Independent variables Similar to S3.1, we include $\mathrm{Treated}_i$, D^k, and Post_{it} for DiD estimation.

Control variables In addition to author and time fixed effects (α_i and λ_t defined in 3.1), we also calculated nb_{it}, defined as the number of non-book references cited by author i in quarter t. By doing so, our analysis of reference types is not simply driven by changes in the number of cited references.

We estimated the impact of LLM use on the number of cited books using Poisson regressions:

$$b_{it} \sim \text{Poisson}\left(\alpha_i + \lambda_t + \sum_{k\neq -1} \gamma_k D^k \text{Treated}_i + \sum_{k\neq -1} \beta_k D^k + nb_{it} + \epsilon_{it}\right)$$

$$b_{it} \sim \text{Poisson}\left(\alpha_i + \lambda_t + \gamma \text{Post}_{it} \times \text{Treated}_i + \beta \text{Post}_{it} + nb_{it} + \epsilon_{it}\right)$$

The impact on reference age and citations are estimated using OLS regressions:

$$a_{it} \sim \alpha_i + \lambda_t + \sum_{k\neq -1} \gamma_k D^k \text{Treated}_i + \sum_{k\neq -1} \beta_k D^k + \epsilon_{it}$$

$$a_{it} \sim \alpha_i + \lambda_t + \gamma \text{Post}_{it} \times \text{Treated}_i + \beta \text{Post}_{it} + \epsilon_{it}$$

$$c_{it} \sim \alpha_i + \lambda_t + \sum_{k\neq -1} \gamma_k D^k \text{Treated}_i + \sum_{k\neq -1} \beta_k D^k + \epsilon_{it}$$

$$c_{it} \sim \alpha_i + \lambda_t + \gamma \text{Post}_{it} \times \text{Treated}_i + \beta \text{Post}_{it} + \epsilon_{it}$$

S3.4 Online access analysis

Finally, we examine changes in researchers' discovery of prior literature in arXiv following the introduction of Bing Chat (powered by GPT-4) on Feb 7th 2023 as a natural experiment. In particular, we adopt a Differences-in-Differences framework, comparing searchers' accesses to arXiv papers through references by Bing (`https://www.bing.com/`, the treatment group) and Google (`https://www.google.com/`, the control group). As visits referred by Bing represent a mix of information search-engine results and GPT-4-guided searches post-treatment, our DiD estimator almost surely represents a conservative approach to understanding the effect of AI-based search on scientific discovery.

For consistency with our citation analysis, we focused on non-AI documents referred from either Google or Bing. We then constructed a panel of users, which we identified as unique pairs of hashed session id × referral domain, who accessed at least one paper both before and after the launch of GPT4-powered Bing. We excluded users with more than 1,000 accesses in this period because they likely represent robots. For each remaining user, we aggregated their weekly accesses (referred by Google or Bing) over the first 25 weeks of 2023. The variables are defined as follows:

Dependent variables Similar to S3.3, we examine (i.) a_{it}, the median age of scientific works accessed by user i in week t; (ii.) b_{it}, the number of book / book chapters accessed by user i in week t; (iii.) c_{it}, the average (logged) citation of works accessed by user i in week t.

Independent variables In our DiD estimation, we define $\text{Treated}_i = 1$ for visits referred from Bing. We also include relative event-time dummy $D^k = 1_{k=t-6}$ for $k \neq 1$ and $\text{Post}_t = 1_{t \geq 6}$.

Control variables Our regressions also include user fixed effects α_i, week fixed effects λ_t, and nb_{it}, the number of non-book works accessed by user i in week t.

Our regression models are analogous to those in S3.3:

$$b_{it} \sim \text{Poisson}\left(\alpha_i + \lambda_t + \sum_{k \neq -1} \gamma_k D^k \text{Treated}_i + nb_{it} + \epsilon_{it}\right)$$

$$a_{it} \sim \alpha_i + \lambda_t + \sum_{k \neq -1} \gamma_k D^k \text{Treated}_i + \epsilon_{it}$$

$$c_{it} \sim \alpha_i + \lambda_t + \sum_{k \neq -1} \gamma_k D^k \text{Treated}_i + \epsilon_{it}$$

S4 Data limitations

Our datasets and methods represent the state-of-the-art in the field, but there are a number of limitations that readers should keep in mind when interpreting the results.

First, our text-based classifier offers a systematic approach to identifying LLM use, but it has several shortcomings. First, the inference framework relies on a ground-truth corpus generated by an LLM in an abstract rewriting task. The resulting token distribution may vary depending on the specific language model, the prompts executed, and the hyperparameters used. Although the LLM detector has demonstrated remarkable accuracy, for a number of reasons it will not achieve error-free detection, nor can it be used for error-free ”treatment” assignment. First, as we run the LLM detection algorithm on the text in abstracts, we should expect false negatives for preprints in which LLMs are predominantly used in other sections of a manuscript. Second, we may observe over-time drift in the accuracy of the detector as new models are introduced. Third, to estimate many of the regressions we assume LLM adoption at the level based on detection of LLM use at the *preprint* level, although the latter are majority written by teams. It is not necessarily the case that all team members will be simultaneous adopters of LLMs, which adds significant noise to the treatment effect. Likewise, we infer adoption date by relying on submission dates which almost certainly introduces a lag.

Second, scientific output increasingly has been produced by teams (*S40*). In line with the existing literature in science of science (*S41, S42, S43, S44, S45*), we have defined individual productivity by

considering all papers with which a researcher has been associated. Our results remain robust when we limit the data to first- and last-authors of preprints. However, variations related to team size and authorship ranks may not be fully captured. As *all* authors on a LLM-written paper are considered to be LLM adopters, our framework may give rise to false positives of adoption, where uses of LLMs by coauthors may lead to (i.) non-adopters being misclassified as LLM users, or (ii.) late adopters are miss assigned to an earlier treatment time. Given that the post-adoption productivity increase appears widespread across diverse disciplines – from biological sciences where large teams are the norm to social sciences where solo-author papers still constitute a significant proportion – it is unlikely that our findings can be solely attributed to measurement issues. To further validate our results, we focused on fields dominated by solo authors or small teams – defined as fields where the mode of team size across all papers is 1 – finding qualitatively consistent results, with an estimated productivity increase of 25.1%. S5.8-5.9 further documents several related exercises, confirming the robustness of our results.

Lastly, our analysis in Fig. 4 relies on widely used metrics in scientometrics and the science of science – publication format, reference age, and citation counts. While these measures have proven valuable as proxies for knowledge consumption, each captures only partial aspects of knowledge use and, at best, approximates how scholars engage with prior literature. Citations and accesses to scientific papers, for instance, may reflect many motivations beyond true intellectual uptake, which our data cannot fully disentangle (*S46*). A promising direction for future research is computational content analysis of the specific claims being cited, which could further clarify how LLMs are shaping the flow of scientific knowledge.

S5 Robustness checks

S5.1 Validation of α using GPTZero

To test the robustness of our LLM detection, we compare our results with GPTZero, a commercial AI detection tool using an end-to-end deep learning approach. It is trained on text datasets from the web, educational sources, and content generated by various LLMs, including ChatGPT, GPT4, Google-Gemini, Llama, and other new AI models. For a given text, GPTZero calculates the probability that the text was created by AI, $p_{\text{AI-written}}$. It also has been employed in recent studies

on LLM uses in science and has demonstrated strong performance (*S47, S48*).

To quantify the relationship between α and $p_{\text{AI-written}}$, we randomly selected 1,000 arXiv papers per month in 2023 and 1,500 arXiv paper per month in 2024 (January - June) 2024 and fed them into the GPTZero API. [1] Comparing LLM use estimated by our detection framework (α) to GPTZero ($p_{\text{AI-written}}$). The Pearson correlation coefficient between these measures is 0.64 (Fig. S9A), documenting a high level of consistency. Fig. S9B compares the distributions of $p_{\text{AI-written}}$ conditional on α, highlighting a clear difference between $P(p_{\text{AI-written}}|\alpha < 0.1)$ and $P(p_{\text{AI-written}}|\alpha \geq 0.1)$. Manuscripts classified as LLM-written by our method are also 12.01 times more likely to have a $p_{\text{AI-written}}$ exceeding the 0.5 threshold. Similarly, comparing $P(\alpha|p_{\text{AI-written}} < 0.5)$ and $P(\alpha|p_{\text{AI-written}} \geq 0.5)$ (Fig. S9C), manuscripts flagged as LLM-written by GPTZero exhibit significantly higher α values.

Together, Fig. S9 suggests remarkable consistency between α and $p_{\text{AI-written}}$, which is somewhat surprising, as they reflect two fundamentally different approaches. α estimates the fraction of LLM-generated content using a simple, transparent statistical model based on unigram distributions, while $p_{\text{AI-written}}$ estimates the probability that a given text is AI-written using a complex, black-box model that integrates a range of linguistic features. Our main analysis relies on the α measure for several reasons:

Scalability. GPTZero is proprietary and only available through a commercial API. Therefore, computing $p_{\text{AI-written}}$ in a large dataset is expensive (e.g., over $10K for arXiv abstracts alone). In contrast, α only requires training on a relatively small sample, and the pretrained models are publicly available. α is far more cost effective and reproducible for the research community.

Stability. α is derived from maximum likelihood estimation based on observed token frequencies in human- and LLM-generated text. In comparison, complex classifiers like GPTZero are sensitive to hyperparameter tuning and model updates. While a proper specification of these hyperparameters might yield higher accuracy and recall, they also create higher uncertainty and noisy estimates – especially on short text such as scientific abstracts. Notably, we have observed substantial fluctuations in GPTZero's predictions over the past few months, raising concerns about its replicability as the underlying technology evolves.

Transparency. Unlike $p_{\text{AI-written}}$, which yields a binary or probabilistic classification, α offers

[1] Model version: v-2024-01-09.

a continuous and interpretable estimate of LLM contribution. This enables detection of partial LLM use, which we believe is more representative of real-world writing practices. To illustrate this, we constructed synthetic test samples composed of varying proportions of human- and LLM-generated text (ranging from 0% to 100% LLM-generated content, which we do based on verified 2022 content). Applying both measures, we find that $p_{\text{AI-written}}$ acts more like a conservative binary classifier, flagging only the most heavily LLM-generated samples while misclassifying lighter uses as fully human-written (Fig. S10). Taken together, these findings suggest that while $p_{\text{AI-written}}$ may be useful for high-precision classification, the α metric offers a more nuanced and scalable tool for studying the continuum of LLM-assisted writing.

While a full replication using GPTZero is not possible for the reasons above, we are able to conduct a further robustness check. We leverage the conservative nature of GPTZero to construct a more stringent definition of the treated group. Specifically, we apply GPTZero to 10% arXiv papers likely to be LLM-assisted (i.e., those with $\alpha > 0.1$ and published after 2022), and then restrict our analysis to the subset with $p_{\text{AI-written}} > 0.5$. Considering these as "high-confidence" cases of LLM use, we redefine the treated group of authors as those having first LLM-assisted paper ($\alpha > 0.1$) within this subset ($p_{\text{AI-written}} > 0.5$). We then replicate our analyses in Fig. 1, again finding a robust productivity increase (Fig. S11).

S5.2 Validation of α using Biber model

As a cross-validation of our method, we have also implemented another LLM detector developed in (*S49*), which we refer to as the Biber method. Similar to GPTZero, this model combines a variety of language features in a machine learning framework to predict LLM-assisted writing.

We trained and applied this model on our datasets, finding the predicted probability of LLM-assisted writing is positively correlated with our α (Pearson's $r = 0.30$ for arXiv, 0.15 for bioRxiv, and 0.27 for SSRN). We find this model is also a conservative binary classifier, flagging only the most heavily LLM-generated samples while misclassifying lighter uses as fully human-written (Fig. S10). Still, when we replicate the estimates of scientific productivity growth using this alternative metric (Fig. S12), we find qualitatively similar patterns. Notably, however, in bioRxiv and SSRN we find suggestive evidence of a jump in productivity before the "treated" month. This again suggests that the new metric conservatively flags LLM use. Assuming this to be the case, the method assigns

treatment timing to be later than initial adoption, which explains why there is evidence of a pre-trend and the estimated coefficients are slightly smaller than our original results.

S5.3 Validation of different detection thresholds

In the main text, we applied a threshold $\alpha_0 = 0.1$ to determine the uses of LLM in scientific text. Other values of α_0 may also serve as reasonable thresholds and can influence the exact estimates. Broadly speaking, a lower α_0 threshold increases recall, it captures a higher fraction of LLM-assisted cases, but at the cost of reduced precision as more human writing may be misclassified. Conversely, a higher-threshold α_0 improves classification precision but reduces recall, likely leading to many false negatives.

To address this, here we present sensitivity analyses using a wide range of α thresholds= 0.05, 0.15, 0.2, and 0.5. Table S5 presents the corresponding estimates. We find our core results remain statistically significant across different thresholds. At the same time, the estimated effect sizes taper with higher thresholds (e.g., $\alpha_0 = 0.5$), which reflects a trade-off: when the threshold is set high, papers that are partially LLM-assisted (e.g., $0 < \alpha_0 < 0.5$) are labeled as non-LLM-assisted. This results in a misclassification of early LLM adopters ("treated") into the late adopter or non-adopter pool (i.e., controls). This will attenuate estimates of the treatment effect and bias them toward zero.

Dynamic event study estimates support this interpretation (Fig. S13). Under higher α_0 values, we see suggestive signs of a productivity increase before the first detected LLM use. This indicates that the true timing of LLM adoption likely occurred earlier than is detected when we employ a high α_0 value because we miss earlier LLM use that did not trigger the threshold. This again reinforces the conservative nature of estimates based on high α_0 values. These insights motivated us to conduct a third set of threshold analyses, where we avoid the "grey area" (papers with α values in the middle range) and only focus on samples with high- confidence predictions. In particular, we consider a paper as:

- LLM-written if $\alpha > \alpha_{0,H}$
- Human-written if $\alpha < \alpha_{0,L}$
- Unclear if its α falls between $\alpha_{0,L}$ and $\alpha_{0,H}$

We then define authors as adopters if they have published at least one "LLM-written" paper and no "Unclear" papers prior to that point; and non-adopters if they have no "LLM-written" or "Unclear" papers during the entire observation window. By doing so, we exclude people whose adoption status and timing is ambiguous because they wrote at least one "unclear" paper before any strong evidence of adoption (publishing any "LLM-written" paper). This approach allows us to focus on cases with higher classification confidence. Table S6 presents our estimates under this strategy, which shows robust (and slightly higher) parameter estimates.

S5.4 Estimating α on alternative language models

In our original analysis, we followed recent literature and used GPT-3.5 turbo-0125 to construct an LLM-written vocabulary. A potential limitation is that the model was released in early 2024 and may not precisely match the LLMs used by authors in 2023. Motivated by this concern, we revisited the ChatGPT release history and assessed potential differences between model generations. First, we note that the web-based version of ChatGPT was largely based on the GPT-3.5-turbo family, which includes the model variant we used. Due to evolving model versions and differences between the web interface and API access, the exact model available to users at any given time may differ slightly from the one used in our analysis. However, we think it is likely that most users had access to a family of models that shared core architectural and training features.

More broadly, we note that recent work in LLM detection suggests that inter-model variation (across different LLMs or versions) appears to be smaller than the human–machine distinction. For example, Cheng et al. (*S50*) developed a classifier trained on Falcon-7B and found it generalizable to texts produced by other models. More recently, Reinhart et al. (*S49*) conducted a classification task across seven classes (human-written text, two GPT-4 variants, and four LLaMA-3 variants), concluding that "little of the error was due to confusion between human texts and LLMs." These papers conclude that model-specific variation plays a relatively minor role compared with the fundamental difference between human- and machine-created writing styles. That said, we have conducted an additional robustness check using an older model from the GPT-3.5-turbo-instruct family, accessed through OpenAI's deprecated Completions endpoint. Released on September 14, 2023, this is the earliest version in the GPT-3.5-turbo family that remains accessible through OpenAI. We did not use GPT-2, as it is not instruction-tuned and cannot reliably generate rewritten

abstracts via prompting.

Fig. S14 presents the α scores computed using this older model. While the α values tend to be higher overall (likely reflecting differences in verbosity or generation style) we find a high correlation (Pearson's $r = 0.58$ for arXiv, 0.55 for bioRxiv, and 0.57 for SSRN) with our original estimates. To adjust for the upward shift in α values, we applied a recalibrated threshold ($\alpha_0 = 0.3$) for binary classification and replicated our main analyses using this adjusted scheme. The resulting estimates remain qualitatively similar across all key figures (Fig. S15 – S18).

We conducted another robustness check using LLaMA. Since LLaMA-3 was not released until Apr 2024, we used meta-llama/Llama-2-7b-chat-hf model–a version of LLaMA-2 fine-tuned for chat–from the Hugging Face library. Similar to our approach with GPT-3.5-turbo, we used LLaMa to rewrite abstracts of papers published before 2023 to generate a token distribution characteristic of AI-generated text. We then applied MLE to estimate the α parameter for all papers published since 2023. We observe that the responses from LLaMA-2 are more sensitive to prompt formulation. In particular, our original prompts (that were sufficiently clear for different GPT-3.5-turbo models) may still lead to incomplete, repetitive, or instruction-parroting outputs when used with LLaMA 2. To this end, we carefully refined our prompts with explicit and detailed instructions to ensure clean outputs from LLaMA 2.

