nature human behaviour

Article

https://doi.org/10.1038/s41562-025-02273-8

# Quantifying large language model usage in scientific papers

Received: 2 May 2024

Accepted: 19 June 2025

Published online: 04 August 2025

Check for updates

Weixin Liang[1,9] ✉, Yaohui Zhang[2,9], Zhengxuan Wu[1], Haley Lepp[3], Wenlong Ji[4], Xuandong Zhao[5], Hancheng Cao[1,6], Sheng Liu[7], Siyu He[7], Zhi Huang[7], Diyi Yang[1], Christopher Potts[1,8,10], Christopher D. Manning[1,8,10] & James Zou[1,2,7,10] ✉

Scientific publishing is the primary means of disseminating research findings. There has been speculation about how extensively large language models (LLMs) are being used in academic writing. Here we conduct a systematic analysis across 1,121,912 preprints and published papers from January 2020 to September 2024 on *arXiv*, *bioRxiv* and *Nature* portfolio journals, using a population-level framework based on word frequency shifts to estimate the prevalence of LLM-modified content over time. Our findings suggest a steady increase in LLM usage, with the largest and fastest growth estimated for computer science papers (up to 22%). By comparison, mathematics papers and the *Nature* portfolio showed lower evidence of LLM modification (up to 9%). LLM modification estimates were higher among papers from first authors who post preprints more frequently, papers in more crowded research areas and papers of shorter lengths. Our findings suggest that LLMs are being broadly used in scientific writing.

The release of ChatGPT in late 2022 has coincided with a growing number of reports describing the presence of large language model (LLM)-generated content in scientific manuscripts[1,2] and peer review reports[3]. While some instances are identifiable through explicit textual markers—such as the inclusion of interface prompts such as 'regenerate response'[4,5] or generic disclaimers such as 'as an artificial intelligence (AI) language model'[6]—many others are too subtle to identify at the individual level. However, from a birds-eye view, broader trends can become apparent. Prior work has also raised concerns regarding the fairness and validity of existing detection tools, particularly in relation to their disparate impact on non-native English authors[7]. Although model-level classifiers continue to improve, reliably identifying LLM-generated or LLM-modified content at the individual level remains a complex and unresolved task[8].

Examining LLM-modified content in the aggregate, however, presents an opportunity to understand broader shifts in scholarly communication. Liang et al.[9] introduce a methodological framework for estimating the proportion of LLM-modified text within a corpus, without relying on the identification of individual instances. By population-level approach is designed to move beyond case-based classification, enabling insights into the prevalence and distribution of generative model use. When applied to academic texts, such methods facilitate the identification of systemic conditions that may shape engagement with LLMs, while also revealing linguistic and stylistic trends that may not be apparent at smaller scales.

Measuring the extent of LLM use on scientific publishing has urgent applications. Concerns about accuracy, plagiarism, anonymity, ownership and scientific independence have prompted some prominent scientific institutions to take a stance on the use of LLM-modified content in academic publications. The 2023 International Conference on Machine Learning (ICML) has prohibited the inclusion of

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [3]Graduate School of Education, Stanford University, Stanford, CA, USA. [4]Department of Statistics, Stanford University, Stanford, CA, USA. [5]Department of Computer Science, University of California, Santa Barbara, Santa Barbara, CA, USA. [6]Goizueta Business School, Emory University, Atlanta, GA, USA. [7]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [8]Department of Linguistics, Stanford University, Stanford, CA, USA. [9]These authors contributed equally: Weixin Liang, Yaohui Zhang. [10]These authors jointly supervised this work: Christopher Potts, Christopher D. Manning, James Zou. ✉e-mail: wxliang@stanford.edu; jamesz@stanford.edu

LLM-generated content in submitted manuscripts, except where it forms part of the experimental methodology[10]. The journal *Science* introduced editorial policies prohibiting the use of LLM-generated text, images or graphics in submitted or published material[11]. Despite such policies, little is currently known about the extent to which LLMs have been adopted in scientific writing practices. Developing scalable, methodologically robust tools for monitoring LLM use can inform institutional decision-making and supporting evidence-based policy.

In addition to normative concerns, understanding the distribution of LLM use may offer insight into structural pressures that shape scientific writing practices. For example, reliance on generative tools may reflect linguistic marginalisation, resource constraints or the demands of high publication volume. Identifying where and how LLMs are used can help clarify the social and institutional contexts in which they are being adopted.

In this study, we present a large-scale empirical analysis of LLM-modified content across scientific preprints and journal articles. Building on the distributional GPT quantification framework proposed in prior work[9], we estimate the proportion of academic text that has been substantially modified by LLMs. For the purposes of this analysis, 'LLM-modified' refers to text content substantially altered by ChatGPT beyond basic orthographic and grammatical corrections. Modifications captured in our analysis may include, for example, summarization of existing writing or prose generation based on structural outlines.

A key characteristic of the framework is its population-level inference, which enables corpus-wide quantification without requiring individual-document classification. As validated in the prior paper, the framework is orders of magnitude more computationally efficient and thus scalable, produces more accurate estimates and generalizes better than its counterparts under substantial temporal distribution shifts and other realistic distribution shifts.

We apply this framework to abstracts and main texts (Fig. 1 and Supplementary Fig. 1) of academic papers across multiple disciplines, including *arXiv*, *bioRxiv* and 15 journals within the *Nature* portfolio, such as *Nature*, *Nature Biomedical Engineering*, *Nature Human Behaviour* and *Nature Communications*. Our analysis encompasses a total of 1,121,912 papers published between January 2020 and September 2024, comprising 861,253 papers from *arXiv*, 205,094 from *bioRxiv* and 55,565 from the *Nature* portfolio journals. The papers from *arXiv* span multiple academic disciplines, including computer science, electrical engineering and systems science, mathematics, physics and statistics. These datasets allow us to quantify the prevalence of LLM-modified academic writing over time and across a broad range of academic fields.

Our results indicate that the largest and fastest growth in LLM use was observed in computer science papers, with $\alpha$ reaching 22.5% for abstracts and 19.6% for introductions by September 2024. By contrast, the mathematics papers and the *Nature* portfolio showed the least increase, with $\alpha$ reaching 7.7% and 8.9% for abstracts and 4.1% and 9.4% for introductions, respectively.

Moreover, our analysis reveals at the aggregate level that higher levels of LLM-modification are associated with papers whose first authors publish preprints more frequently and papers with reduced length. Results also demonstrate stronger correlations between papers with LLM-modifications, which may indicate increased use in more competitive research areas (as measured by proximity to the nearest neighbouring paper in the embedding space) or that generated text is reducing writing diversity. We adapt the distributional LLM quantification framework from Liang et al.[9] to quantify the prevalence of LLM-modified academic writing (Methods).

## Results

### Overview of the *arXiv*, *bioRxiv* and *Nature* portfolio data

We collected data from three sources: *arXiv*, *bioRxiv* and 15 journals from the *Nature* portfolio. For *bioRxiv* and the *Nature* portfolio, we randomly sampled up to 2,000 papers per month from January 2020 to September 2024. For *arXiv*, which covers multiple academic fields including computer science, electrical engineering and systems science, mathematics, physics and statistics, we randomly sampled up to 2,000 papers per month for each main category during the same time period. We then generated LLM-produced training data using the two-stage approach described in the Methods.

For the main analysis, we focused on the introduction sections, as the introduction was the most consistently and commonly occurring section across diverse categories of papers. However, for the computer science category on *arXiv*, which showed the highest estimated LLM-modified content, we conducted a more detailed analysis by examining various sections of the papers, including abstracts, introductions, related works, methods, experiments and conclusions (Supplementary Fig. 2). See Supplementary Section A for comprehensive implementation details.

### Data split, model fitting and evaluation

For model fitting, we count word frequencies for scientific papers written before the release of ChatGPT and the LLM-modified corpora. We fit the model with data from 2020 and use data from January 2021 onwards for validation and inference. We fit separate models for abstracts and introductions for each major category.