We observe a strong, positive correlation between the α estimated using GPT-3.5 and LLaMA-2, confirming the robustness of our LLM detection (Fig. S19 – S22, Pearson's $r = 0.54$ for arXiv, 0.53 for bioRxiv, and 0.58 for SSRN)). Again using the recalibrated threshold ($\alpha_0 = 0.3$) for binary classification.

S5.5 Estimating α on introduction

Another essential trade-off in LLM use detection arises when one chooses a subset of text to focus on. Our approach uses the abstract, which is a relatively short paragraph where LLMs are disproportionately used to improve writing. On the other hand, one may develop a similar pipeline on the full text, which may yield more stable inference yet runs a risk of under-detection if an LLM is used to write a small fraction of the text. While we were not able to replicate the entire pipeline on full-text due to the cost of collecting, cleaning, and retraining our model on the full text data, here we present three analyses to explore the potential weakness of our focus on abstracts.

First, we collected LaTeX source files from a sample of arXiv papers. Since the full text is largely unstructured and contains a high fraction of non-natural language content (for example, equations), we used the S2ORC doc2json parser (https://github.com/allenai/s2orc-doc2json) to extract structured sections. From this, we successfully obtained introduction sections for 217,798 papers.

We then applied our detection pipeline to these introductions to compute α values and compared them to the α scores based on abstracts (Fig. S23A). We observe a positive correlation between the two sets of estimates (Pearson's $r = 0.59$), suggesting that LLM usage in abstracts is meaningfully associated with usage in the broader paper. Notably, we also find that α values estimated from introductions are systematically lower than those from abstracts, reinforcing the idea that abstracts are more likely to exhibit detectable LLM usage.

Next, we conducted a similar analysis using data from bioRxiv by downloading and parsing the full-text XML data dump. For 146,723 papers with successfully extracted and cleaned introduction sections, we applied the same LLM detection pipeline. The results mirror our findings from arXiv (Fig. S23B): α values estimated from the introduction are systematically lower than those from the abstract, yet remain positively correlated (Pearson's $r = 0.45$), reinforcing the idea that LLM use detected in abstracts is a meaningful proxy for broader use.

Lastly, we test the robustness of our tea results. Fig. S24 plots the productivity dynamics on bioRxiv data using the α values estimated from introductions. The results show a post-adoption productivity boost that mirrors the original Fig. 1B, albeit with a slightly smaller magnitude. This suggests that, despite the limitations of each detection strategy, our core conclusion remains broadly robust.

S5.6 Alternative definitions of LLM adoption

The assignment of author-level treatment time here is inherently imperfect (especially in large teams), which we flag as a key limitation that readers should keep in mind. While pinpointing the exact LLM user within a collaboration remains challenging, we now present two robustness checks to gauge the impact of such misclassifications on our results:

We conducted an independent validation exercise by consulting author contribution statements. In particular, we downloaded and parsed the full text XML dump of bioRxiv, finding 53,428

papers (43% of the corpus) include a section of author contribution statements. Since statements are unstructured (e.g., "A.B. and C.D. wrote the first draft"), we developed a Gemini-2.5-based pipeline to extract and normalize author names and roles (see Fig. S25 for an illustration of the pipeline). By so doing, we are able to restrict treatment to the subset of authors that drafted the manuscript. For example, a paper of team size 10 has an average of 5.76 primary contributors in writing (Fig. S26).

We then restricted the treated group to include only researchers who are tagged as principal contributors to manuscript writing on their first LLM-assisted paper, yielding a subset for which treatment assignment may be more precise. Re-estimating our difference-in-differences model on this subset (Fig. S27), the productivity dynamics remain similar to Fig. 1B. These results suggest that the productivity increases we observed in our main analyses are unlikely to be explained by the imperfect assignment of treatment month.

Unfortunately, the analyses shown above are limited to papers with author contribution statements, which are largely missing in arXiv and SSRN manuscripts. To this end, in our second robustness check, we approximate writing effort with authorship rank. Consistent with practices in the recent science of science literature, we consider first and last authors as principal contributors to writing. (Indeed, in the bioRxiv data, 84% first / last authors self-report leading roles in manuscript writing, in contrast to just 20% for other authors). Building on this idea, in Fig. S28 we restricted the treated group to include only first or last author on their first LLM-assisted paper. The estimates from this approach again replicate both the magnitude and significance of the productivity boost.

The two analyses discussed above are based on the idea of only sampling authors who have likely contributed to the writing process. This should reduce noise in adoption timing for teams with many co-authors. The results are highly consistent with our main analyses, which suggests that assigning treatment to all authors in large teams does not introduce bias.

S5.7 Alternative measures of individual productivity

Team size Many projects in science are produced by large teams, raising the question of how collaborations may affect or even confound our estimates of productivity changes. First, following related research on scientific collaboration, we consider the sum of inversely weighted team size – a paper written by n authors will contribute a $\frac{1}{n}$ unit of productivity to each coauthor. This measure

should be more robust against the inflation of team size. Fig. S29 shows our estimates on this weighted productivity metric, showing highly consistent results across all three preprint servers.

Lead authors Second, noting that a few "lead authors" are disproportionally responsible for large-team collaborations, we follow prior work on team dynamics in science production and redefine productivity as the number of first- or last- author papers produced by individual scientists, while keeping their adoption time unchanged. The estimated coefficients are again remarkably consistent with our original results (Fig. S30.)

Definition of AI papers Further, the definition of "AI" papers is evolving. For example, cs.CL (Computation and Language) historically included many linguistic works, but is now dominated by LLM-focused research. At the same time, categories such as cs.NE (Neural and Evolutionary Computing) have been foundational to AI but are less directly impacted by the recent surge of foundation models. Indeed, our primary goal here was to isolate AI-adjacent fields experiencing unprecedented growth in knowledge production, which we must do to ensure that our findings on LLM use and productivity dynamics are not simply driven by these background trends. We leverage the co-listing feature between arXiv categories, which allows for a paper to be cross-listed across multiple categories. This feature makes it possible to calculate the Jaccard similarity between any pair of categories A and B, defined as

$$J(\text{category } A, \text{category } B) = \frac{|\{\text{Papers in } A\} \cap \{\text{Papers in } B\}|}{|\{\text{Papers in } A\} \cup \{\text{Papers in } B\}|}$$

Fig. S31 visualizes the undirected weighted network, where nodes represent arXiv fields in computer science or statistics, with edge weights proportional to the Jaccard similarity. As the figure shows, categories like cs.CL and cs.CV cluster closely with core AI fields (e.g., cs.LG and cs.AI), providing empirical support for our rationale. At the same time, it suggests that stat.ML is a highly related field (and cs.IR a less relevant one).

In light of these observations, we now conduct robustness checks using two alternative definitions of "AI fields" : (v1) the set of AI field following (*S51*) (cs.AI, cs.LG, cs.NE and stat.ML) and (v2) the data-driven community of fields derived from the manuscript co-listing network (cs.CV, cs.LG, cs.AI, cs.CL and stat.ML). Our results appear largely consistent across different specifications, with slightly lower estimates when we consider cs.CL and cs.CV as AI research (Fig. S32).

Including AI papers It is possible that AI-related subfields may benefit from LLM tools that accelerate productivity and experience a rapid influx of new research questions in the LLM-era. Both phenomena could potentially drive productivity changes. In our original analysis, we excluded papers from AI-related subfields (cs.CV, cs.LG, cs.AI, cs.IR, cs.CL) from our analysis for precisely this reason.

That said, we reran the productivity analysis on a sample that includes all papers, including those explicitly about AI. Fig. S33 plots the results, based on the original non-AI sample (green) and the expanded sample including AI papers (purple). Adding AI papers increases the estimated productivity growth to 0.476. This also suggests that excluding AI-related research yields a conservative estimate of the LLM-related productivity effect, further demonstrating that our core finding remains robust across different field compositions.

S5.8 Heterogeneity in productivity changes

Fields of study The productivity changes observed in Fig. 1 may be correlated with specific scientific fields of study, raising the question of whether this trend is driven by certain fast-moving fields. To this end, we first assign a field label to each author in our arXiv panel, defined as the modal field of their publications prior to 2022. We then included an interaction term between this researcher-level field and post_treated in our Diff-in-Diff model. Fig. S34 visualizes our estimates for 148 fields on arXiv. While the effects vary substantially across fields, we find positive coefficients in 138 out of 148 fields. To ensure robustness against noise from small sample sizes, we then restricted the analysis to fields with at least 200 authors. All 86 such fields exhibit positive effects, again suggesting the results are not driven by a small number of outlier research communities.

Native vs non-native english speaking countries Figure 2 shows substantial heterogeneity across race, ethnicity, and research affiliations, suggesting a link between productivity changes and English proficiency. We also performed a similar analysis at the country level, where we focused on 43 countries with at least 100 authors in our arXiv panel data, repeating our analysis on each of them separately. Fig. S35 presents the distribution of productivity gains of native and non-native English-speaking countries, again suggesting non-native speakers are associated with a higher productivity gain.

Researcher experience In our analysis, we focused on researchers with at least 4 papers in

a 4-year period (2018-2021). Conditioning on 4+ manuscripts means that our results should be interpreted as an estimate on the productivity of "incumbent scientists".

That said, the pool of "incumbent scientists" remains heterogeneous, allowing us to perform an additional analysis to examine productivity changes between more and less experienced researchers. For each author we calculated the number of published papers between 2018-2021, which provides a noisy proxy for productivity. For each preprint server, we then median-split authors based on a cutoff of prior productivity of 7 (arXiv), 5 (bioRxiv), and 6 (SSRN), respectively. Table S7 report our estimates for above- and below-median scholars. Results show less experienced subgroups appear to benefit more from LLM adoption, but the differences appear relatively small.

S5.9 Robustness checks on writing complexity and publication outcomes

To examine whether our measurements of writing complexity are simply driven by author race and ethnicity, Fig. S36 repeats Fig. 3A-C for authors with Asian and non-Hispanic White names separately. Across both racial groups and all three preprint platforms, we see robustness evidence that LLM-assisted manuscripts are associated with higher writing complexity scores.

We also examined the relationship between writing and publication quality, by running regressions with an interaction term between LLM use and writing complexity.

For ICLR submissions we estimated:

$$r_{ij} \sim \alpha \mathrm{WC}_i \times \mathrm{LLM}_i + \beta_c \mathrm{WC}_i + \beta_l \mathrm{LLM}_i + \epsilon_{ij}$$

For preprints we estimated:

$$p_i \sim \mathrm{Logit}\left(\alpha \mathrm{WC}_i \times \mathrm{LLM}_i + \beta_c \mathrm{WC}_i + \beta_l \mathrm{LLM}_i + \epsilon_i\right)$$

Finally, our analysis traced the peer review outcomes by mid 2024, which may include false negatives for papers that are still in the pipeline. This prompts us to run additional regression analysis, controlling for preprint month and field fixed effects.

$$p_i \sim \mathrm{Logit}\left(\alpha \mathrm{WC}_i \times \mathrm{LLM}_i + \beta_c \mathrm{WC}_i + \beta_l \mathrm{LLM}_i + \sum_f \beta_f F_{fi} + \sum_t \beta_t M_{ti} + \epsilon_i\right)$$

Table S4 reports a statistically significant positive coefficient for writing complexity and negative coefficients for interaction term across all 7 models, indicating that writing style serves as a rapidly diminishing signal of manuscript quality.

S5.10 Alternative measures of writing complexity

Notably, the Flesch reading ease score, while very widely used, is one of many potential measures of writing complexity. Since our primary interest is the possible decline in the signal value of writing features that have historically been interpreted as indicators of scientific quality, here we consider several additional metrics that capture writing style, which we associate with publication outcomes.

Lexical complexity: We first examine lexical complexity by calculating the number of syllables per word. Fig. S37 replicates our findings across all four datasets, showing LLM-assisted papers use more lexically complex words. However, there is a negative correlation between lexical complexity and peer review outcomes for LLM-assisted manuscripts. The opposite is true for researcher-written projects.

Syntactic complexity: We next re-examine the average sentence length – the syntactic component of the Flesch score. Interestingly, we find much weaker divergence for this measure (Fig. S38), indicating the effects we observed are mainly driven by lexical (vocabulary) rather than purely syntactic factors.

Morphological complexity: We also examine morphological complexity – another important dimension of writing composition. Guided by recent work on LLM writing styles (*S49*), here we focus on a salient morphological marker that separates human- from LLM-written text: the fraction of present participial clauses. Fig. S39 correlates this metric with LLM uses and publication outcomes, showing two patterns. First, LLM-assisted abstracts have a higher fraction of present participial clauses (1.08% vs 0.48% in arXiv, 0.81% vs 0.37% in bioRxiv, 0.74% vs 0.25% in SSRN, and 1.19% vs 0.69% in ICLR). Second, comparing human-written and LLM-written abstracts, we once again observe a reversal in the relationship between writing complexity and scientific merit, as proxied by publication outcomes. This finding recapitulates our broader conclusion, suggesting that LLM usage may erode traditional quality signals across multiple dimensions.

Promotional language: As an extension of these language features, we also examined another

potential signal of scientific quality – the use of promotional language. Following the work by Peng et al (*S52*), we have estimated the fraction of promotional words in scientific abstracts, and estimate its effect on manuscript placement (Fig. S40). First, we successfully replicate the positive relationship between promotional words and peer review outcomes in non-LLM papers, confirming the Peng et al (*S52*) result that promotional language signals perceived scientific quality. Second, we find LLM-assisted abstracts to contain a higher fraction (1.64% vs 0.97% in arXiv, 1.52% vs 1.03% in bioRxiv, 2.05% vs 1.32% in SSRN, and 2.11% vs 1.55% in ICLR) of promotional words, suggesting this quality signal may attenuate in LLM generated text. More interestingly, for LLM-assisted papers we observe a reversal of the positive relationship between promotional language and publication success. LLM-assisted manuscripts with extensive use of promotional words are less likely to clear peer review. Together, these results provide strong empirical evidence that promotional language, another signal of scientific quality, may lose its relevance in the post-LLM era of scientific production.

S5.11 Robustness checks on citation analysis

We conducted three additional robustness checks to validate our citation analysis (Fig. 4).

First, to explore the relationship between LLM use and references to prior works, we extended our analysis beyond the DiD framework by performing a series of paper-level regressions for all post-ChatGPT manuscripts:

Dependent variables Similar to S3.3, for a given paper i we examine (i.) a_i, defined as the median age of references cited by paper i; (ii.) b_i, defined as the number of book / book chapters references cited by paper i; and (iii.) c_i, defined as the mean (logged) count of citation received by the references cited by paper i.

Independent variable LLM_i, a dummy variable representing whether paper i is classified as LLM-written.

Control variables Our baseline model controls for the preprint month and field of study. In an extended model, we also include team size and individual fixed effects for the first and last authors.

We estimated these models for manuscripts published in 2023 and 2024. Results remain consistent across datasets and regression specifications (Tables S8, S9, S10).

Further, we ask whether our results are sensitive to citation data sources. The coefficients

visualized in Fig. 4 were estimated using a combination of citation data from Semantic Scholar and OpenAlex (S1.6). To test robustness, we replicated the results in Fig. 4D,F using only Semantic Scholar (Fig. S41A,D) or OpenAlex (Fig. S41B,E) respectively. These analyses produced highly consistent results.

Lastly, following standard practice in the literature (*S53*), we treat missing author-quarter cells as "NA" (null) for reference-related metrics, rather than imputing them as zeros. To probe the potential impact of this choice on the results, we conduct a robustness check using a 2×2 difference-in-differences design. Instead of aggregating citation-based metrics on the author-quarter level, here we aggregate all pre-treatment (and post-treatment) papers for each author as the unit of analysis. We then estimate the effect of LLM adoption using a TWFE model, which automatically drops authors with only pre-treatment or post-treatment papers and therefore eliminates the issue of missing data. This alternative specification suggests authors using LLMs cite more books (25.6% ↑), more recent references (0.505 years ↓) and slight less impactful references (1.48% ↓), yielding results that are consistent in both direction and magnitude with our main estimates.

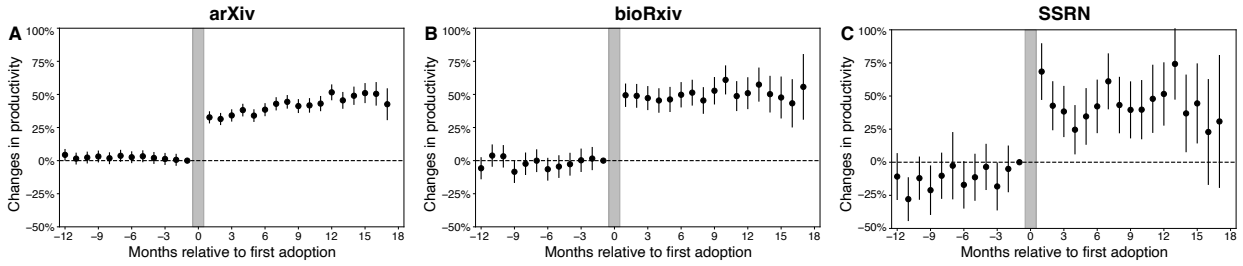


Figure S1: **LLM usage and scientific productivity.** We track the productivity dynamics (measured as the number of preprints published monthly) of 168,553 authors between Jan 2022 and July 2024, distributed across arXiv (109,965 authors), bioRxiv (43,218 authors), and SSRN (15,370 authors). For each author, we apply a text-based detector to their preprints to determine whether and when they "adopted" LLMs in scientific writing. (A-C) Using a stacked difference-in-difference regression, we estimate the impact of LLM on individual productivity. Comparing researcher-level pre- and post-adoption, we conservatively observe significant productivity increases, with boosts of 36.2% (arXiv), 52.9% (bioRxiv), and 59.8% (SSRN) relative to non-adopters. Panel A is identical to the figure shown in the main manuscript (the first figure, top).