To evaluate model accuracy and calibration under temporal distribution shift, we use 3,000 papers from 1 January 2022 to 29 November 2022, a time period before the release of ChatGPT, as the validation data. We construct validation sets with LLM-modified content proportions ($\alpha$) ranging from 0% to 25%, in 5% increments and compared the model's estimated $\alpha$ with the ground truth $\alpha$ (Fig. 2). Full vocabulary, adjectives, adverbs and verbs all performed well in our application, with a prediction error consistently less than 3.5% at the population level across various ground truth $\alpha$ values (Fig. 2).

### Temporal trends in AI-modified academic writing

We applied the model to estimate the fraction of LLM-modified content ($\alpha$) for each paper category each month, for both abstracts and introductions. Each point in time was independently estimated, with no temporal smoothing or continuity assumptions applied.

Our findings reveal a steady increase in the fraction of LLM-modified content ($\alpha$) in both the abstracts (Fig. 1) and the introductions (Supplementary Fig. 1), with the largest and fastest growth observed in computer science papers. By September 2024, the estimated $\alpha$ for computer science had increased to 22.5% for abstracts (bootstrapped 95% confidence interval (CIs) (21.7%, 23.3%)) and 19.6% for introductions (bootstrapped 95% CIs (19.2%, 20.0%)). The second-fastest growth was observed in electrical engineering and systems science, with the estimated $\alpha$ reaching 18.0% for abstracts (bootstrapped 95% CIs (16.7%, 19.3%)) and 18.4% for introductions (bootstrapped 95% CIs (17.8%, 19.0%)) during the same period. By contrast, mathematics papers and the *Nature* portfolio showed the smallest increase. By the end of the studied period, the estimated $\alpha$ for mathematics had increased to 7.7% for abstracts (bootstrapped 95% CIs (7.1%, 8.3%)) and 4.1% for introductions (bootstrapped 95% CIs (3.9%, 4.3%)), while the estimated $\alpha$ for the *Nature* portfolio had reached 8.9% for abstracts (bootstrapped 95% CIs (8.2%, 9.6%)) and 9.4% for introductions (bootstrapped 95% CIs (9.0%, 9.8%)).

The November 2022 estimates serve as a pre-ChatGPT reference point for comparison, as ChatGPT was launched on 30 November 2022. The estimated $\alpha$ for computer science in November 2022 was 2.4% (bootstrapped 95% CIs (2.1%, 2.7%)), while for electrical engineering and systems science, mathematics and the *Nature* portfolio, the estimates were 2.9% (bootstrapped 95% CIs (2.3%, 3.5%)), 2.5% (bootstrapped 95% CIs (2.1%, 2.9%)) and 3.4% (bootstrapped 95% CIs (2.8%, 4.0%)), respectively. These values are consistent with the false positive rate we found in the modal validations (Fig. 2).

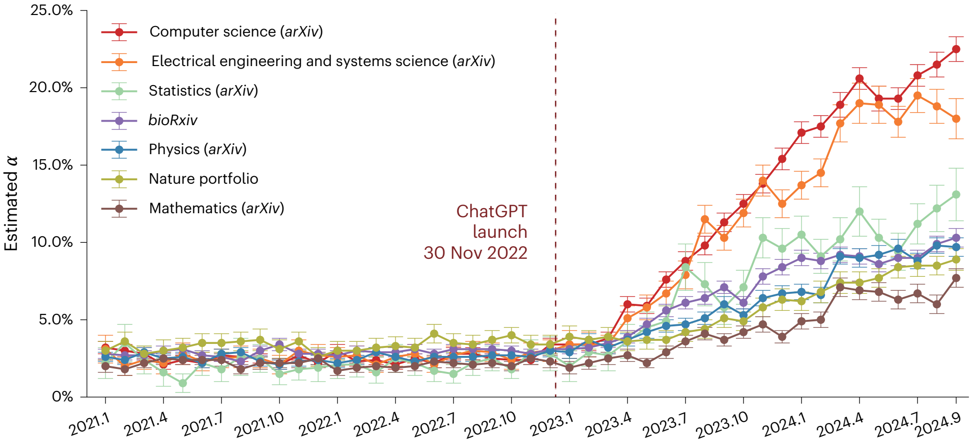

**Fig. 1 | Estimated fraction of LLM-modified sentences across research paper venues over time.** This figure displays the fraction ($\alpha$) of sentences estimated to have been substantially modified by LLM in abstracts from various academic writing venues. The vertical brown dashed line marks the release date of ChatGPT (November 30, 2022). The analysis includes five areas within *arXiv* (computer science, electrical engineering and systems science, mathematics, physics and statistics), articles from *bioRxiv* and a combined dataset from 15 journals within the *Nature* portfolio. The estimates are based on the distributional GPT quantification framework, which provides population-level estimates rather than individual document analysis. Each point in time is independently estimated, with no temporal smoothing or continuity assumptions applied. The data are presented as the mean ± 95% CI based on 1,000 bootstrap iterations. For computer science (*arXiv*), $n$ = 2,000 independent paper abstracts per month. For electrical engineering and systems science (*arXiv*), the monthly sample size of independent paper abstracts varied (mean of 708; minimum (min) of 388; maximum (max) of 1,041). For statistics (*arXiv*), the monthly sample size of independent paper abstracts varied (mean of 337; min of 203; max of 513). For *bioRxiv*, $n$ = 2,000 independent paper abstracts per month. For physics (*arXiv*), $n$ = 2,000 independent paper abstracts per month. For *Nature* portfolio, the monthly sample size of independent paper abstracts varied (mean of 1,039; min of 601; max of 1,537). For mathematics (*arXiv*), the monthly sample size of independent paper abstracts varied (mean of 1,958; min of 1,444; max of 2,000).

As computer science papers from *arXiv* show the highest estimated $\alpha$, we further stratified the main paper content by section (Supplementary Fig. 2). We found a higher fraction of LLM-modified content in abstracts, introductions, related works and conclusions compared with experiment and method sections (similar results were also obeserved in electrical engineering and systems science papers from *arXiv*) (Supplementary Fig. 3). This observation aligns with the current strengths of LLMs in summarization tasks, which might inspire scholars to use the tool for writing abstracts.

## Relationship between first-author preprint posting frequency and GPT usage

We found a notable correlation between the number of preprints posted by the first author on *arXiv* and the estimated number of LLM-modified sentences in their academic writing. The papers were stratified into two groups based on the number of first-authored *arXiv* computer science preprints by the first author in the year: those with two or fewer (≤2) preprints and those with three or more (≥3) preprints (Fig. 3a). We used the 2023 author grouping for the 2024 data, as we do not have the complete 2024 author data yet.

By September 2024, abstracts of papers whose first authors had ≥3 preprints in 2023 showed an estimated 22.9% (bootstrapped 95% CIs (21.7%, 24.1%)) of sentences modified by LLMs, compared with 20.0% (bootstrapped 95% CIs (19.2%, 20.8%)) for papers whose first authors had ≤2 preprints (Fig. 3a). We observe a similar trend in the introduction sections, with first authors posting more preprints having an estimated 20.9% (bootstrapped 95% CIs (20.4%, 21.4%)) LLM-modified sentences, compared with 17.8% (bootstrapped 95% CIs (17.5%, 18.1%)) for first authors posting fewer preprints (Fig. 3a). Since the first-author preprint posting frequency may be confounded by research field, we conduct an additional robustness check for our findings. We find that the observed trend holds for each of the three *arXiv* computer science subcategories: cs.CV (computer vision and pattern recognition), cs.LG (machine learning) and cs.CL (computation and language) (Supplementary Fig. 4a–c).

Our results suggest that researchers posting more preprints tend to utilize LLMs more extensively in their writing. One interpretation of this effect could be that the increasingly competitive and fast-paced nature of CS research communities incentivizes taking steps to accelerate the writing process. We do not evaluate whether these preprints were accepted for publication.

## Relationship between paper similarity and LLM usage

We investigate the relationship between a paper's similarity to its closest peer and the estimated LLM usage in the abstract. To measure similarity, we first embed each abstract from the *arXiv* computer science papers using OpenAI's text-embedding-three-small model, creating a vector representation for each abstract. We then calculate the distance between each paper's vector and its nearest neighbour within the *arXiv* computer science abstracts. Based on this similarity measure we divide papers into two groups: those more similar to their closest peer (below median distance) and those less similar (above median distance).

The temporal trends of LLM usage for these two groups are shown in Fig. 3b. After the release of ChatGPT, papers most similar to their closest peer consistently showed higher LLM usage compared with those least similar. By September 2024, the abstracts of papers more similar to their closest peer had an estimated 23.0% (bootstrapped 95% CIs (22.3%, 23.7%)) of sentences modified by LLMs, compared with 18.7% (bootstrapped 95% CIs (18.0%, 19.4%)) for papers less similar to their closest peer. To account for potential confounding effects of research fields, we conducted an additional robustness check by measuring the nearest neighbour distance within each of the three *arXiv* computer science subcategories: cs.CV (computer vision and pattern recognition), cs.LG (machine learning) and cs.CL (computation and language) and found that the observed trend holds for each subcategory (Supplementary Fig. 5a–c).

There are several ways to interpret these findings. First, LLM-use in writing could cause the similarity in writing or content. The similarity we observe could be incidental or sought after: community pressures could motivate scholars to incorporate LLM-generated text if they perceive the 'style' of generated text to be more prestigious than their own. Alternatively the crowded nature of fields could cause the uptick in use: LLMs may be more commonly used in research areas

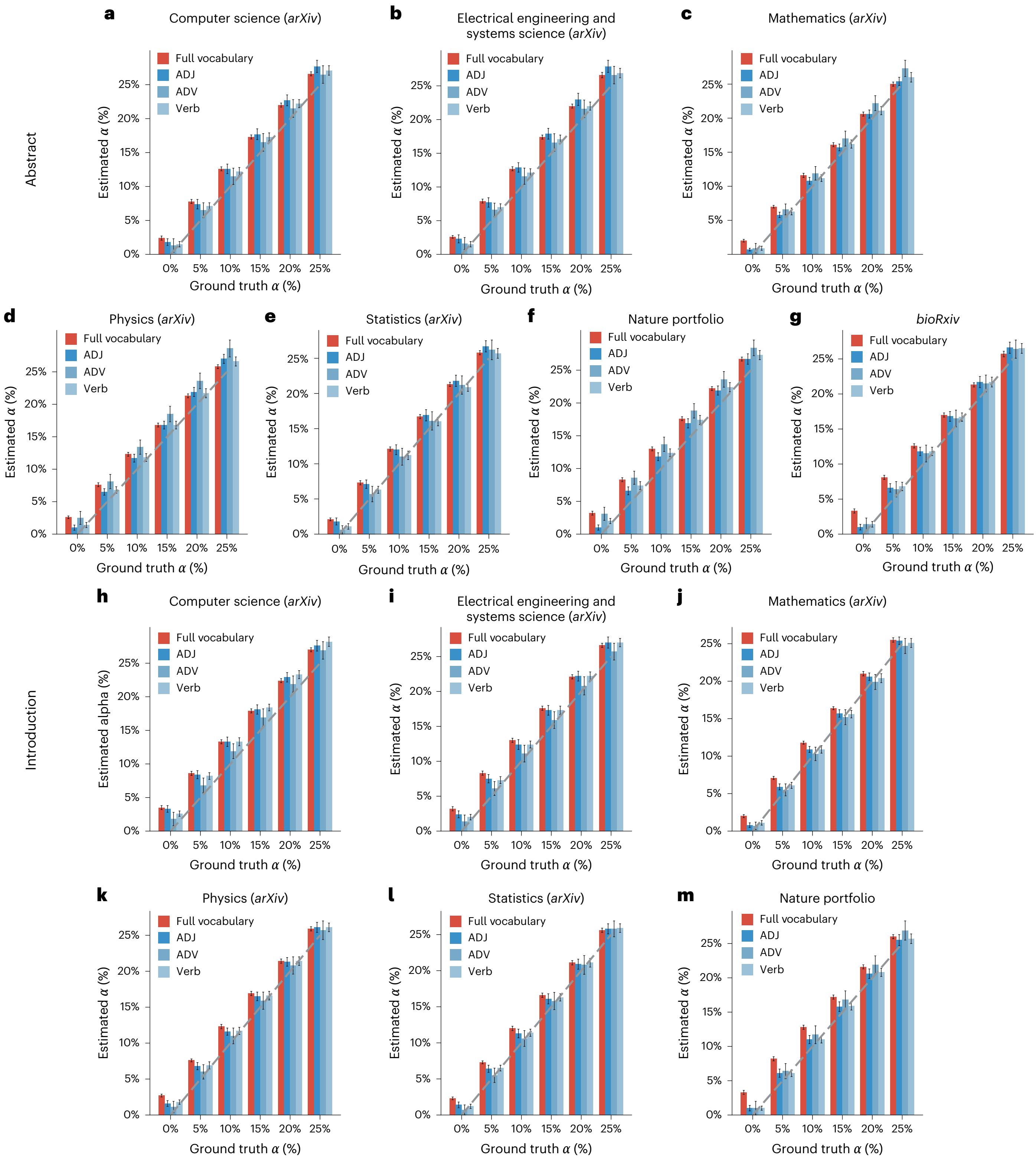

**Fig. 2 | Fine-grained validation of estimation accuracy under temporal distribution shift.** Panels **a**–**g** show the validation results on abstracts from each academic writing venue, while panels **h**–**m** show the validation on introductions. ADJ and ADV refer to adjectives and adverbs, respectively. We evaluate the accuracy of our models in estimating the fraction of LLM-modified content ($\alpha$) under a challenging temporal data split, where the validation data (sampled from 1 January 2022 to 29 November 2022) are temporally separated from the training data (collected up to 31 December 2020) by at least a year. The $x$ axis indicates the ground truth $\alpha$, while the $y$ axis indicates the model's estimated $\alpha$. In all cases, the estimation error for $\alpha$ is less than 3.5%. We did not include *bioRxiv* introductions due to the unavailability of bulk PDF downloads. The data are presented as the mean ± 95% CIs based on 1,000 bootstrap iterations. For each ground truth $\alpha$, $n$ = 30,000 sentences.