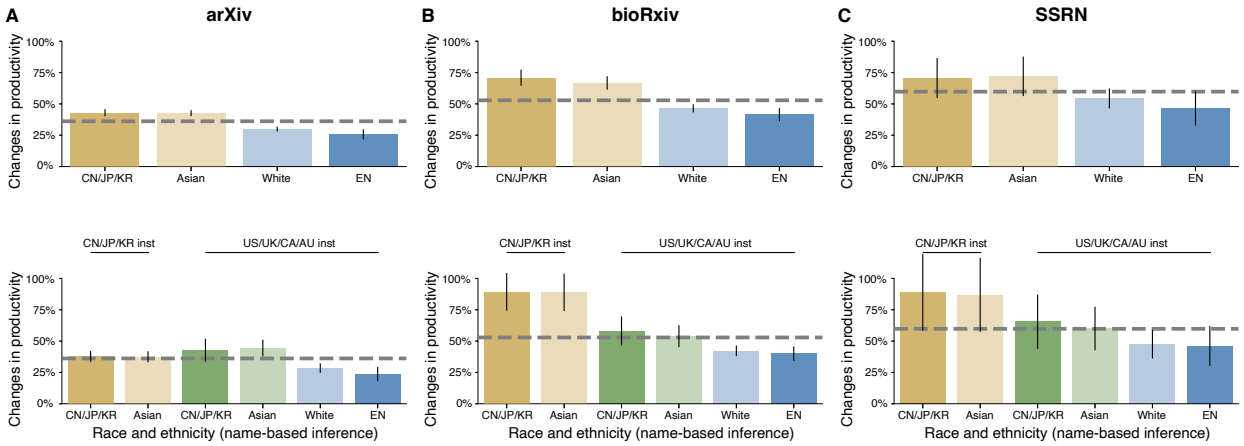


Figure S2: **Heterogeneity across races, ethnicities and home geographies.** (A-C) The productivity gains are pronounced for authors with East Asian names, showing increases of 43.0% (arXiv), 70.9% (bioRxiv), and 70.1% (SSRN), but remain meaningful for authors with Caucasian names, with gains of 25.7% (arXiv), 41.5% (bioRxiv), and 46.5% (SSRN). The effect is most pronounced for authors with Asian names affiliated with institutions in Asia, showing productivity boosts of 37.7% (arXiv), 89.3% (bioRxiv), and 88.9% (SSRN). By comparison, the productivity boost for authors with Caucasian names affiliated with institutions in English-speaking countries, are 23.7% (arXiv), 40.0% (bioRxiv), and 46.2% (SSRN). The gray dashed line represents the average effect across all authors.

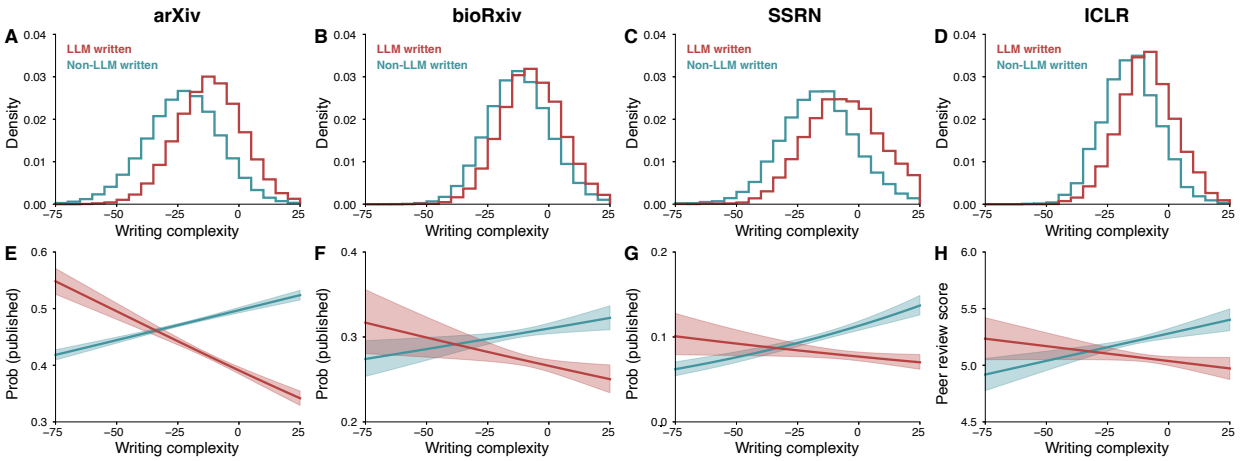


Figure S3: **LLM usage, scientific writing, and publication outcomes.** For 264,125 manuscripts written since 2023 (177,880 on arXiv, 58,958 on bioRxiv, 31,959 on SSRN, and 7,243 submitted to ICLR 2024). We measure writing complexity as the negative of the Flesch Reading Ease Score, a metric defined as a joint function of a text's mean sentence and word length. We also measure paper quality using (i) the probability of peer-reviewed publication for preprints (prior to censoring), and (ii) peer review scores for ICLR submissions. (A-D) Distribution of writing complexity for LLM-assisted (red) and non-LLM-assisted (blue) manuscripts. LLM-assisted manuscripts exhibit significantly higher writing complexity across all four datasets. (E-H) Relationship between writing complexity and paper quality. For non-LLM-assisted manuscripts, writing complexity is positively correlated with measures of perceived manuscript quality. In sharp contrast, for LLM-assisted manuscripts, greater writing complexity correlates with lower manuscript quality. Predicted outcomes are based on logistic regressions (publication probability in E-G) and OLS regressions (peer review scores in H), respectively. Panel E is identical to the figure shown in the main manuscript (the first figure, bottom).

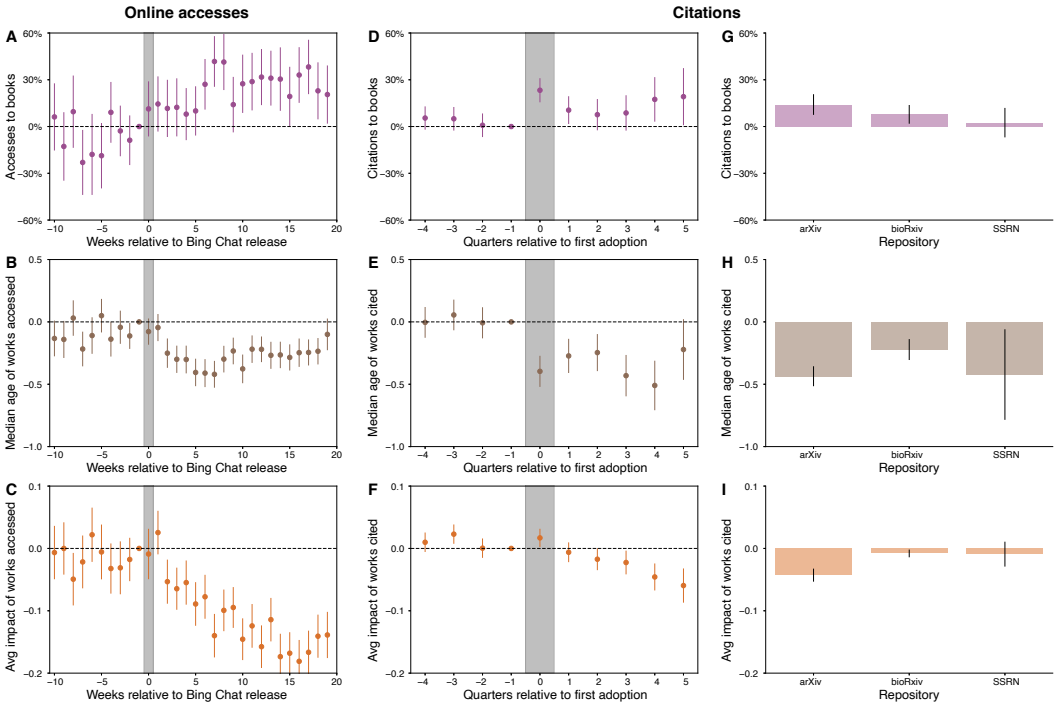


Figure S4: **LLM usage and references to prior works.** (A-C) We examine user-level changes in access to arXiv manuscripts following the release of Bing Chat (powered by GPT-4) in Feb 2023. Comparing online accesses redirected from Google and Bing, we find users access (A) more books, (B) more recent works, and (C) less highly cited works, post-event. (D-I) We examine author-level changes in referencing / citation patterns following authors' initial adoption of LLMs, using a Diff-in-Diff strategy as in Figure 1. Due to data sparsity, all outcomes are measured quarterly. (D) Authors using LLMs cite 11.9% more books post-adoption, showcasing LLM's advanced ability to process and integrate a more diverse range of knowledge sources. Estimates are from Poisson regressions. (E) Authors using LLMs cite more recent references post-adoption, with the median reference age decreasing by 0.379 years. (F) Contrary to concerns that LLMs may reinforce reliance on well-established scientific works, we find no increase in citation to high-impact works, where "reference impact" is measured by log(#citations+1) averaged over all cited references. (G) Authors using LLMs on arXiv, bioRxiv, and SSRN cite 14.1% ($P < 0.001$), 7.81% ($P = 0.011$), and 2.50% ($P = 0.605$) more books post-adoption, respectively. (H) Authors using LLMs on arXiv, bioRxiv, and SSRN cite references that are 0.436 years ($P < 0.001$), 0.222 years ($P < 0.001$), and 0.422 years ($P = 0.023$) more recent, respectively. (I) Authors using LLMs on arXiv, bioRxiv, and SSRN cite references that are 4.29% ($P < 0.001$), 0.800% ($P = 0.012$), and 0.916% ($P = 0.369$) less highly cited, respectively. Panel A and B are identical to the figures shown in the main manuscript (the second figure).

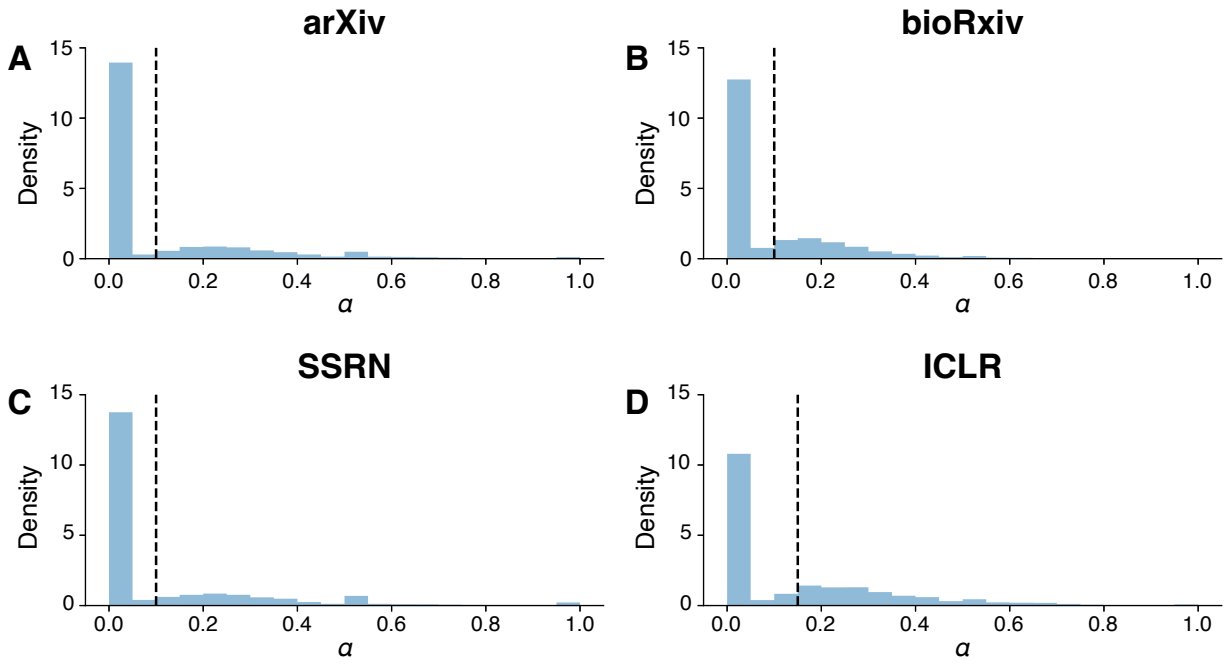


Figure S5: **Distribution of α for post-ChatGPT manuscripts in 2023 and 2024.** The dashed lines indicate the threshold α_0.

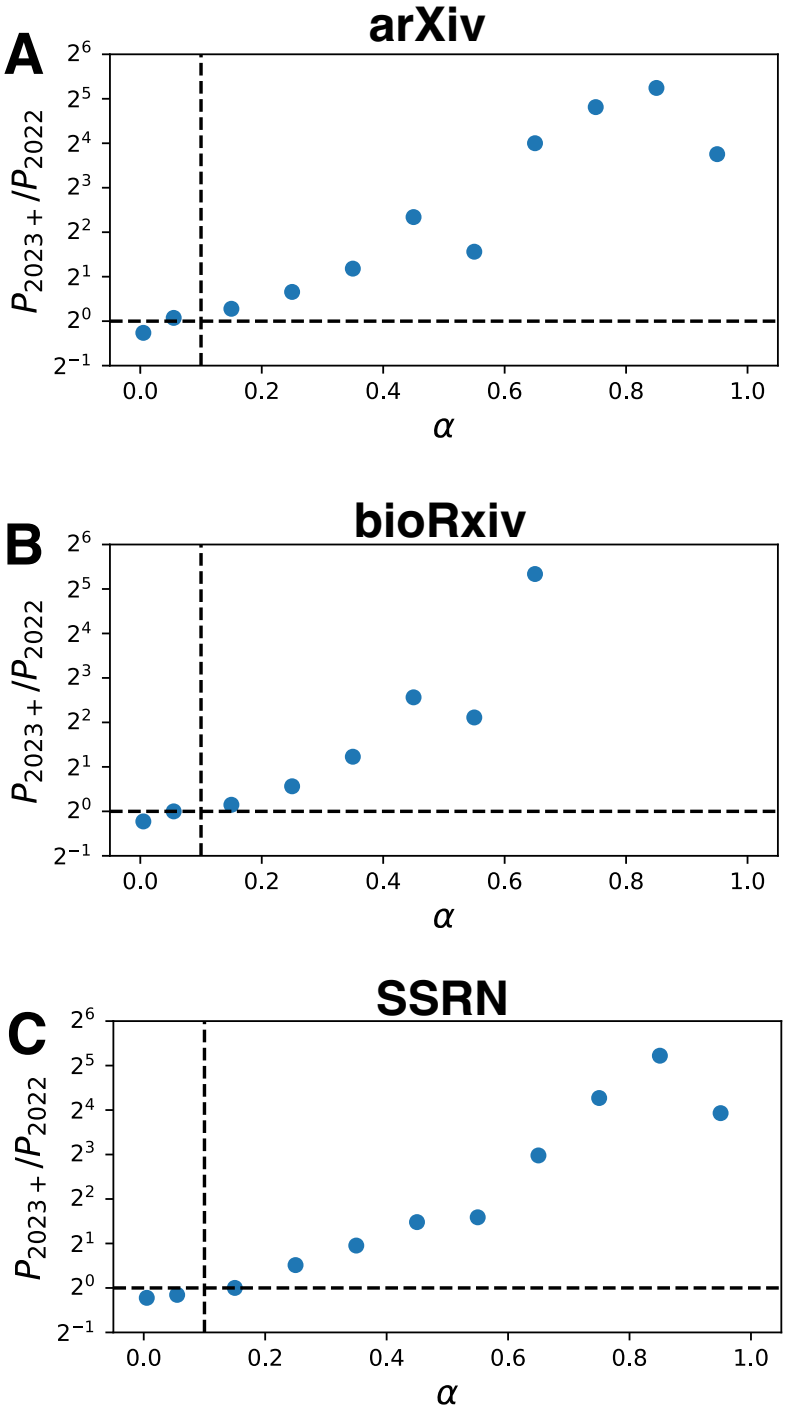


Figure S6: **Comparing the distribution of α between papers published before / after 2023**.

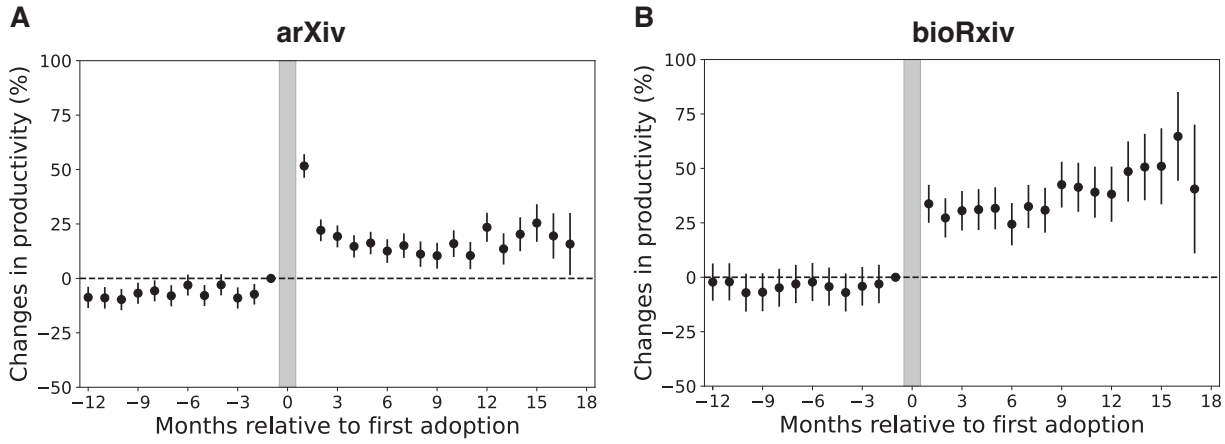


Figure S7: **Changes in scientific productivity** on arXiv (A) and bioRxiv (B) using OpenAlex author id.

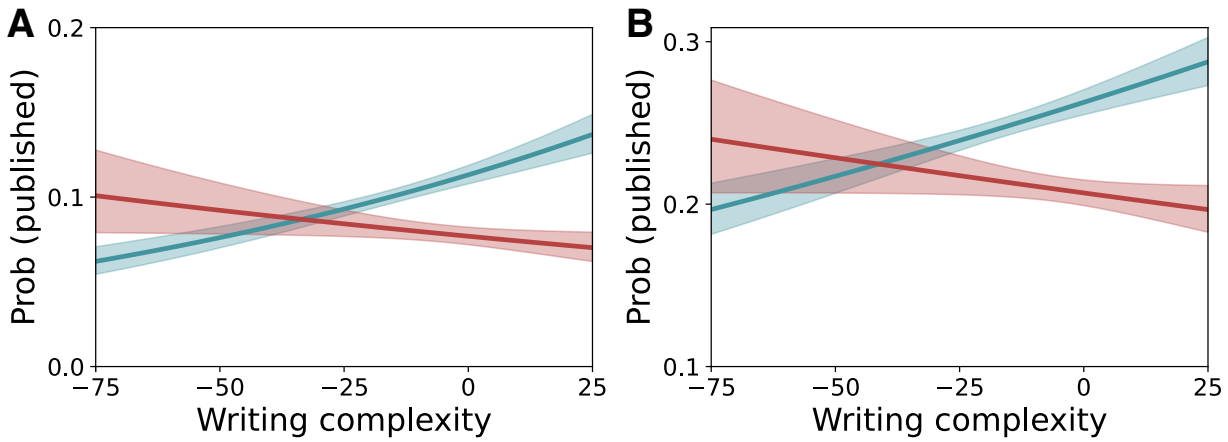


Figure S8: **Writing complexity and publication outcomes of SSRN preprints.** We find qualitatively consistent results using exact matches only (A) or including fuzzy matches (B).

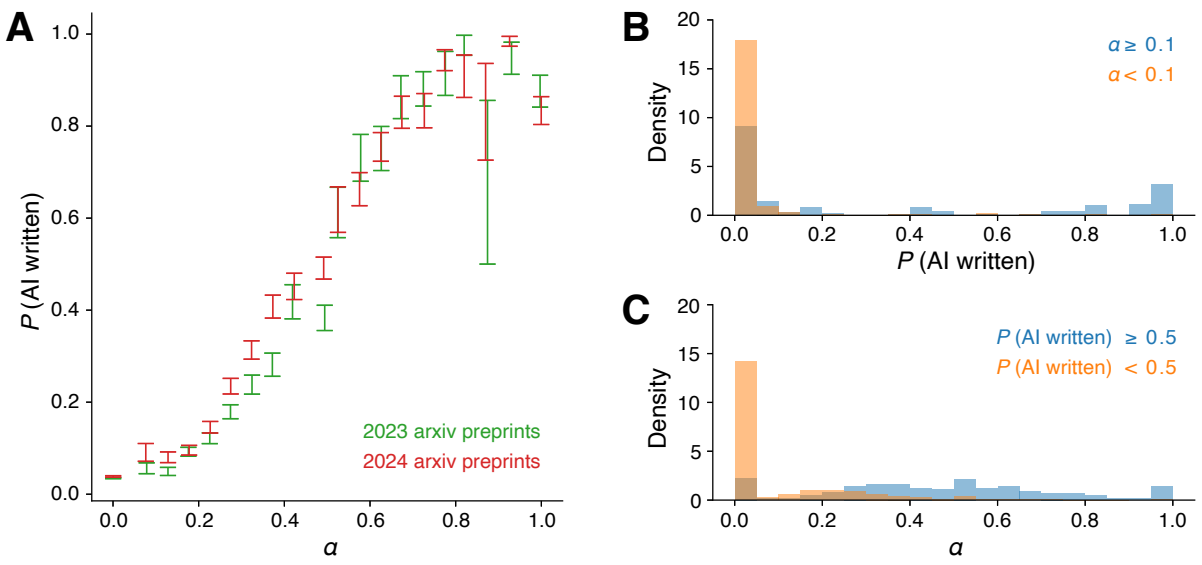


Figure S9: **Comparing α and GPTZero on arXiv data.**

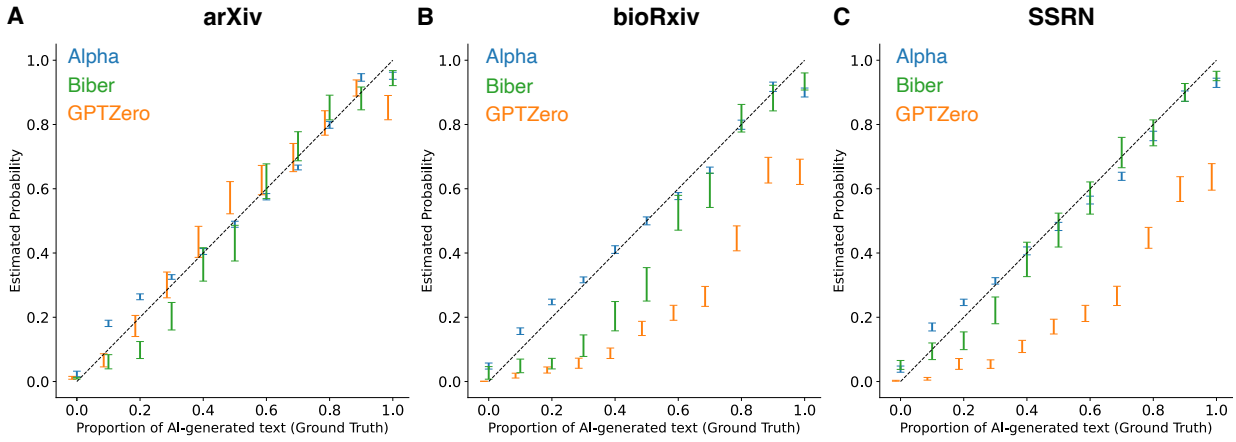


Figure S10: **Comparing α, Biber, and GPTZero on synthetic test samples composed of varying proportions of human- and LLM-generated text.** GPTZero and Biber systematically underestimate partial uses of LLM.

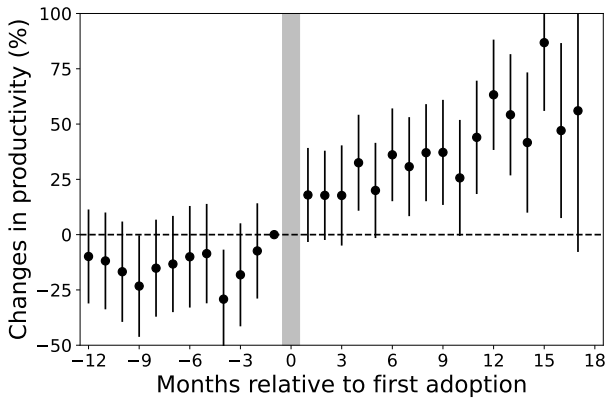


Figure S11: **Changes in scientific productivity**, where LLM usage is detected by GPTZero.

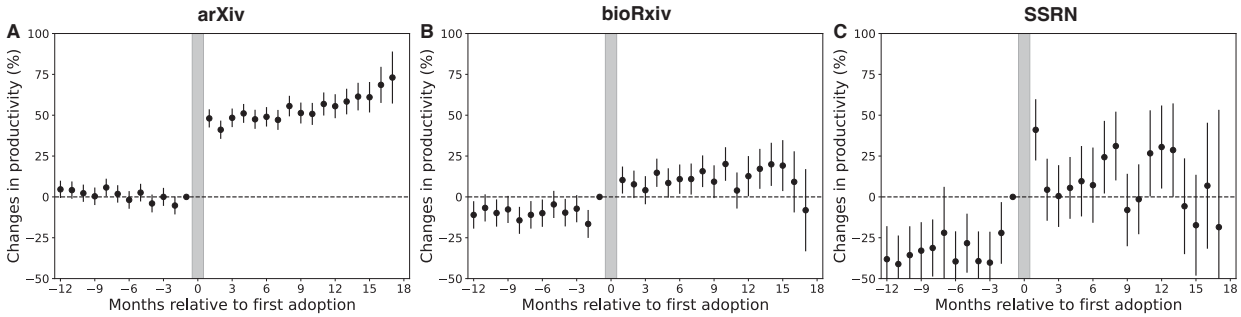


Figure S12: **Changes in scientific productivity**, where LLM usage is detected by Biber method.

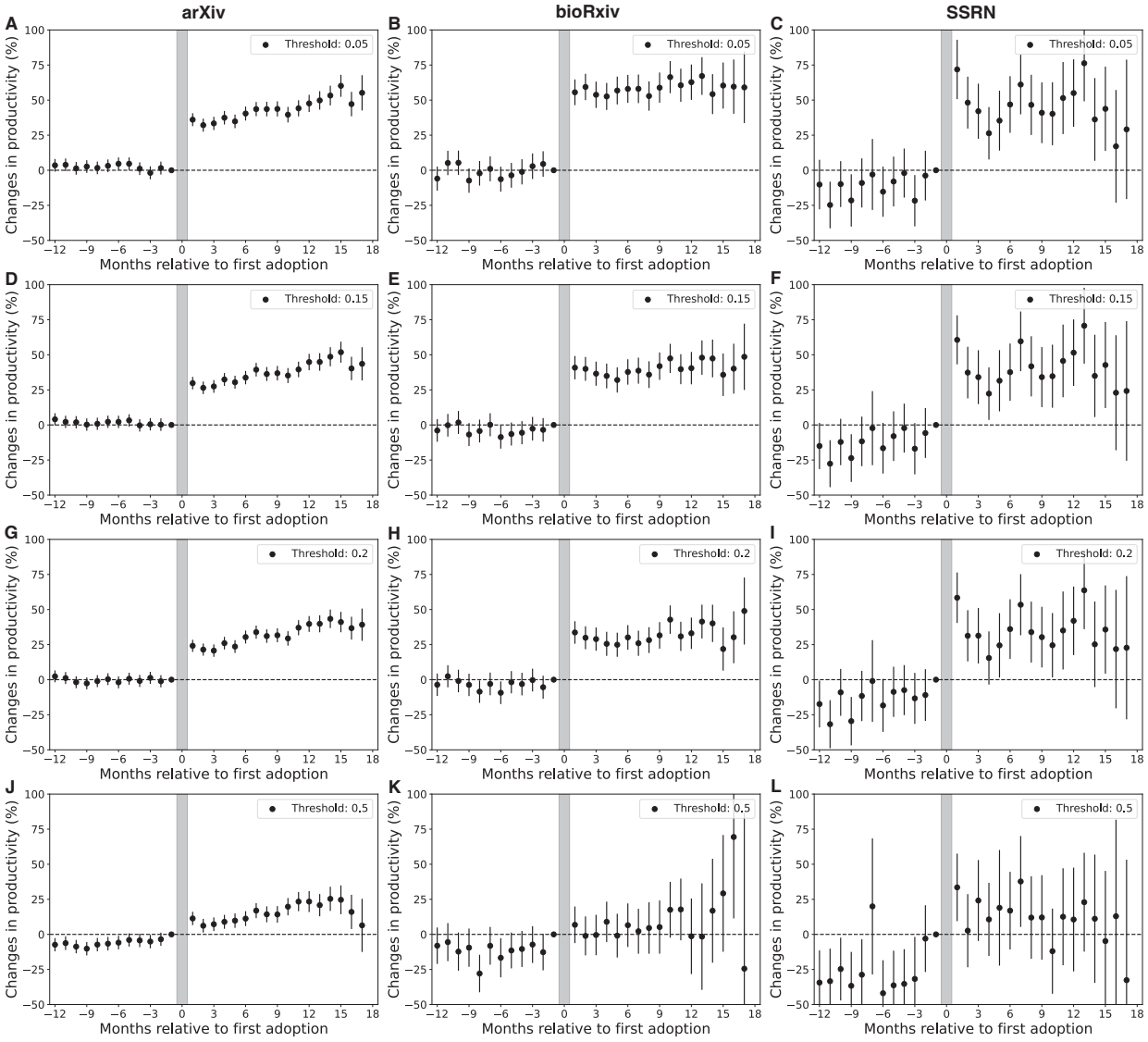


Figure S13: **Changes in scientific productivity** with different thresholds.

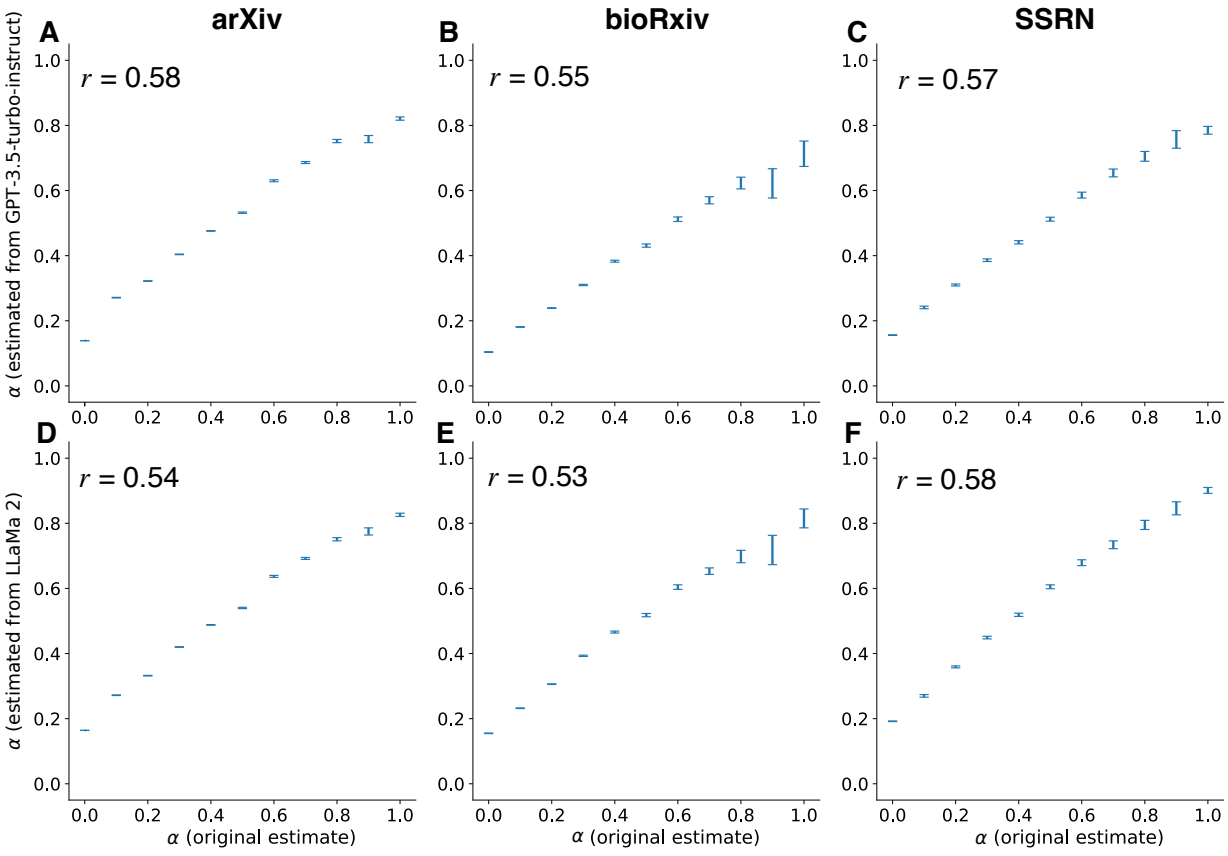


Figure S14: **Correlation between original α and α estimated by GPT-3.5-turbo-instruct and LLaMa 2.**

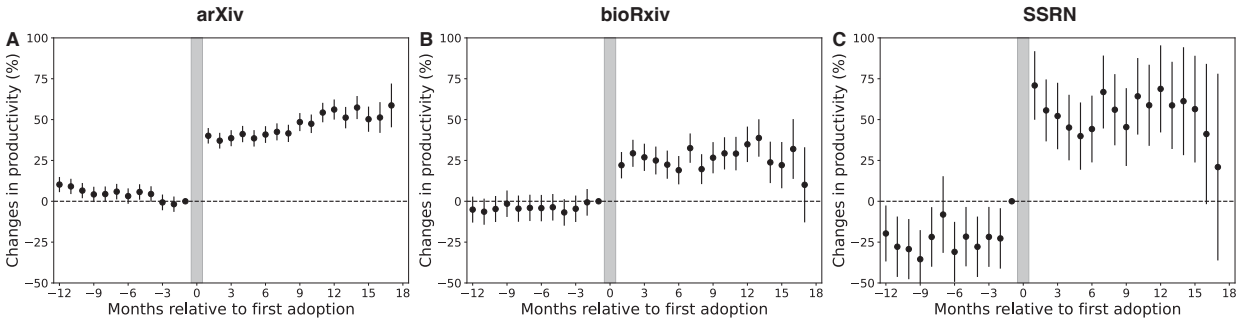


Figure S15: **Changes in scientific productivity**, where LLM usage is detected by alpha estimated using GPT-3.5-turbo-instruct.