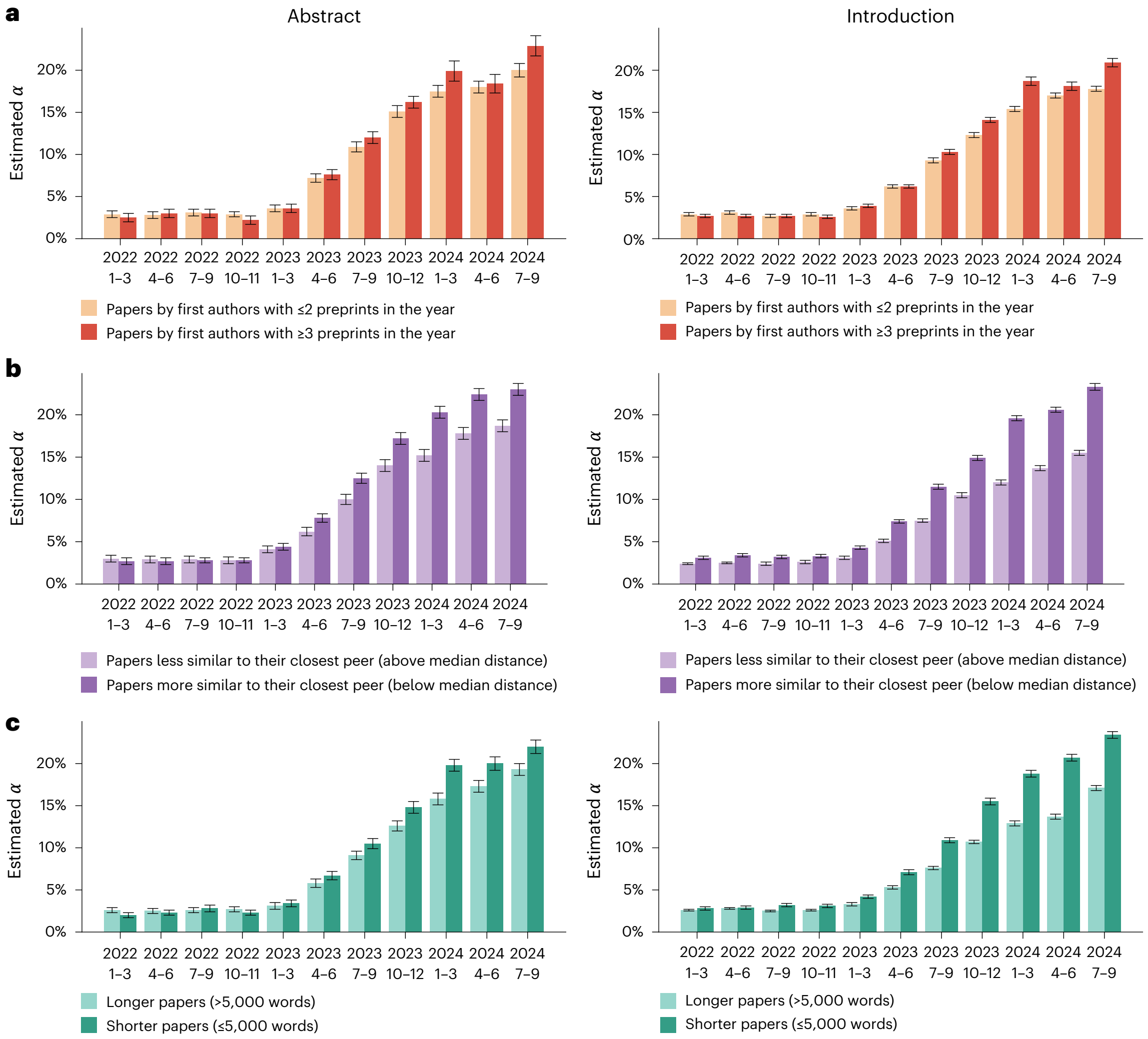

**Fig. 3 | Associations between LLM-modification and scientific publishing characteristics in *arXiv* computer science papers. a**, The papers authored by first authors who post preprints more frequently tend to have a higher fraction of LLM-modified content. The papers in *arXiv* computer science are stratified into two groups based on the preprint posting frequency of their first author, as measured by the number of first-authored preprints in the year. The sample size of papers by first authors with ≤2 preprints is $n = 2{,}000$ per quarter. For papers by first authors with ≥3 preprints, the quarterly sample size varied (mean of 1,202; minimum of 870; maximum of 1,849). **b**, The papers in more crowded research areas tend to have a higher fraction of LLM-modified content. The papers in *arXiv* computer science are divided into two groups based on their abstract's embedding distance to their closest peer: the papers more similar to their closest peer (below median distance) and papers less similar to their closest peer (above median distance). In both groups, $n = 2{,}000$ independent papers per quarter. **c**, Shorter papers tend to have a higher fraction of LLM-modified content. *arXiv* computer science papers are stratified by their full text word count, including appendices, into two bins: below or above 5,000 words (the rounded median). In both groups, $n = 2{,}000$ independent papers per quarter. The findings also hold when stratified by more fine-grained subject categories (Supplementary Figs. 4–6). The data are presented as mean ± 95% CIs based on 1,000 bootstrap iterations.

in which papers tend to be more similar to each other. If a subfield is more crowded, then multiple research teams could be studying the same topic and producing similar writing. The resulting competition may coerce researchers to make use of LLM-generated text to speed up the publication of findings. To further explore these hypotheses, our comparative analysis (Supplementary Section B.4) offers suggestive evidence in favour of the first: papers with high and low LLM usage had comparable nearest neighbour distances to 2022 publications—indicating similar baseline field competitiveness—yet exhibited a more pronounced gap when comparing nearest neighbour distances within 2023. This pattern supports the interpretation that LLM usage itself may be contributing to increased similarity in academic writing.

### Relationship between paper length and AI usage

We also explored the association between paper length and LLM usage in *arXiv* computer science papers. The papers were stratified by their full text word count, including appendices, into two bins: below or above 5,000 words (the rounded median).

Figure 3c shows the temporal trends of LLM usage for these two groups. After the release of ChatGPT, shorter papers consistently showed higher LLM usage compared with longer papers. By September 2024, the abstracts of shorter papers had an estimated 22.0% (bootstrapped 95% CIs (21.2%, 22.8%)) of sentences modified by LLMs, compared with 19.3% (bootstrapped 95% CIs (18.6%, 20.0%)) for longer papers (Fig. 3c).

We observe a similar trend in the introduction sections (Fig. 3c). To account for potential confounding effects of research fields, we conducted an additional robustness check. The finding holds for both cs.CV (computer vision and pattern recognition) and cs.LG (machine learning) (Supplementary Fig. 6a–c). However, for cs.CL (computation and language), we found no consistent difference in LLM usage between shorter and longer papers, possibly due to the limited sample

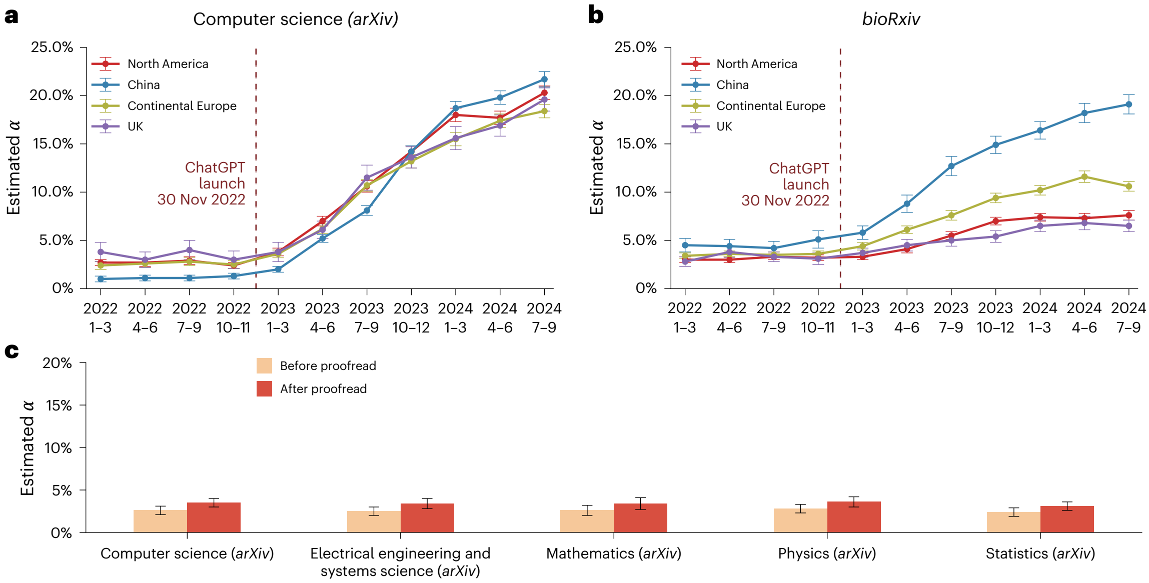

**Fig. 4 | Regional trends in the adoption of LLMs for academic writing.** **a**, The quarterly growth of LLM usage in computer science publications on the *arXiv* by first author affiliation region. Distinct regions (North America, China, Continental Europe and the UK) exhibit consistent upward trends in LLM adoption. For North America, the quarterly sample size $n = 2{,}000$. For China, the quarterly sample size varied (mean of 1,752; minimum (min) of 1,232; maximum (max) of 2,000). For Continental Europe, the quarterly sample size varied (mean of 1,929; min of 1,491; max of 2,000). For the UK, the quarterly sample size varied (mean of 558; min of 346; max of 835). **b**, The quarterly growth of LLM usage in biology publications on *bioRxiv* by first author affiliation region. Different regions (North America, China, Continental Europe and the UK) display consistent increases in LLM usage, with papers from regions with lower rates of English speakers, including China and Continental Europe, showing slightly higher estimated usage rates. For North America, the quarterly sample size $n = 2{,}000$. For China, the quarterly sample size varied (mean of 688; min of 439; max of 872). For Continental Europe, the quarterly sample size $n = 2{,}000$. For the UK, the quarterly sample size varied (mean of 830; min of 541; max of 972). **c**, Robustness of LLM-modified content prevalence quantification to proofreading. The plot illustrates similar proportions of LLM-modified content estimated after employing LLMs for 'proofreading' across various *arXiv* main categories. For each area, the sample size is $n = 1{,}000$ independent abstracts. This finding confirms the robustness of our method to minor text edits generated by LLMs, such as those introduced by simple proofreading tasks. The data are presented as the mean ± 95% CIs based on 1,000 bootstrap iterations.

size, as we only parsed a subset of the LaTeX sources and calculated their full length.