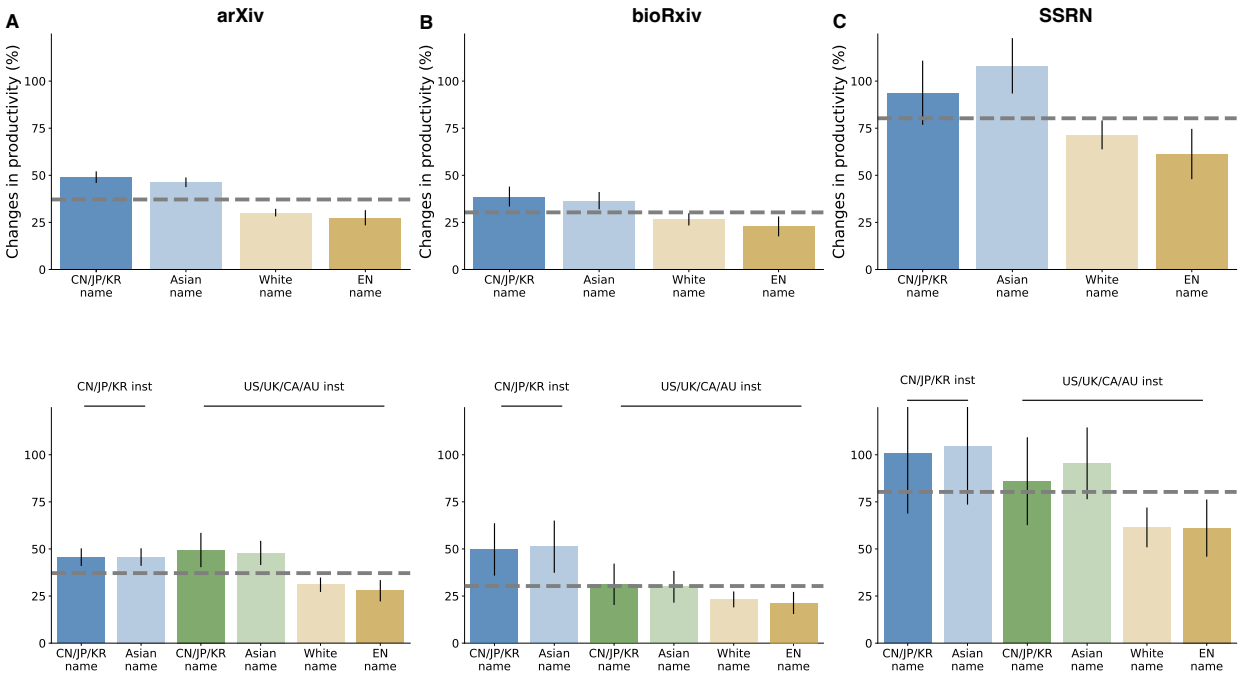


Figure S16: **Heterogeneity across races, ethnicities, and home geographies**, where LLM usage is detected by alpha estimated using GPT-3.5-turbo-instruct.

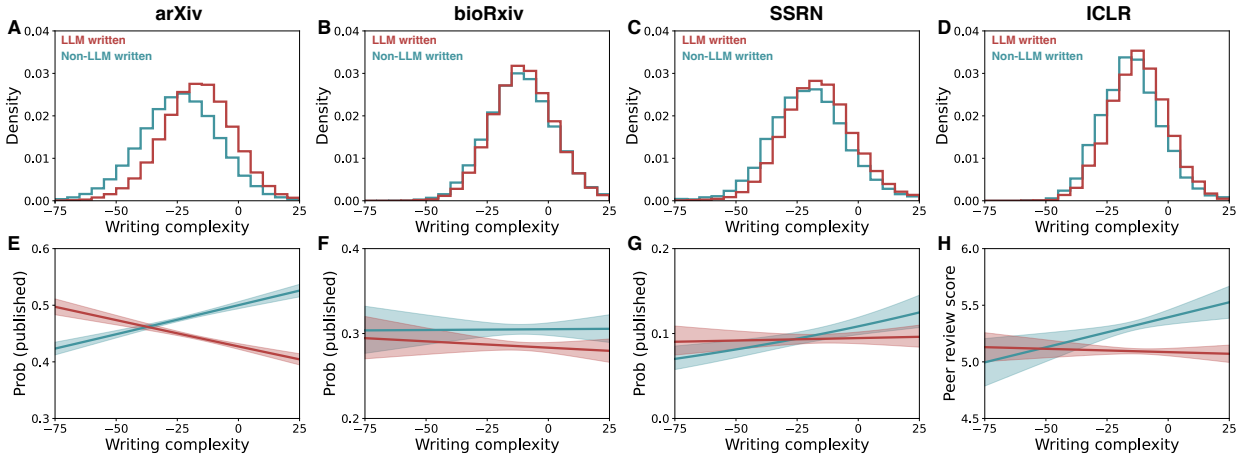


Figure S17: **LLM usage, scientific writing, and publication outcomes**, where LLM usage is detected by alpha estimated using GPT-3.5-turbo-instruct.

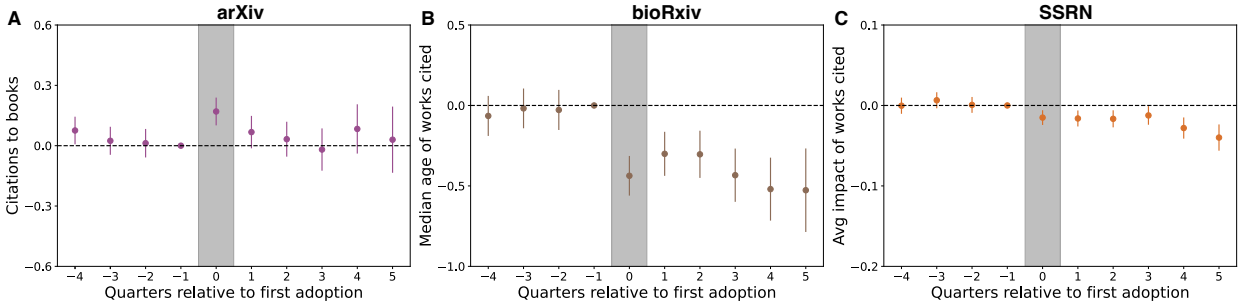


Figure S18: **LLM usage and references to prior works**, where LLM usage is detected by alpha estimated using GPT-3.5-turbo-instruct.

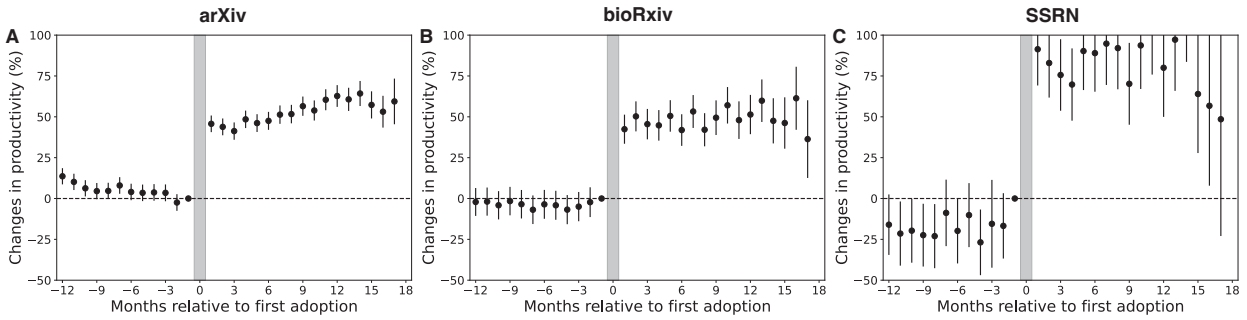


Figure S19: **Changes in scientific productivity**, where LLM usage is detected by alpha estimated using LLaMA 2.

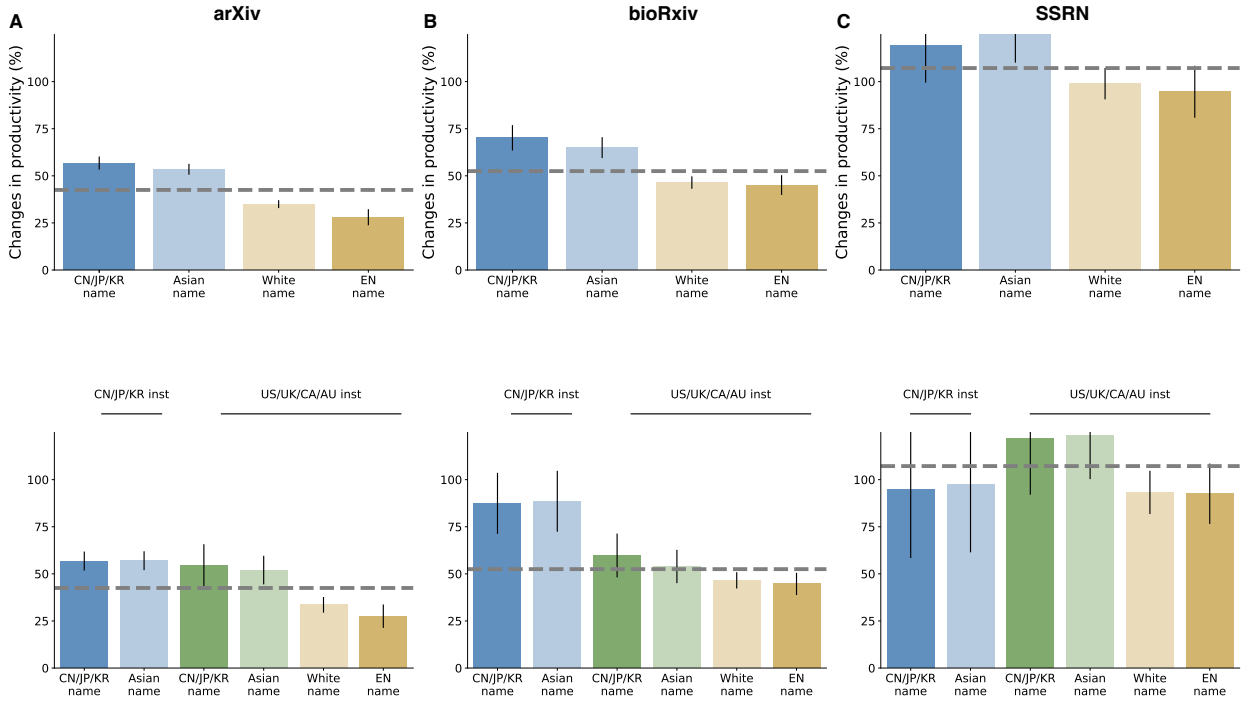


Figure S20: **Heterogeneity across races, ethnicities, and home geographies**, where LLM usage is detected by alpha estimated using LLaMA 2.

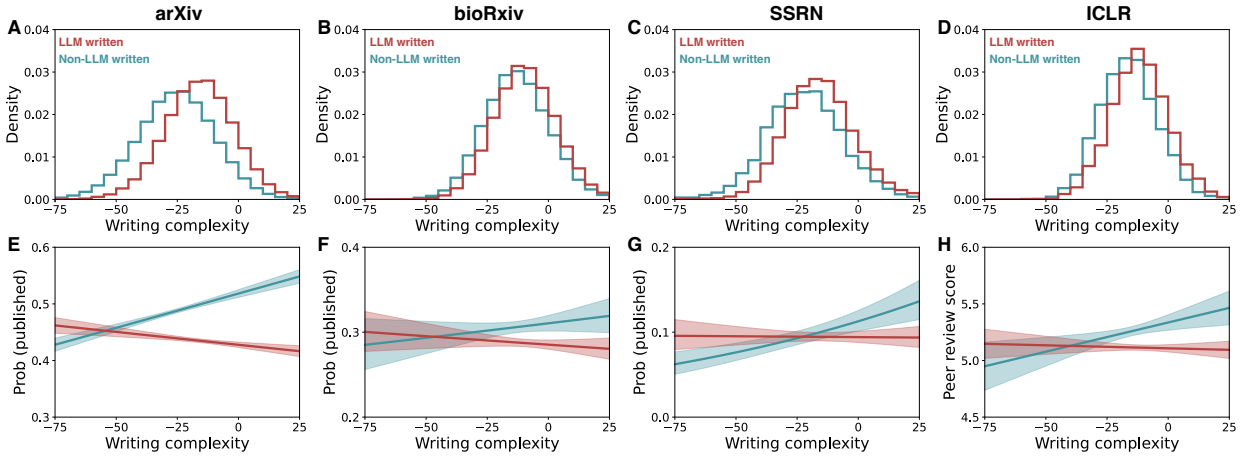


Figure S21: **LLM usage, scientific writing, and publication outcomes**, where LLM usage is detected by alpha estimated using LLaMA 2.

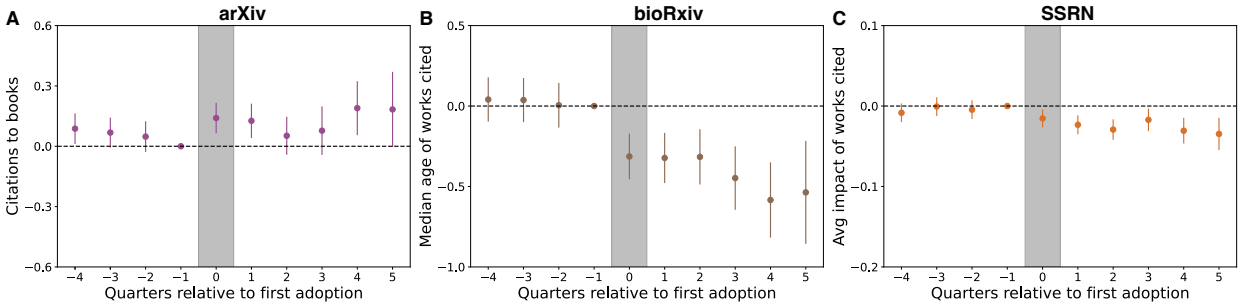


Figure S22: **LLM usage and references to prior works)**, where LLM usage is detected by alpha estimated using LLaMA 2.

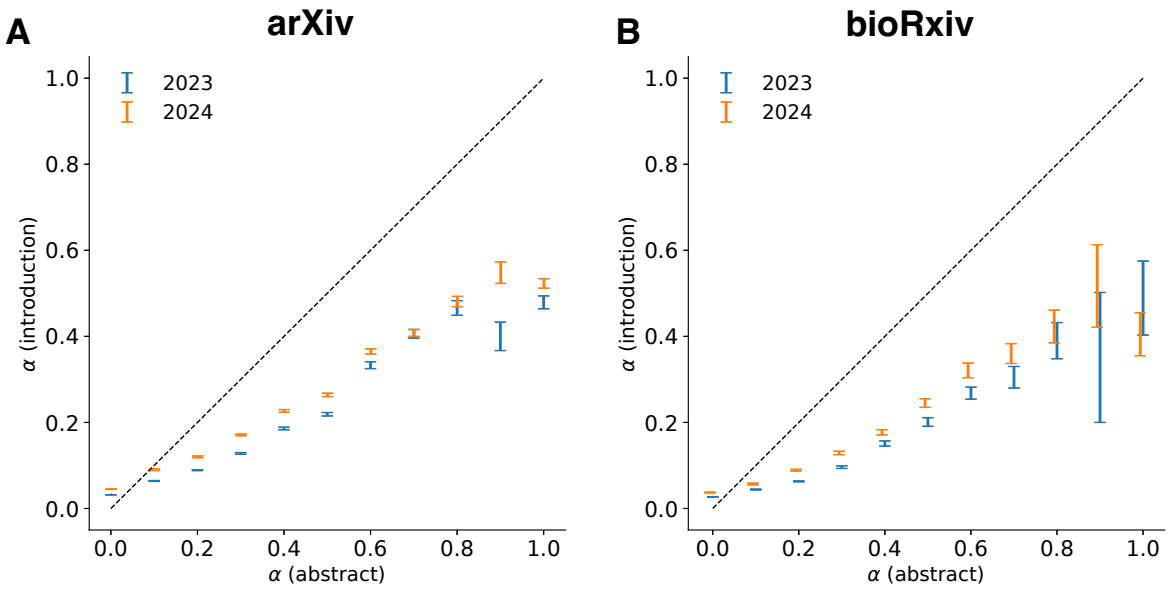


Figure S23: **Correlation of α measured between abstract and introduction in arXiv (A) and bioRxiv (B).**

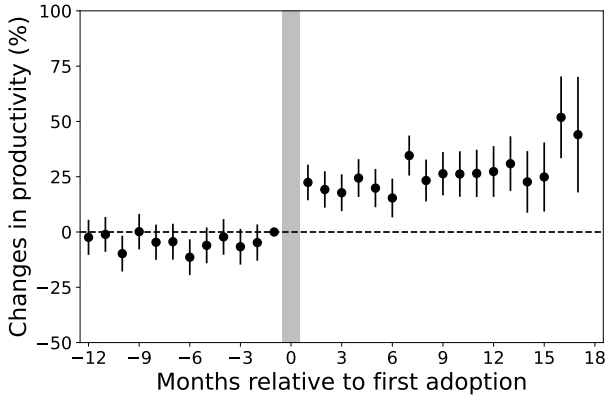


Figure S24: **Changes in scientific productivity**, where α is estimated by introduction texts in bioRxiv.

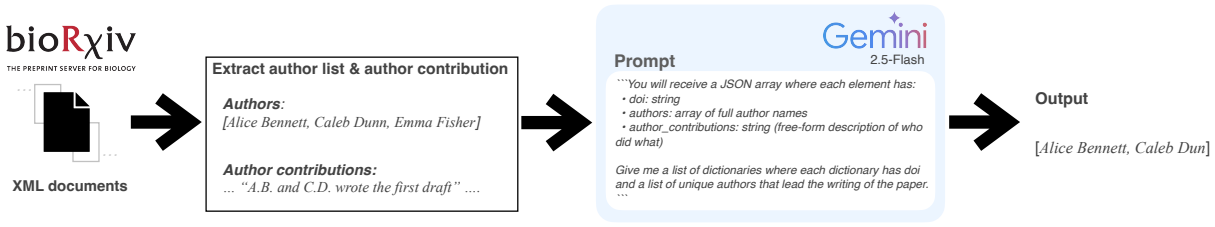


Figure S25: **Identifying writing authors from bioRxiv author contribution statements.**

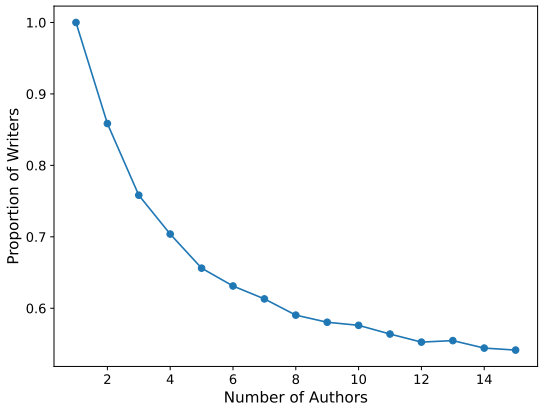


Figure S26: **The fraction of contributing-to-writing authors across team sizes**.

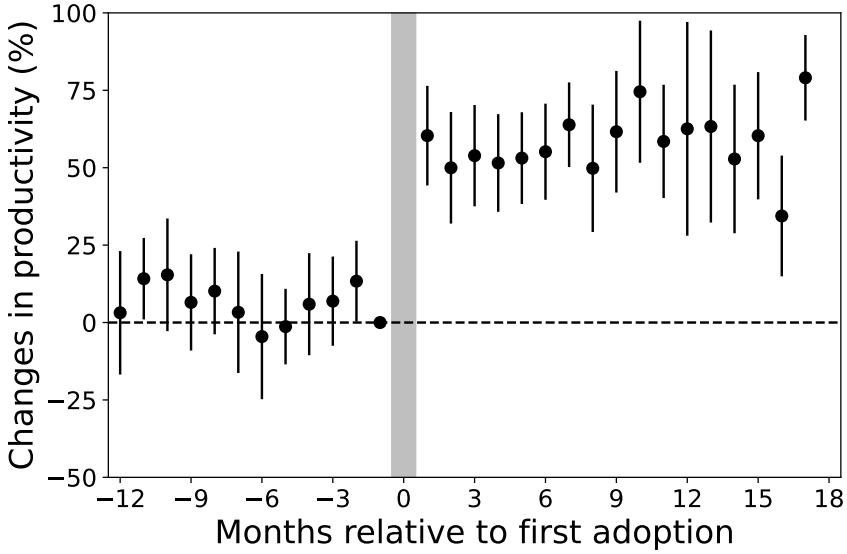


Figure S27: **Robustness checks on productivity dynamics.** . We restricted the treated group to include only researchers who are tagged as principal contributors (bioRxiv) to manuscript writing on their first LLM-assisted paper.

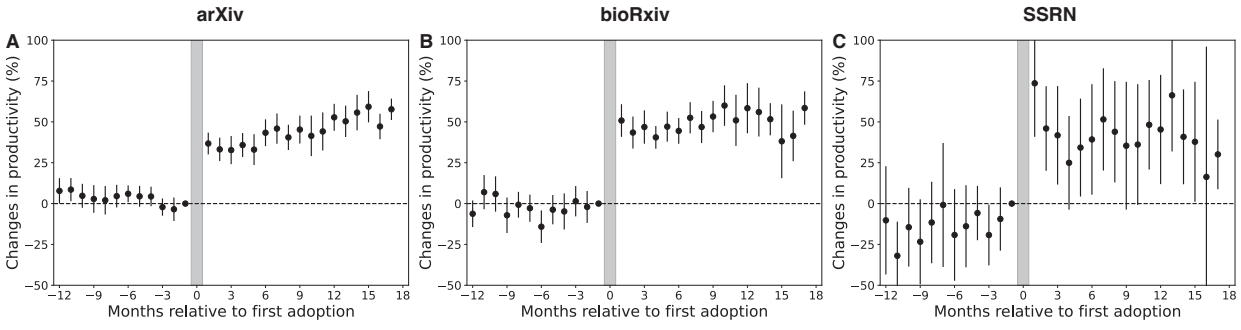


Figure S28: **Robustness checks on productivity dynamics.** . We restricted the treated group to include only researchers who are listed as first or last authors on their first LLM-assisted paper.

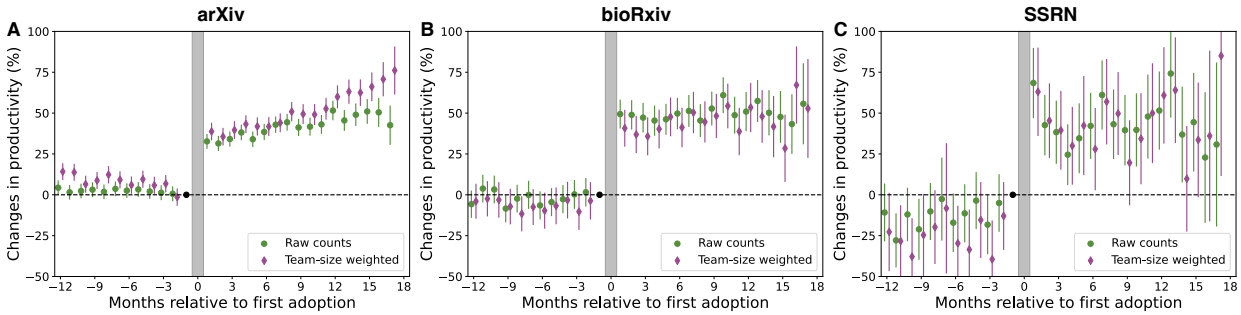


Figure S29: **Changes in scientific productivity**, measured by raw counts (green) and inverse team-size weighting (purple).

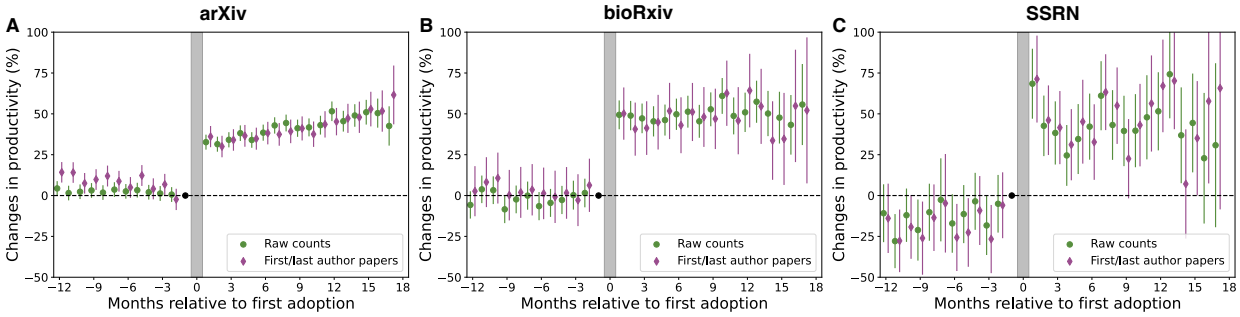


Figure S30: **Changes in scientific productivity**, measured by all papers (green) and first / last author papers (purple).

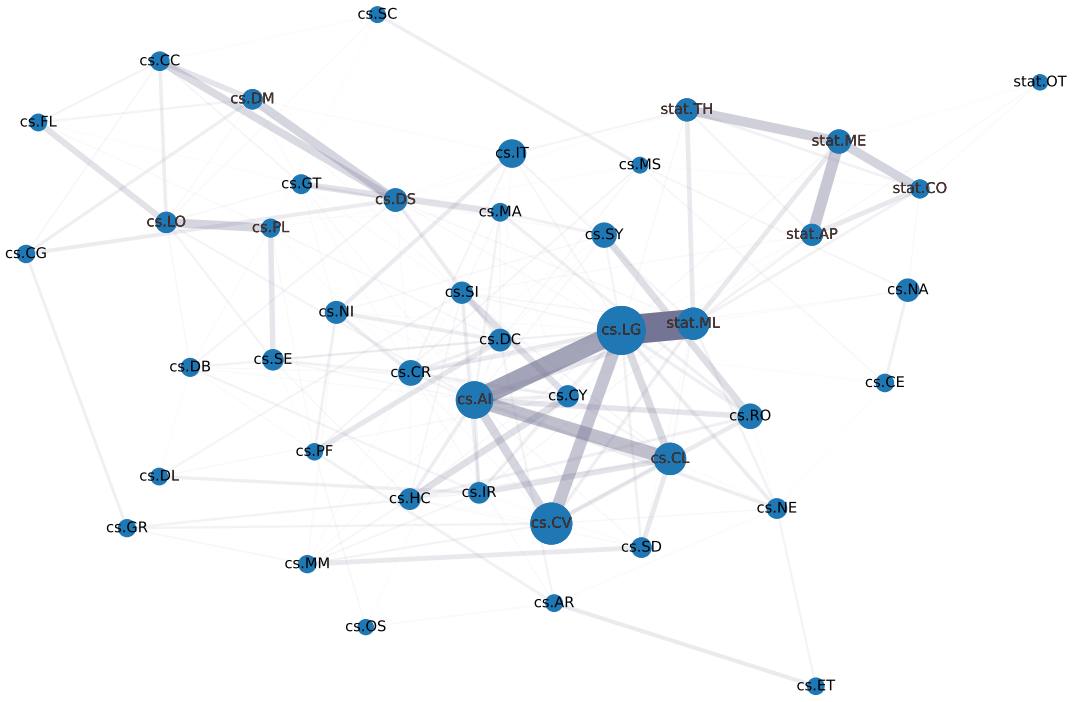


Figure S31: **Network of arXiv CS and Stats fields.** Nodes represent fields, and edge weights indicate the frequency with which two fields are co-tagged in the same paper.

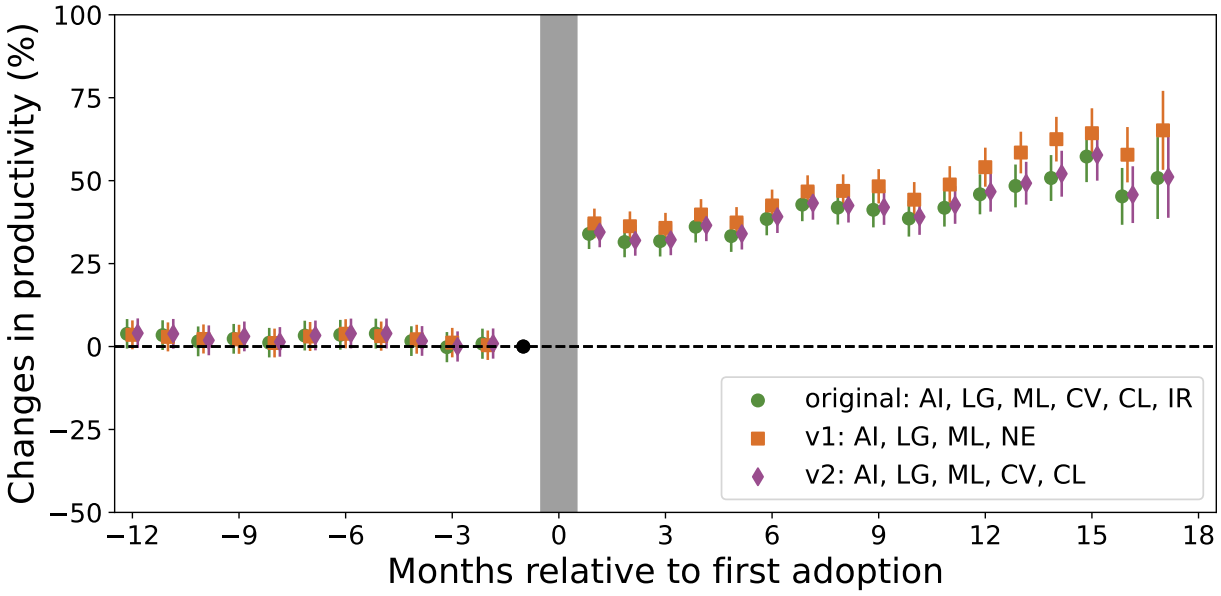


Figure S32: **LLM usage and scientific productivity**, where we exclude 'AI-related' fields defined as (1) AI, LG, ML, CV, CL, and IR, (2) AI, LG, ML, and NE, and (3) AI, LG, ML, CV, and CL.

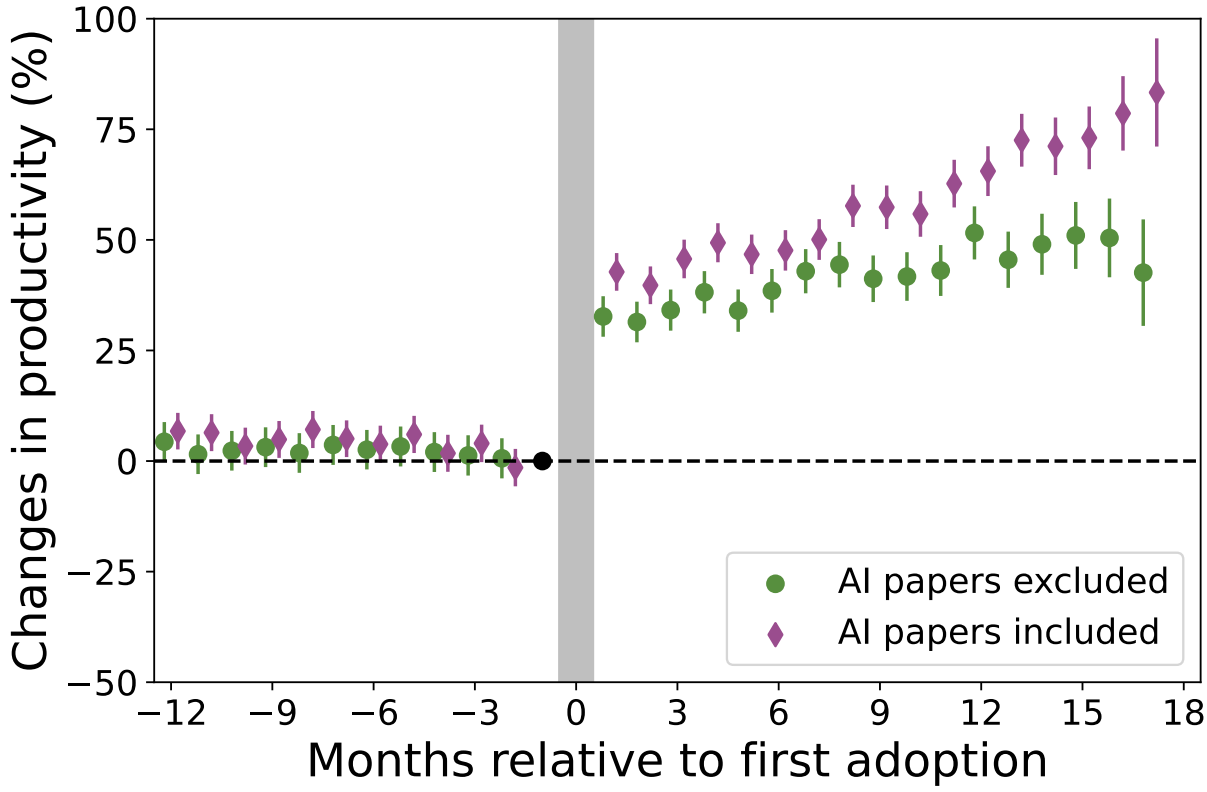


Figure S33: **Changes in scientific productivity**, with AI papers included (purple) or excluded (green).