As computer science conference papers typically have a fixed page limit, longer papers probably have more substantial content in the appendix. The lower LLM usage in these papers may suggest that researchers with more comprehensive work rely less on LLM-assistance in their writing. However, further investigation is needed to determine the relationship between paper length, content comprehensiveness and the quality of the research.

## Regional trends in LLM adoption for academic writing

To investigate the regional trends in the adoption of LLMs for academic writing, we analysed the quarterly growth of LLM usage in computer science papers on *arXiv* (by first author affiliation) and biology papers on *bioRxiv* (by corresponding author affiliation) across different regions (Fig. 4a,b). Interestingly, we observed higher estimated usage rates in *bioRxiv* papers from regions with lower populations of English-language speakers, including China and Continental Europe, compared with those from North America and the UK (Fig. 4b). The number of papers from Africa and South America is too low to include in our calculations, demonstrating the urgent importance of efforts to increase geographic diversity in scientific publishing. This difference may be attributed to authors using ChatGPT for English-language assistance. In the *arXiv* data, although the absolute estimates of LLM usage are similar across regions by the end of the study, the patterns of relative growth reveal a notable distinction. In particular, our results show that China exhibits the largest relative increase when the false positive rate is reduced, which aligns with our findings in bioRxiv.

To further validate the robustness of our method, we examined the effect of employing LLMs for 'proofreading' on the estimated proportion of LLM-modified content across various *arXiv* main categories (Fig. 4c and Supplementary Fig. 7). The similarities in the fraction of estimated LLM-modified content after proofreading, with only a slight measurable increase of approximately 1%, confirms that our approach is robust to minor text edits generated by LLMs during simple proofreading tasks. Overall, these findings highlight the growing utilization of LLMs in academic writing across different regions and research fields, emphasizing the need for further research on the implications associated with their use.

Based on our observations of LLM use in academic writing, we conducted a brief analysis of how scholars disclose this use in their writing. We manually inspected 200 randomly sampled computer science papers uploaded to *arXiv* in February 2024. We found that only 2 out of the 200 papers explicitly disclosed the use of LLMs during paper writing. Further analysis of disclosure motivation might help determine an explanation. For example, policies around disclosing LLM usage in academic writing may still be unclear, or scholars may have other motivations for intentionally avoiding to disclose use.

# Discussion

Our analysis of LLM-modified content in academic writing across various platforms (*arXiv*, *bioRxiv* and the *Nature* portfolio) reveals a sharp increase in the estimated fraction of LLM-modified content, beginning ~5 months after the release of ChatGPT. The 5-month lag and the slopes of the increased usage reflect the speed of diffusion and adoption of LLMs. We identified the fastest growth in computer science papers, a trend that may be partially explained by computer science researchers' familiarity with and access to LLMs. In addition, the fast-paced nature of LLM research and the associated pressure to publish quickly may incentivize the use of LLM writing assistance[12].

We quantified several other factors associated with higher LLM usage in academic writing. First, authors who post preprints

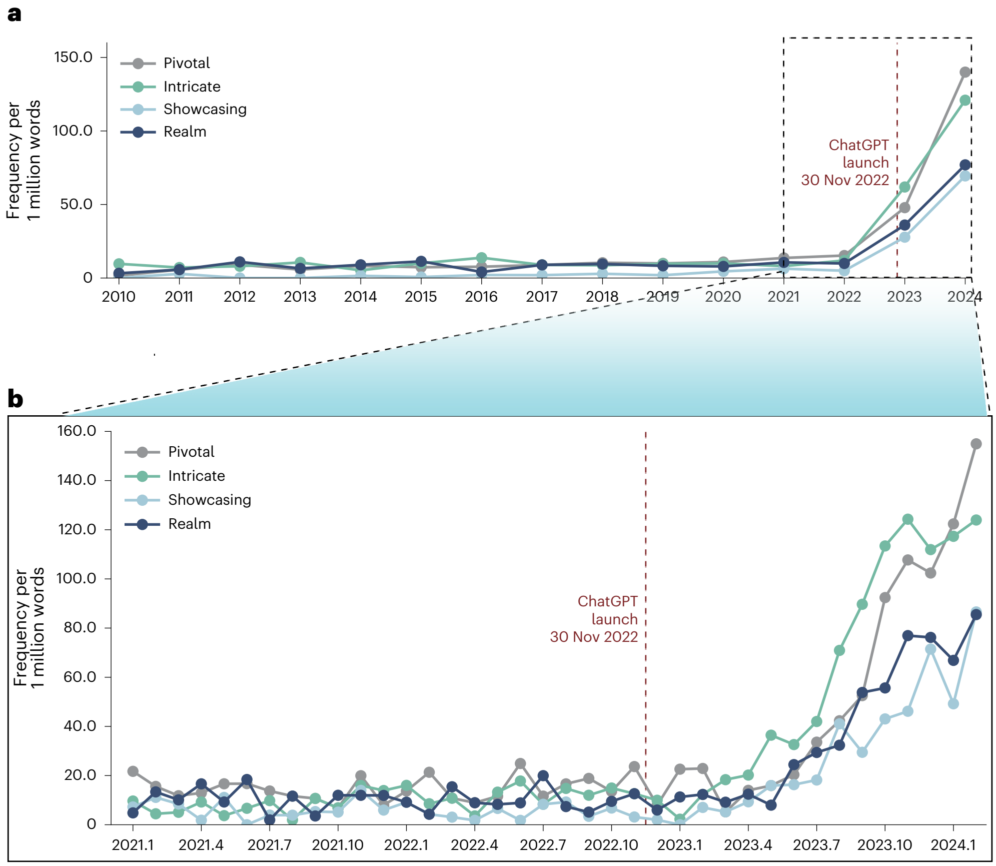

**Fig. 5 | Word frequency shift in *arXiv* computer science abstracts over 14 years (2010–2024). a**, The frequency over time for the top four words most disproportionately generated by LLMs in comparison to use in pre-ChatGPT corpora, as measured by the log odds ratio. The words are: 'realm', 'intricate', 'showcasing' and 'pivotal'. These terms maintained a consistently low frequency in *arXiv* CS abstracts over more than a decade (2010–2022) but experienced a sudden surge in usage starting in 2023. **b**, Zoomed in frequency between 2021 and 2024.

more frequently show a higher fraction of LLM-modified content in their writing. Second, papers in more crowded research areas, where papers tend to be more similar, showed higher LLM modification compared with those in less crowded areas. Third, shorter papers consistently showed higher LLM modification compared with longer papers, which may indicate that researchers trying to produce a higher quantity of writing are more likely to rely on LLMs. These results may be an indicator of the competitive nature of certain research areas and the pressure to publish quickly. We also found a higher fraction of AI-modified content in abstracts, introductions, related works and conclusions compared with the experiment and method sections. This suggests that researchers may be more comfortable using LLM for summarization tasks, such as writing abstracts, which traditionally provide a concise overview of the entire paper.

Furthermore, our regional analysis of LLM adoption in academic writing revealed higher estimated usage rates in *bioRxiv* papers from regions with lower populations of English-language speakers, including China and Continental Europe, compared with those from North America and the UK. In CS *arXiv* papers, the increase in LLM usage is consistently high across the different regions, potentially reflecting differences across disciplines. It is important to note that using author affiliation as a proxy for country of origin has inherent limitations, as it may not accurately reflect an author's linguistic or cultural background. Furthermore, papers posted on *arXiv* and *bioRxiv* may not be fully representative of all research output from each region, and patterns of LLM usage could differ for papers published in regional journals or other venues not captured by our analysis.