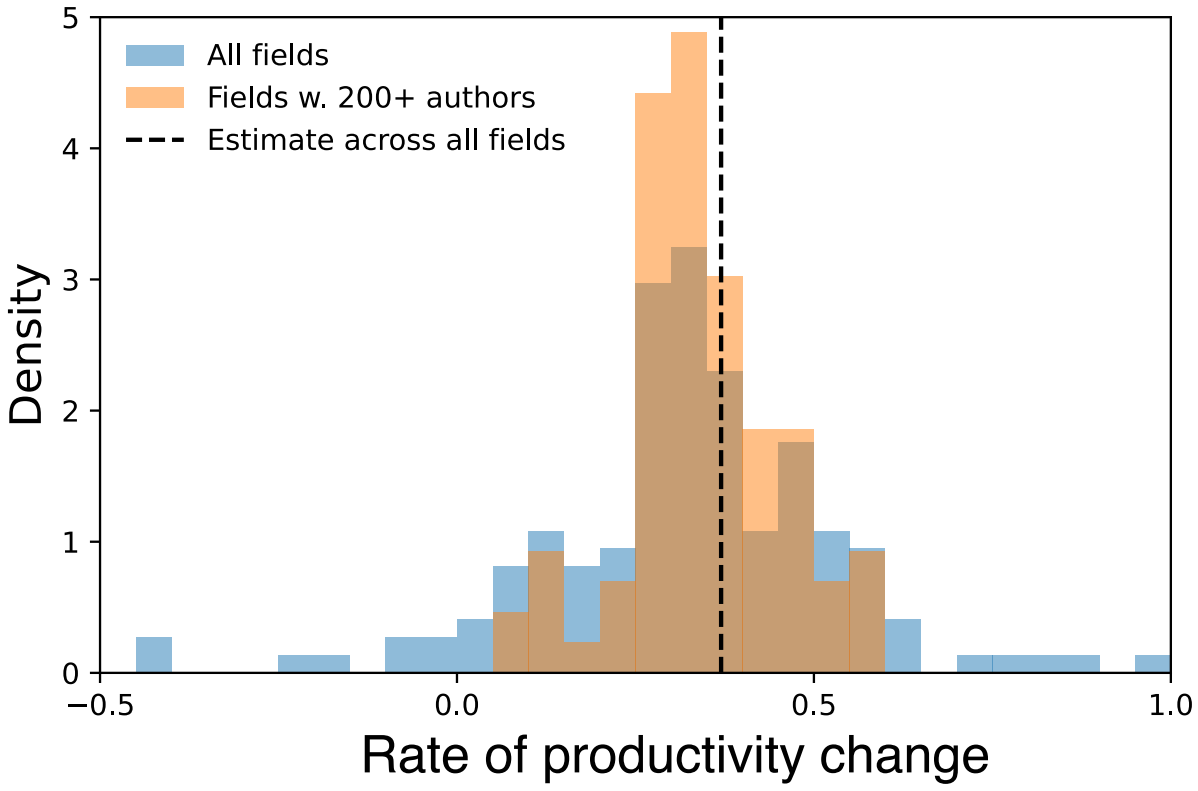


Figure S34: **Rate of productivity changes across 148 fields in arXiv**.

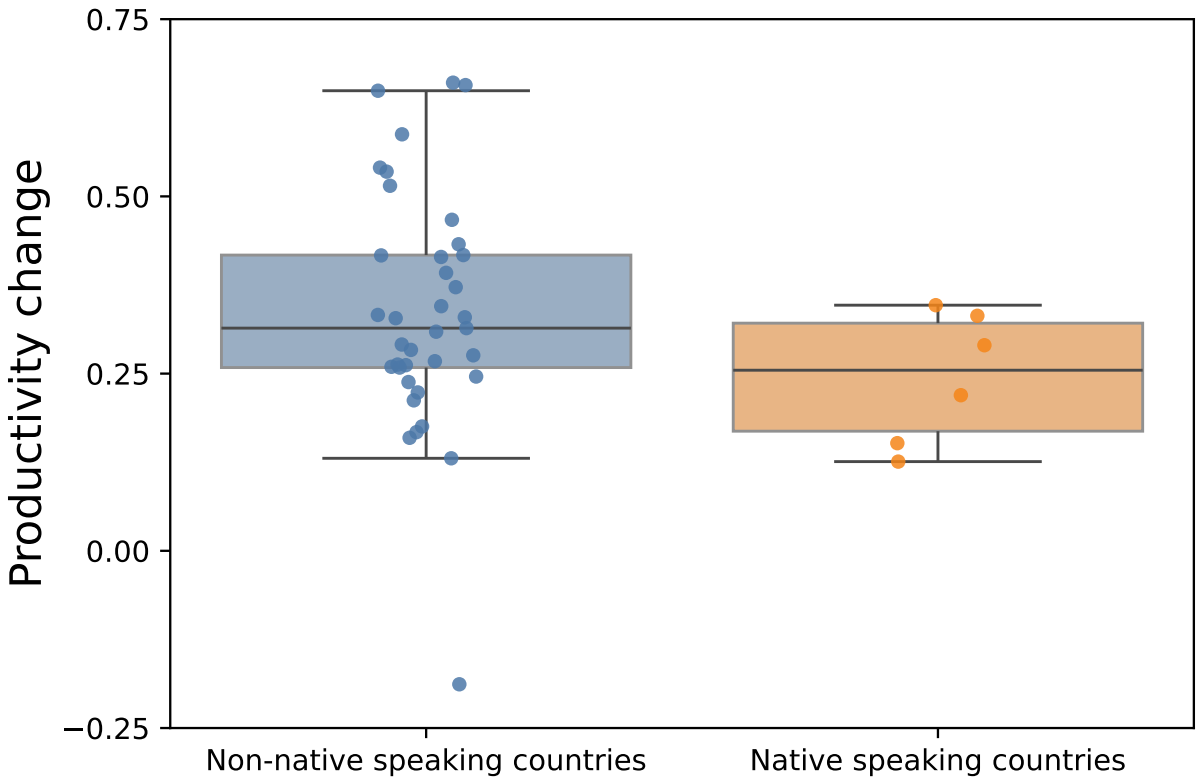


Figure S35: **Productivity changes in native and non-native English-speaking countries**.

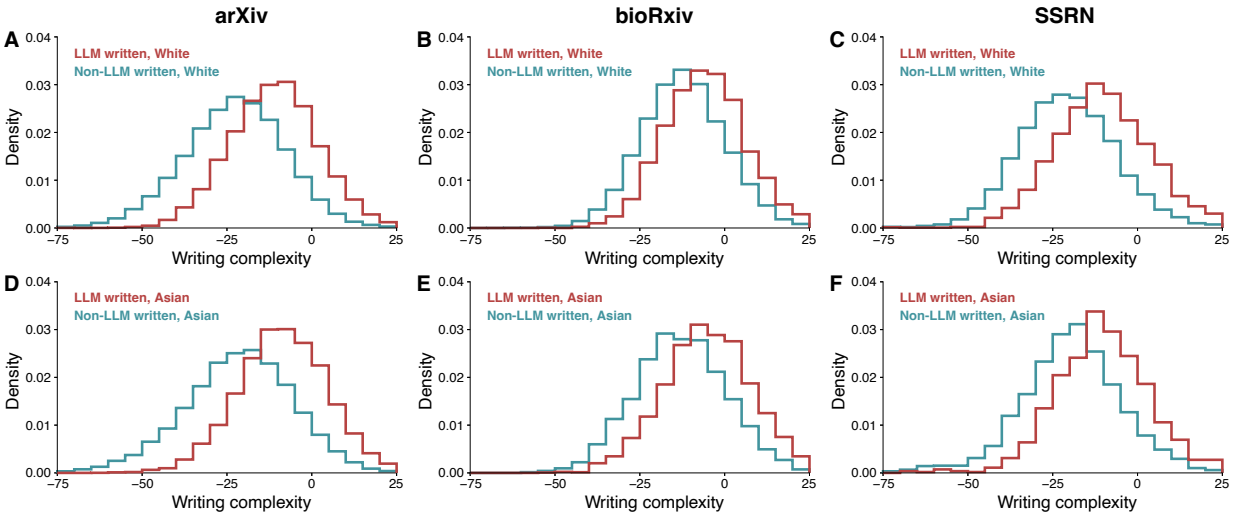


Figure S36: **Distribution of writing complexity.** We replicate the measures in Fig. 3A-C, separately for White authors (A-C) and Asian authors (D-F).

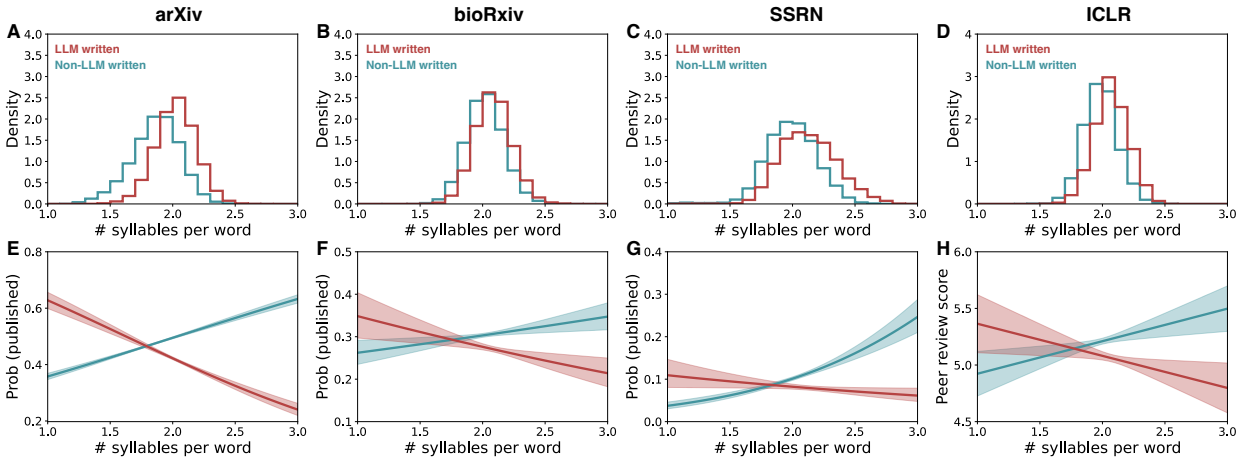


Figure S37: **LLM usage, lexical complexity, and publication outcomes.**

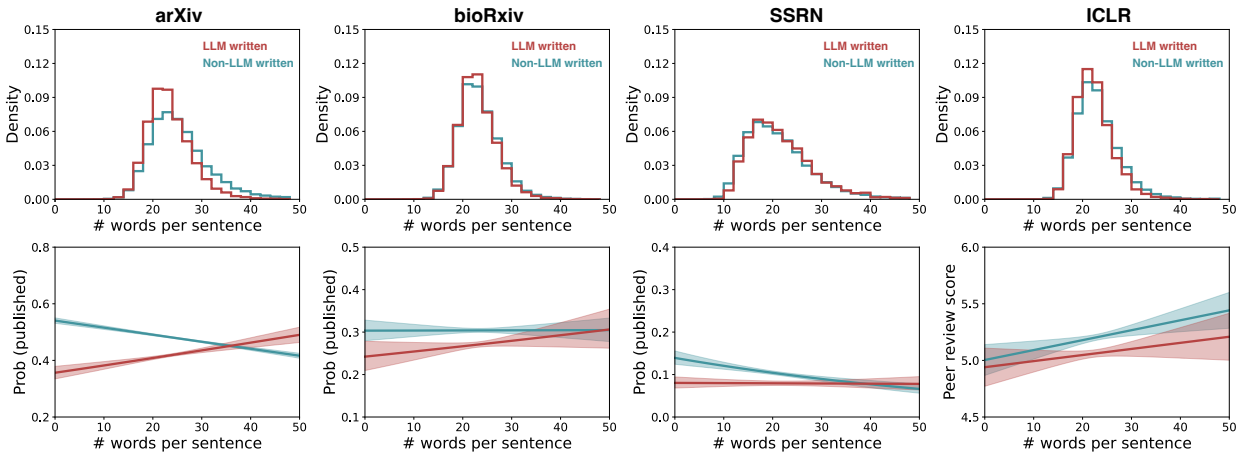


Figure S38: **LLM usage, syntactic complexity, and publication outcomes.**

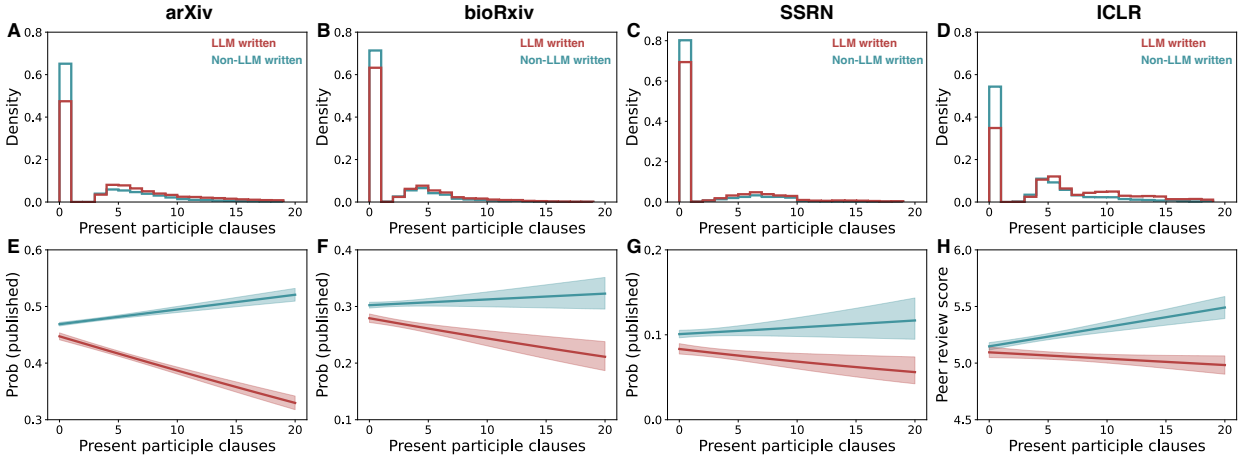


Figure S39: **LLM usage, morphological complexity, and publication outcomes.**

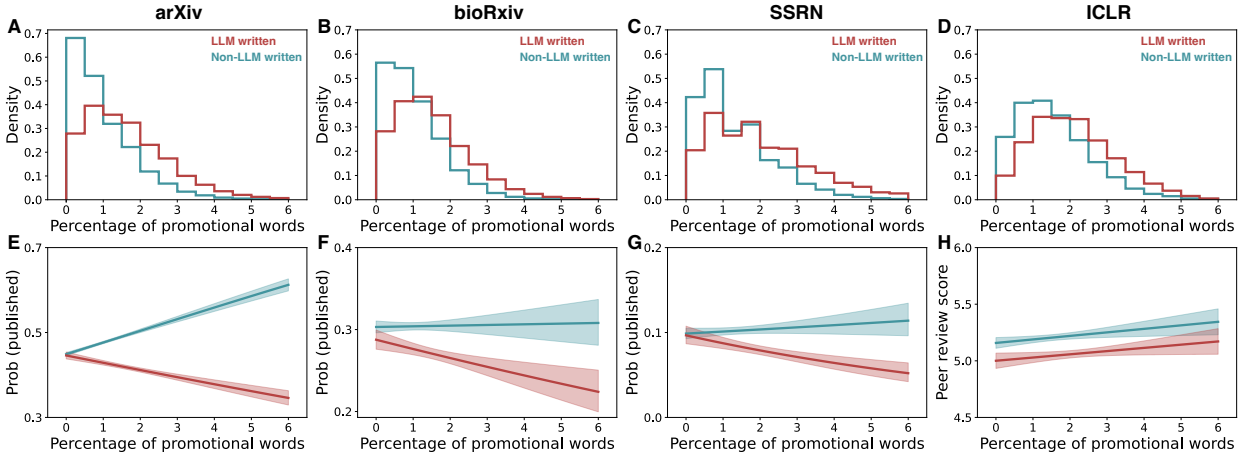


Figure S40: **LLM usage, promotional language, and publication outcomes.**

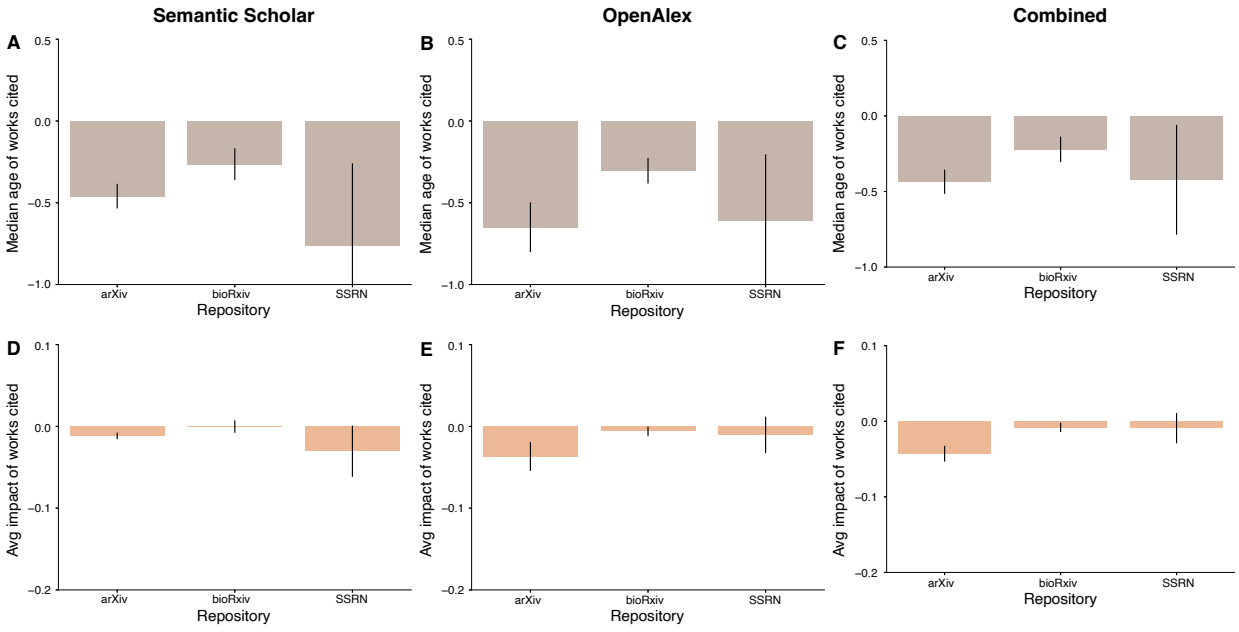


Figure S41: **Robustness checks on citation analysis.** We replicate the results in Fig. 4D,F, using data from Semantic Scholar (A, D), OpenAlex (B, E), and a combination of both sources (C, F).

Data source	arXiv	bioRxiv	SSRN
OpenAlex	12.26%	91.98%	40.61%
Semantic Scholar	91.45%	77.54%	35.45%
Combined	92.15%	98.09%	48.01%

Table S1: **Percentage of papers with reference data.**

Category	Type	arXiv	bioRxiv	SSRN
Race	Asian	32.9%	28.1%	47.1%
	Hispanic	8.6%	10.0%	7.5%
	Black	2.9%	4.0%	4.0%
	White	55.1%	57.4%	40.9%
	Others	0.4%	0.39%	0.53%
Ethnicity	Chinese	18.2%	14.0%	30.7%
	Japanese	3.8%	4.3%	3.0%
	Korean	2.0%	1.7%	2.4%
	English	14.8%	20.8%	13.1%

Table S2: **Distribution of races and ethnicity across datasets.**

Preprint title (SSRN)	**Published title**
Educational Job Mismatch, Job Satisfaction, On-the-Job Training, and Employee Quit **Behavior**: A Dynamic Analytical Approach	Educational job mismatch, job satisfaction, on-the-job training, and employee quit **behaviour**: a dynamic analytical approach
Investors' Responses to **Macro-Economic** News: The Role of Mandatory Derivatives and Hedging Activities Disclosure	Investors' responses to **macroeconomic** news: the role of mandatory derivatives and hedging activities disclosure
Freedom Not to See a Doctor: The Path **to** Over-the-Counter Abortion Pills	Freedom Not to See a Doctor: The Path **Toward** Over-The-Counter Abortion Pills

Table S3: **Examples of mismatching titles.**

	Published (Logistic regression)						**Rating** (OLS)
	arXiv		**bioRxiv**		**SSRN**		**ICLR**
LLM written	-0.432***	-0.233***	-0.214***	-0.083***	-0.425***	-0.197**	-0.243***
	(0.016)	(0.019)	(0.024)	(0.025)	(0.046)	(0.048)	(0.003)
Writing complexity	0.004***	0.002**	0.002**	0.002**	0.008***	0.006**	0.005***
	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
LLM written ×	-0.013***	-0.004***	-0.006***	-0.004**	-0.013***	-0.009***	-0.008***
Writing complexity	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
Month F.E.		Y		Y		Y	
Field F.E.		Y		Y		Y	

Table S4: **Robustness checks on writing complexity and publication outcomes.**

Threshold α_0	arXiv	bioRxiv	SSRN
0.05	0.371	0.598	0.612
0.1 (original)	0.362	0.529	0.598
0.15	0.325	0.417	0.564
0.2	0.297	0.330	0.532
0.5	0.201	0.158	0.425

Table S5: **Estimated productivity boosts with different thresholds.**

Threshold α_0	arXiv	bioRxiv	SSRN
$\alpha_{0,L} = \alpha_{0,H} = 0.1$ (original)	0.362	0.529	0.598
$\alpha_{0,L} = 0.1$, $\alpha_{0,H} = 0.15$	0.353	0.532	0.596
$\alpha_{0,L} = 0.1$, $\alpha_{0,H} = 0.2$	0.356	0.548	0.670
$\alpha_{0,L} = 0.1$, $\alpha_{0,H} = 0.5$	0.375	0.637	0.719

Table S6: **Performance across different asymmetric threshold values $(\alpha_{0,L}, \alpha_{0,H})$ for three repositories.**

Subgroup	arXiv	bioRxiv	SSRN
Less experienced	0.444 (0.015)	0.570 (0.029)	0.631 (0.059)
More experienced	0.339 (0.009)	0.520 (0.016)	0.600 (0.045)

Table S7: **Estimated average productivity change for less and more experienced subgroups**.

	arXiv			bioRxiv			SSRN		
LLM written	0.076**	0.042	0.018	0.071***	0.014	0.006	0.245***	0.068*	0.047
	(0.033)	(0.056)	(0.046)	(0.018)	(0.036)	(0.023)	(0.031)	(0.037)	(0.038)
Month F.E.	Y	Y	Y	Y	Y	Y	Y	Y	Y
Field F.E.	Y	Y	Y	Y	Y	Y	Y	Y	Y
# non-book refs	Y	Y	Y	Y	Y	Y	Y	Y	Y
Team size F.E.		Y	Y		Y	Y		Y	Y
First author F.E.		Y			Y			Y	
Last author F.E.			Y			Y			Y

Table S8: **Paper-level Poisson regressions on citations to books.**

	arXiv			bioRxiv			SSRN		
LLM written	-1.646***	-0.495***	-0.468***	-0.551***	-0.248**	-0.269***	-1.424***	-0.322	-0.617**
	(0.039)	(0.064)	(0.048)	(0.032)	(0.084)	(0.046)	(0.104)	(0.223)	(0.227)
Month F.E.	Y	Y	Y	Y	Y	Y	Y	Y	Y
Field F.E.	Y	Y	Y	Y	Y	Y	Y	Y	Y
Team size F.E.		Y	Y		Y	Y		Y	Y
First author F.E.		Y			Y			Y	
Last author F.E.			Y			Y			Y

Table S9: **Paper-level OLS regressions on median reference age.**

	arXiv			bioRxiv			SSRN		
LLM written	0.022***	-0.010**	-0.006	-0.004	-0.013	-0.008	-0.054**	-0.006	-0.017
	(0.003)	(0.005)	(0.005)	(0.007)	(0.023)	(0.013)	(0.017)	(0.038)	(0.038)
Month F.E.	Y	Y	Y	Y	Y	Y	Y	Y	Y
Field F.E.	Y	Y	Y	Y	Y	Y	Y	Y	Y
Team size F.E.		Y	Y		Y	Y		Y	Y
First author F.E.		Y			Y			Y	
Last author F.E.			Y			Y			Y

Table S10: **Paper-level OLS regressions on average reference impact.**

S1. N. Rasmussen, *Picture Control: The Electron Microscope and the Transformation of Biology in America, 1940–1960* (Stanford Univ. Press) (1999).

S2. W. W. Ding, S. G. Levin, P. E. Stephan, A. E. Winkler, The impact of information technology on academic scientists' productivity and collaboration patterns. *Manage. Sci.* **56**, 1439–1461 (2010). doi:10.1287/mnsc.1100.1195

S3. J. L. Furman, F. Teodoridis, Automation, research technology, and researchers' trajectories: Evidence from computer science and electrical engineering. *Organ. Sci.* **31**, 330–354 (2020). doi:10.1287/orsc.2019.1308

S4. D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, E. R. Mardis, The next-generation sequencing revolution and its impact on genomics. *Cell* **155**, 27–38 (2013). doi:10.1016/j.cell.2013.09.006 Medline

S5. H. Bao, M. Sun, M. Teplitskiy, Where there's a will there's a way: ChatGPT is used more for science in countries where it is prohibited. *Quant. Sci. Stud.* **6**, 1–16 (2025). doi:10.1162/qss_a_00368

S6. D. Kobak, R. González-Márquez, E.-A. Horvát, J. Lause, Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Sci. Adv.* **11**, eadt3813 (2025). doi:10.1126/sciadv.adt3813 Medline

S7. W. Zhu, L. W. Cong, "Divergent LLM Adoption and Heterogeneous Convergence Paths in Research Writing," *Cornell SC Johnson College of Business Research Paper Forthcoming* (2024).

S8. H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, A. Anandkumar, K. Bergen, C. P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T.-Y. Liu, A. Manrai, D. Marks, B. Ramsundar, L. Song, J. Sun, J. Tang, P. Veličković, M. Welling, L. Zhang, C. W. Coley, Y. Bengio, M. Zitnik, Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023). doi:10.1038/s41586-023-06221-2 Medline

S9. E. Callaway, What's next for AlphaFold and the AI protein-folding revolution. *Nature* **604**, 234–238 (2022). doi:10.1038/d41586-022-00997-5 Medline

S10. A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, E. D. Cubuk, Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023). doi:10.1038/s41586-023-06735-9 Medline

S11. B. Romera-Paredes, M. Barekatain, A. Novikov, M. Balog, M. P. Kumar, E. Dupont, F. J. R. Ruiz, J. S. Ellenberg, P. Wang, O. Fawzi, P. Kohli, A. Fawzi, Mathematical discoveries from program search with large language models. *Nature* **625**, 468–475 (2024). doi:10.1038/s41586-023-06924-6 Medline

S12. C. Si, D. Yang, T. Hashimoto, Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. arXiv:2409.04109 [cs.CL] (2024).

S13. J. Kim, B. Lee, Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. arXiv:2305.09620 [cs.CL] (2023).

S14. S. Lai, J. Kim, N. Kunievsky, Y. Potter, J. Evans, Biased AI improves human decision-making but reduces trust. arXiv:2508.09297 [cs.HC] (2025).

S15. C. Xie *et al*., Can large language model agents simulate human trust behavior? *Adv. Neural Inf. Process. Syst.* **37**, 15674–15729 (2024).

S16. S. Peng, E. Kalliamvakou, P. Cihon, M. Demirer, The impact of ai on developer productivity: Evidence from github copilot. arXiv:2302.06590 [cs.CE] (2023).

S17. D. N. McCloskey, The rhetoric of economics. *J. Econ. Lit.* **21**, 481–517 (1983).

S18. D. I. Hanauer, K. Englander, Quantifying the burden of writing research articles in a second language: Data from Mexican scientists. *Writ. Commun.* **28**, 403–416 (2011). doi:10.1177/0741088311420056

S19. V. Ramírez-Castañeda, Disadvantages in preparing and publishing scientific papers caused by the dominance of the English language in science: The case of Colombian researchers in biological sciences. *PLOS ONE* **15**, e0238372 (2020). doi:10.1371/journal.pone.0238372 Medline

S20. D. D. Belcher, Seeking acceptance in an English-only research world. *J. Second Lang. Writ.* **16**, 1–22 (2007). doi:10.1016/j.jslw.2006.12.001

S21. V. Berdejo-Espinola, T. Amano, AI tools can improve equity in science. *Science* **379**, 991–991 (2023). doi:10.1126/science.adg9714 Medline

S22. H. Hu, D. Wang, S. Deng, Analysis of the scientific literature's abstract writing style and citations. *Online Inf. Rev.* **45**, 1290–1305 (2021). doi:10.1108/OIR-05-2020-0188

S23. J. Feld, C. Lines, L. Ross, Writing matters. *J. Econ. Behav. Organ.* **217**, 378–397 (2024). doi:10.1016/j.jebo.2023.11.016

S24. W. Liang, Y. Zhang, H. Cao, B. Wang, D. Y. Ding, X. Yang, K. Vodrahalli, S. He, D. S. Smith, Y. Yin, D. A. McFarland, J. Zou, Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI* **1**, AIoa2400196 (2024). doi:10.1056/AIoa2400196

S25. E. A. M. van Dis, J. Bollen, W. Zuidema, R. van Rooij, C. L. Bockting, ChatGPT: Five priorities for research. *Nature* **614**, 224–226 (2023). doi:10.1038/d41586-023-00288-7 Medline

S26. B. C. Lee, J. J. Chung, An empirical investigation of the impact of ChatGPT on creativity. *Nat. Hum. Behav.* **8**, 1906–1914 (2024). doi:10.1038/s41562-024-01953-1 Medline

S27. A. Zeng, Z. Shen, J. Zhou, J. Wu, Y. Fan, Y. Wang, H. E. Stanley, The science of science: From the perspective of complex systems. *Phys. Rep.* **714–715**, 1–73 (2017). doi:10.1016/j.physrep.2017.10.001

S28. D. Crane, *Invisible Colleges: Diffusion of Knowledge in Scientific Communities* (Univ. Chicago Press, 1972).

S29. W. Liang *et al*., "Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews" in *Forty-first International Conference on Machine Learning* (2024).

S30. W. Liang *et al*., Mapping the increasing use of llms in scientific papers. arXiv:2404.01268 [cs.CL] (2024).

S31. N. R. Smalheiser, V. I. Torvik, Author name disambiguation. *Annu. Rev. Inform. Sci. Tech.* **43**, 1–43 (2009). doi:10.1002/aris.2009.1440430113

S32. R. Chintalapati, S. Laohaprapanon, G. Sood, Predicting race and ethnicity from the sequence of characters in a name. arXiv:1805.02109 [stat.AP](2018).

S33. H. R. Hatfield, H. Hao, M. Klein, J. Zhang, Y. Fu, J. Kim, J. Lee, S. J. G. Ahn, Addressing Whiteness in communication scholar composition and collaboration across seven decades of ICA journals (1951–2022). *J. Commun.* **74**, 451–465 (2024). doi:10.1093/joc/jqae019

S34. S. Nilforoushan, Q. Wu, M. Milani, "Entity Matching with AUC-Based Fairness" in *2022 IEEE International Conference on Big Data (Big Data)* (IEEE, 2022), pp. 5068–5075.

S35. J. W. Lockhart, M. M. King, C. Munsch, Name-based demographic inference and the unequal distribution of misrecognition. *Nat. Hum. Behav.* **7**, 1084–1095 (2023). doi:10.1038/s41562-023-01587-9 Medline

S36. T. Chumthong, K. Jitkajornwanich, O. Kraishan, K. F. Kee, A. Narabin, "Leveraging Race Prediction Algorithms to Enhance Team Composition in Big Data Science Teams" in *2024 IEEE International Conference on Big Data (BigData)* (IEEE, 2024), pp. 3105–3113.

S37. P. Treeratpituk, C. L. Giles, Name-ethnicity classification and ethnicity-sensitive name match-ing. *Proc. Conf. AAAI Artif. Intell.* **26**, 1141–1147 (2012). doi:10.1609/aaai.v26i1.8324

S38. R. Flesch, A new readability yardstick. *J. Appl. Psychol.* **32**, 221–233 (1948). doi:10.1037/h0057532 Medline

S39. C. Lu, Y. Bu, J. Wang, Y. Ding, V. Torvik, M. Schnaars, C. Zhang, Examining scientific writing styles from the perspective of linguistic complexity. *J. Assoc. Inf. Sci. Technol.* **70**, 462–475 (2019). doi:10.1002/asi.24126

S40. S. Wuchty, B. F. Jones, B. Uzzi, The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007). doi:10.1126/science.1136099 Medline

S41. R. Hill, Y. Yin, C. Stein, D. Wang, B. F. Jones, Adaptability and the pivot penalty in science. SSRN 3886142 [Preprint] (2021); http://dx.doi.org/10.2139/ssrn.3886142.

S42. S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, A.-L. Barabási, Science of science. *Science* **359**, eaao0185 (2018). doi:10.1126/science.aao0185 Medline

S43. R. Sinatra, D. Wang, P. Deville, C. Song, A.-L. Barabási, Quantifying the evolution of individual scientific impact. *Science* **354**, aaf5239 (2016). doi:10.1126/science.aaf5239 Medline

S44. L. Liu, Y. Wang, R. Sinatra, C. L. Giles, C. Song, D. Wang, Hot streaks in artistic, cultural, and scientific careers. *Nature* **559**, 396–399 (2018). doi:10.1038/s41586-018-0315-8 Medline

S45. A. M. Petersen, Quantifying the impact of weak, strong, and super ties in scientific careers. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E4671–E4680 (2015). doi:10.1073/pnas.1501444112 Medline

S46. M. Teplitskiy, E. Duede, M. Menietti, K. R. Lakhani, How status of research papers affects the way they are read and cited. *Res. Policy* **51**, 104484 (2022). doi:10.1016/j.respol.2022.104484

S47. G. R. Latona, M. H. Ribeiro, T. R. Davidson, V. Veselovsky, R. West, The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates. arXiv:2405.02150 [cs.CY] (2024).

S48. H. Bao, M. Sun, M. Teplitskiy, Where there's a will there's a way: ChatGPT is used more for science in countries where it is prohibited. arXiv:2406.11583 [cs.DL] (2024).

S49. A. Reinhart, B. Markey, M. Laudenbach, K. Pantusen, R. Yurko, G. Weinberg, D. W. Brown, Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2422455122 (2025). doi:10.1073/pnas.2422455122

S50. H.-Z. Cheng *et al*., Have AI-generated texts from LLM infiltrated the realm of scientific writing? A large-scale analysis of preprint platforms. bioRxiv 2024.03.25.586710 [Preprint] (2024); https://doi.org/10.1101/2024.03.25.586710.

S51. M. Krenn, L. Buffoni, B. Coutinho, S. Eppel, J. G. Foster, A. Gritsevskiy, H. Lee, Y. Lu, J. P. Moutinho, N. Sanjabi, R. Sonthalia, N. M. Tran, F. Valente, Y. Xie, R. Yu, M. Kopp, Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. *Nat. Mach. Intell.* **5**, 1326–1335 (2023). doi:10.1038/s42256-023-00735-0

S52. H. Peng, H. S. Qiu, H. B. Fosse, B. Uzzi, Promotional language and the adoption of innovative ideas in science. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2320066121 (2024). doi:10.1073/pnas.2320066121 Medline

S53. E. Zhou, D. Lee, Generative artificial intelligence, human creativity, and art. *PNAS Nexus* **3**, pgae052 (2024). doi:10.1093/pnasnexus/pgae052 Medline