How can we interpret the uneven adoption of LLMs across different world regions? One widely discussed use-case of LLMs in scientific publishing is the 'polishing' of writing by multilingual scientists. Today, English is nearly hegemonic in scientific publishing, creating a 'tax' on scientists who do not speak English as a first language[13]. A technology that can generate dominant varieties of scholarly English could hypothetically lower barriers to entry[14,15]. However, several recent studies have demonstrated the complex interactions between AI-use and scholarly language ideologies. Lepp and Smith[16] find that while ChatGPT may cover up 'errors' in writing, peer reviewers now describe words like 'delve' as indicators that an author might not be a native English speaker. Writers anticipate this, and attempt to remove ChatGPT-markers from their writing. Liang et al.[7] show that GPT detectors also discriminate against people based on language background. Agarwal et al.[17] show that not only does LLM use 'homogenize' towards 'Western' language but content as well. In the context of science, in which diverse ideas lead to a robust research ecosystem, this finding opens new questions about the desirability of using LLMs for linguistic assimilation[18]. There is much to be learned from future system-level studies of how people make use of AI to express and navigate complex language ideologies. Future studies should collect more granular data on author backgrounds, research topics and motivations to better understand the regional variation in LLM adoption and its implications for global scientific communication.

While our study focused on ChatGPT, which accounts for more than three-quarters of worldwide internet traffic in the category[19], we acknowledge that there are other LLMs used for assisting academic writing. Regardless, relying heavily on LLMs owned by private companies raises concerns about safeguarding the security and autonomy of scientific work. We hope our findings will spark further investigations into the widespread use of LLM-assisted writing and encourage discussions on creating scientific publishing environments that value openness, intellectual diversity, factual reliability and scholarly independence.

Furthermore, while previous work[7] demonstrated that GPT-detection methods can falsely identify the writing of language learners as LLM-generated, our results showed consistently low false positives estimates of $\alpha$ in 2022, which contains a substantial fraction of texts written by multilingual scholars. We recognize that substantial author population changes[20] or other language-use shifts could still impact the accuracy of our estimates. In addition, while our model demonstrates high accuracy in detecting LLM-modified content, it has several limitations. First, the detection method is not a direct measure of LLM usage—it identifies statistical patterns consistent with LLM-generated text, which may not always correspond to actual use. Second, the method systematically overestimates LLM usage at the lower end and underestimates it at the higher end of the distribution. These biases can affect absolute prevalence estimates, though the relative trends remain robust. Third, shifts in writing style, evolving research practices or changes in author demographics (for example, increased participation from multilingual scholars) could also influence model predictions. Despite these limitations, the relative increase after subtracting the false positive rate remains substantial (for example 19% for the abstracts of *arXiv* CS papers) and supports our overall findings. Finally, the associations that we observe between LLM usage and paper characteristics are correlations which could be affected by other factors such as research topics. Future studies should explore the causal relationship between LLM use and observed temporal changes.

Prior research suggests that CS researchers adopt AI technologies at a higher rate than those in other fields[21]. This may be due to their greater exposure and familiarity with AI, as AI research primarily originates in CS and disseminates through collaborations with CS researchers[22]. In addition, familiarity with AI may foster greater confidence in its use, as studies have shown a correlation between familiarity and confidence in AI[23,24]. However, our study does not differentiate between these mechanisms, and this limitation may constrain our comprehensive understanding of the underlying factors driving AI technology adoption.

Our observations of the rise of generated or modified papers open many questions for future research. How do such papers compare in terms of accuracy, creativity or diversity? How do readers react to LLM-generated abstracts and introductions? How do citation patterns of LLM-generated papers compare with other papers in similar fields? How might the dominance of a limited number of for-profit organizations in the LLM industry affect the independence of scientific output? We hope our results and methodology inspire further studies of widespread LLM-modified text and conversations about how to promote transparent, diverse and high-quality scientific publishing.

## Methods

No ethics approval was required, as the study did not involve human participants.

### The distributional LLM quantification framework

We adapt the distributional LLM quantification framework from Liang et al.[9] to quantify the use of LLM-modified academic writing. This framework leverages word frequency shifts (Fig. 5a,b and Supplementary Fig. 8) to measure the prevalence of LLM-generated text. More specifically, the framework consists of the following steps:

(1) Problem formulation: denote $X$ as each individual document, $\mathcal{P}$ and $\mathcal{Q}$ as the probability distributions of human-written and LLM-modified documents, respectively. The mixture distribution is given by $\mathcal{D}_\alpha(X) = (1-\alpha)\mathcal{P}(x) + \alpha\mathcal{Q}(x)$, where $\alpha$ is the fraction of AI-modified documents. The goal is to estimate $\alpha$ based on observed documents $\{X_i\}_{i=1}^N \sim \mathcal{D}_\alpha(X)$, where $i$ is an integer index of observed document.

(2) Parameterization: to make $\alpha$ identifiable, the framework models the distributions of token occurrences in human-written and LLM-modified documents, denoted as $\mathcal{P}_T$ and $\mathcal{Q}_T$, respectively, for a chosen list of tokens $T = \{t_i\}_{i=1}^M$. The occurrence probabilities of each token in human-written and LLM-modified documents, $p_t$ and $q_t$, are used to parameterize $\mathcal{P}_T$ and $\mathcal{Q}_T$

$$\mathcal{P}_T(X) = \prod_{t\in T} p_t^{\mathbb{1}\{t\in X\}}(1-p_t)^{\mathbb{1}\{t\notin X\}}, \quad \mathcal{Q}_T(X) = \prod_{t\in T} q_t^{\mathbb{1}\{t\in X\}}(1-q_t)^{\mathbb{1}\{t\notin X\}}.$$

(3) Estimation: the occurrence probabilities $p_t$ and $q_t$ are estimated using collections of known human-written and LLM-modified documents, $\{X_j^P\}_{j=1}^{n_P}$ and $\{X_j^Q\}_{j=1}^{n_Q}$, respectively, here $j$ is an integer index of documents with known sources

$$\hat{p}_t = \frac{1}{n_P}\sum_{j=1}^{n_P} \mathbb{1}\left\{t \in X_j^P\right\}, \quad \hat{q}_t = \frac{1}{n_Q}\sum_{j=1}^{n_Q} \mathbb{1}\left\{t \in X_j^Q\right\}.$$

(4) Inference: the fraction $\alpha$ is estimated by the maximum likelihood estimator (MLE) on the observed documents under the mixture distribution $\hat{\mathcal{D}}_{\alpha,T}(X) = (1-\alpha)\hat{\mathcal{P}}_T(X) + \alpha\hat{\mathcal{Q}}_T(X)$:

$$\hat{\alpha}_T^{\mathrm{MLE}} = \operatorname*{argmax}_{\alpha\in[0,1]} \sum_{i=1}^{N} \log\left((1-\alpha)\hat{\mathcal{P}}_T(X_i) + \alpha\hat{\mathcal{Q}}_T(X_i)\right).$$

Liang et al.[9] demonstrate that the data points $\{X_i\}_{i=1}^N \sim \mathcal{D}_\alpha$ can be constructed either as a document or as a sentence, and both work well. Following their method, we use sentences as the unit of data points for the estimates in the main results. In addition, we extend this framework for our application to academic papers with two key differences:

### Generating realistic LLM-produced training data

We use a two-stage approach to generate LLM-produced text, as simply prompting an LLM with paper titles or keywords would result in unrealistic scientific writing samples containing fabricated results, evidence and ungrounded or hallucinated claims.

Specifically, given a paragraph from a paper known to not include LLM-modification, we first perform abstractive summarization using an LLM to extract key contents in the form of an outline. We then prompt the LLM to generate a full paragraph based the outline (see Supplementary Figs. 9 and 10 for full prompts).

Our two-stage approach can be considered a counterfactual framework for generating LLM text: given a paragraph written entirely by a human, how would the text read if it conveyed almost the same content but was generated by an LLM? This additional abstractive summarization step can be seen as the control for the content. This approach also simulates how scientists may be using LLMs in the writing process, where the scientists first write the outline themselves and then use LLMs to generate the full paragraph based on the outline.

### Using the full vocabulary for estimation

We use the full vocabulary instead of only adjectives, as our validation shows that adjectives, adverbs and verbs all perform well in our application (Fig. 2). Using the full vocabulary minimizes design biases stemming from vocabulary selection. We also find that using the full vocabulary is more sample-efficient in producing stable estimates, as indicated by their smaller confidence intervals by bootstrap.

### Overview of current LLM-generated text detection methods

Various methods have been proposed for detecting LLM-modified text, including zero-shot approaches that rely on statistical signatures characteristic of machine-generated content[25–32] and training-based methods that finetune language models for binary classification of human versus LLM-modified text[33–42]. However, these approaches face challenges such as the need for access to LLM internals, overfitting to training data and language models, vulnerability to adversarial attacks[43] and bias against non-dominant language varieties[7]. The effectiveness and reliability of publicly available LLM-modified text detectors have also been questioned[28,29,44–52], with the theoretical possibility of accurate instance-level detection being debated[53–55]. In this study, we apply the recently proposed distributional GPT quantification framework[9], which estimates the fraction of LLM-modified content in a text corpus at the population level, circumventing the need for classifying individual documents or sentences and improving upon the stability, accuracy and computational efficiency of existing approaches. The method also preserves the privacy of writers—that is, it does not detect individual cases of use. A more comprehensive discussion of related work can be found in Supplementary Section C.

### Overview of the *arXiv*, *bioRxiv* and *Nature* portfolio data

We collected data for this study from three publicly accessible sources: official application programming interfaces (APIs) provided by *arXiv* and *bioRxiv* and web pages from the *Nature* portfolio. For each of the five major *arXiv* categories (computer science, electrical engineering and systems science, mathematics, physics and statistics), we randomly sampled up to 2,000 papers per month from January 2020 to September 2024. Similarly, from *bioRxiv*, we randomly sampled up to 2,000 papers for each month within the same timeframe. To extract regional information from the *arXiv* preprints, we exploited the existing tool S2ORC[56] to parse author affiliations from LaTeX sources. We used the public affiliation metadata from *bioRxiv* preprints.

For the *Nature* portfolio, encompassing 15 *Nature* journals including *Nature*, *Nature Biomedical Engineering*, *Nature Human Behaviour* and *Nature Communications*, we followed the same sampling strategy, selecting 2,000 papers randomly from each month, from January 2020 to September 2024. When there were not enough papers to reach our target of 2,000 per month, we included all available papers. The *Nature* portfolio encompasses the following 15 *Nature* journals: *Nature*, *Nature Communications*, *Nature Ecology and Evolution*, *Nature Structural and Molecular Biology*, *Nature Cell Biology*, *Nature Human Behaviour*, *Nature Immunology*, *Nature Microbiology*, *Nature Biomedical Engineering*, *Communications Earth and Environment*, *Communications Biology*, *Communications Physics*, *Communications Chemistry*, *Communications Materials* and *Communications Medicine*.

### Robustness analysis of model variants using restricted word subsets

To assess the robustness of the model variants against the use of restricted word subsets, we conducted an analysis using vocabularies limited to adjectives, adverbs or verbs (Fig. 2). These subsets were obtained by the most-frequent-tag baseline approach, which assigns each word its most commonly observed part-of-speech tag. This method achieves fairly good accuracy[57]. As our validation showed that adjectives, adverbs and verbs all perform well in our application (Fig. 2), we used the full vocabulary in our analysis. Using the full vocabulary minimizes design biases stemming from vocabulary selection. We also find that using the full vocabulary is more sample-efficient in producing stable estimates, as indicated by their smaller confidence intervals by bootstrap.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The datasets analysed in the current study are public at the following links: via *arXiv* at https://www.kaggle.com/datasets/Cornell-University/arxiv (ref. 58), via *bioRxiv* at https://github.com/nicholasmfraser/rbiorxiv (ref. 59) and via *Nature* portfolio at https://www.nature.com/nature-portfolio (ref. 60).

## Code availability

The code can be accessed via GitHub at https://github.com/Weixin-Liang/Mapping-the-Increasing-Use-of-LLMs-in-Scientific-Papers (refs. 15,61). The study was conducted using Python 3.8.19, R 4.4.1.

## References

1. Okunytė, P. Google search exposes academics using ChatGPT in research papers. *Cybernews* https://cybernews.com/news/academic-cheating-chatgpt-openai/ (2023).
2. Deguerin, M. AI-generated nonsense is leaking into scientific journals. *Popular Science* https://www.popsci.com/technology/ai-generated-text-scientific-journals/ (2024).
3. Oransky, I. & Marcus, A. Papers and peer reviews with evidence of ChatGPT writing. *Retraction Watch* https://retractionwatch.com/papers-and-peer-reviews-with-evidence-of-chatgpt-writing/ (2024).
4. Conroy, G. Scientific sleuths spot dishonest ChatGPT use in papers. *Nature* https://doi.org/10.1038/d41586-023-02477-w (2023).
5. Conroy, G. How ChatGPT and other AI tools could disrupt scientific publishing. *Nature* https://doi.org/10.1038/d41586-023-03144-w (2023).
6. Vincent, J. 'As an AI language model': the phrase that shows how AI is pollulating the web. *The Verge* https://www.theverge.com/2023/4/25/23697218/ai-generated-spam-fake-user-reviews-as-an-ai-language-model (2023).
7. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E. & Zou, J. Y. GPT detectors are biased against non-native English writers. *Patterns (N Y)* https://doi.org/10.1016/j.patter.2023.100779 (2023).
8. Yu, S., Luo, M., Madasu, A., Lal, V. & Howard, P. Is your paper being reviewed by an LLM? Investigating AI text detectability in peer review. In *NeurIPS Safe Generative AI Workshop* https://openreview.net/forum?id=f2G7C2fKxV (2024).
9. Liang, W. 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* https://openreview.net/forum?id=bX3J7ho18S (ICML, 2024).
10. Clarification on large language model policy LLM. *ICML* https://icml.cc/Conferences/2023/llm-policy (2023).
11. Thorp, H. H. ChatGPT is fun, but not an author. *Science* **379**, 313 (2023).
12. Foster, J. G., Rzhetsky, A. & Evans, J. A. Tradition and innovation in scientists' research strategies. *Am. Soc. Rev.* **80**, 875–908 (2015).
13. Amano, T., González-Varo, J. P. & Sutherland, W. J. Languages are still a major barrier to global science. *PLoS Biol.* **14**, e2000933 (2016).
14. Lee, M. et al. A design space for intelligent and interactive writing assistants. In *Proc. 2024 CHI Conference on Human Factors in Computing Systems* (eds Mueller, F. F. et al.) https://doi.org/10.1145/3613904.3642697 (Association for Computing Machinery, 2024)
15. Liang, W. et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI* https://ai.nejm.org/doi/full/10.1056/AIoa2400196 (2023).
16. Lepp, H. & Smith, D. S. 'You cannot sound like GPT': signs of language discrimination and resistance in computer science publishing. In *Proc. 2025 ACM Conference on Fairness, Accountability, and Transparency* https://doi.org/10.1145/3715275.3732202 (Association for Computing Machinery, 2025).
17. Agarwal, D., Naaman, M. & Vashistha, A. AI suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proc. 2025 CHI Conference on Human Factors in Computing Systems* 1–21 (2025).

18. Lepp, H. & Sarin, P. A global AI community requires language-diverse publishing. Preprint at https://arxiv.org/abs/2408.14772 (2024).
19. Van Rossum, D. Generative AI top 150: the world's most used AI tools. *FlexOS* https://www.flexos.work/learn/generative-ai-top-150 (2024).
20. MacroPolo. The flobal AI talent tracker https://archivemacropolo.org/interactive/digital-projects/the-global-ai-talent-tracker/ (2024).
21. Wiley. ExplanAItions: an artificial intelligence study by Wiley. https://www.wiley.com/en-us/ai-study (2023).
22. Bianchini, S., Müller, M. & Pelletier, P. Drivers and barriers of ai adoption and use in scientific research. Preprint at https://arxiv.org/abs/2312.09843 (2023).
23. Horowitz, M. C., Kahn, L., Macdonald, J. & Schneider, J. Adopting AI: how familiarity breeds both trust and contempt. *AI Soc.* **39**, 1721–1735 (2024).
24. Topsakal, Y. How familiarity, ease of use, usefulness, and trust influence the acceptance of generative artificial intelligence (AI)-assisted travel planning. *Int. J. Hum. Comput. Interact.* https://doi.org/10.1080/10447318.2024.2426044 (2024).
25. Lavergne, T., Urvoy, T. & Yvon, F. Detecting fake content with relative entropy scoring. *Pan* https://dl.acm.org/doi/10.5555/3053718.3053722 (2008).
26. Badaskar, S., Agarwal, S. & Arora, S. Identifying real or fake articles: towards better language modeling. In *International Joint Conference on Natural Language Processing* https://aclanthology.org/I08-2115/ (2008).
27. Beresneva, D. Computer-generated text detection using machine learning: a systematic review. In *International Conference on Applications of Natural Language to Data Bases* https://doi.org/10.1007/978-3-319-41754-7_43 (Springer, 2016).
28. Solaiman, I. et al. Release strategies and the social impacts of language models. Preprint at https://arxiv.org/abs/1908.09203 (2019).
29. Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D. & Finn, C. DetectGPT: zero-shot machine-generated text detection using probability curvature. In *Proc. 40th International Conference on Machine Learning* Vol. 202 (eds. Krause, A. et al.) 24950–24962 (PMLR, 2023).
30. Yang, X., Cheng, W., Petzold, L., Wang, W. Y. & Chen, H. DNA-GPT: divergent N-gram analysis for training-free detection of GPT-generated text. In *The Twelfth International Conference on Learning Representations* https://openreview.net/forum?id=Xlayxj2fWp(2024).
31. Bao, G., Zhao, Y., Teng, Z., Yang, L. & Zhang, Y. Fast-DetectGPT: efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations* https://openreview.net/forum?id=Bpcgcr8E8Z (ICLR, 2024).
32. Tulchinskii, E. et al. Intrinsic dimension estimation for robust detection of AI-generated texts. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)* https://openreview.net/pdf?id=8uOZ0kNji6 (NeurIPS, 2025).
33. Bhagat, R. & Hovy, E. H. Squibs: what is a paraphrase? *Comput. Linguist.* **39**, 463–472 (2013).
34. Zellers, R. et al. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* https://papers.neurips.cc/paper_files/paper/2019/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf (NeurIPS, 2019).
35. Bakhtin, A. et al. Real or fake? Learning to discriminate machine from human generated text. Preprint at https://arxiv.org/abs/1906.03351 (2019).
36. Uchendu, A., Le, T., Shu, K. & Lee, D. Authorship attribution for neural text generation. In *Conference on Empirical Methods in Natural Language Processing* https://aclanthology.org/2020.emnlp-main.673/ (2020).
37. Chen, Y. et al. GPT-Sentinel: distinguishing human and ChatGPT generated content. Preprint at https://arxiv.org/abs/2305.07969 (2023).
38. Yu, X. et al. GPT Paternity Test: GPT generated text detection with GPT genetic inheritance. Preprint at https://ar5iv.labs.arxiv.org/html/2305.12519 (2023).
39. Li, Y. et al. MAGE: Machine-generated Text Detection in the Wild. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) 36–53 (ACL, 2024).
40. Liu, X., Zhang, Z., Wang, Y., Lan, Y. & Shen, C. CoCo: coherence-enhanced machine-generated text detection under data limitation with contrastive learning. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing* https://aclanthology.org/2023.emnlp-main.1005.pdf (ACL, 2023).
41. Bhattacharjee, A., Kumarage, T., Moraffah, R. & Liu, H. ConDA: contrastive domain adaptation for AI-generated text detection. In *Proc. 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* (Volume 1: Long Papers) https://aclanthology.org/2023.ijcnlp-main.40/ (ACL, 2023).
42. Hu, X., Chen, P.-Y. & Ho, T.-Y. RADAR: robust AI-text detection via adversarial learning. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)* https://openreview.net/forum?id=QGrkbaan79 (NeurIPS, 2023).
43. Wolff, M. Attacking neural text detectors. Preprint at https://arxiv.org/abs/2002.11768 (2022).
44. GPT-2: 1.5B release. *OpenAI* https://openai.com/research/gpt-2-1-5b-release (2019).
45. Jawahar, G., Abdul-Mageed, M. & Lakshmanan, L. V. Automatic detection of machine generated text: a critical survey. In *Proc. 28th International Conference on Computational Linguistics* https://aclanthology.org/2020.coling-main.208.pdf (ACL, 2020).
46. Fagni, T., Falchi, F., Gambini, M., Martella, A. & Tesconi, M. TweepFake: about detecting deepfake tweets. *Plos ONE* **16**, e0251415 (2021).
47. Ippolito, D., Duckworth, D., Callison-Burch, C. & Eck, D. Automatic detection of generated text is easiest when humans are fooled. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* https://aclanthology.org/2020.acl-main.164/ (ACL, 2019).
48. Gehrmann, S., Strobelt, H. & Rush, A. M. GLTR: statistical detection and visualization of generated text. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* 111–116 (2019).
49. Heikkilä, M. How to spot AI-generated text. *MIT Technology Review* https://www.technologyreview.com/2022/12/19/1065596/how-to-spot-ai-generated-text/ (2022).
50. Crothers, E., Japkowicz, N. & Viktor, H. Machine generated text: a comprehensive survey of threat models and detection methods. Preprint at https://arxiv.org/abs/2210.07321 (2022).
51. Kirchner, J. H., Ahmad, L., Aaronson, S. & Leike, J. New AI classifier for indicating AI-written text. *OpenAI* https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/ (2023).
52. Kelly, S. M. ChatGPT creator pulls AI detection tool due to 'low rate of accuracy'. *CNN Business* https://www.cnn.com/2023/07/25/tech/openai-ai-detection-tool/index.html (2023).
53. Weber-Wulff, D. et al. Testing of detection tools for AI-generated text. *Int. J. Educ. Integ.* **19**, 26 (2023).
54. Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W. & Feizi, S. Can AI-generated text be reliably detected? In *The Twelfth International Conference on Learning Representations* https://openreview.net/forum?id=OOgsAZdFOt (2023).

55. Chakraborty, S. et al. On the possibilities of AI-generated text detection. In *Proc. 41st International Conference on Machine Learning Research* https://proceedings.mlr.press/v235/chakraborty24a.html (2024).
56. Lo, K., Wang, L. L., Neumann, M., Kinney, R. & Weld, D. S2ORC: The semantic scholar open research corpus. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 4969–4983 (Association for Computational Linguistics, 2020); https://doi.org/10.18653/v1/2020.acl-main.447
57. Jurafsky, D. & Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* 3rd edn (Prentice Hall, 2025).
58. arXiv.org submitters. arxiv dataset. *kaggle* https://www.kaggle.com/datasets/Cornell-University/arxiv/versions/165 (2024).
59. Fraser, N. rbiorxiv: Client for the 'bioRxiv' API. *GitHub* https://github.com/nicholasmfraser/rbiorxiv (2024).
60. Nature portfolio. *Springer Nature* https://www.nature.com/nature-portfolio (2025).
61. Liang, W. Mapping the increasing use of LLMs in scientific papers. *COLM* https://openreview.net/forum?id=YX7QnhxESU (2024).

## Acknowledgements

We thank D. A. McFarland, D. Jurafsky, Y. Yin, Z. Izzo, X. V. Lin, L. Chen and H. Ye for their helpful comments and discussions. J.Z. is supported by the National Science Foundation (grant nos. CCF 1763191 and CAREER 1942926), the US National Institutes of Health (grant nos. P30AG059307 and U01MH098953) and grants from the Silicon Valley Foundation and the Chan-Zuckerberg Initiative. H.L. is supported by the National Science Foundation (grant nos. 2244804 and 2022435) and the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

## Author contributions

W.L. and Y.Z. designed the study and oversaw the quantification analysis. W.L. and Y.Z. provided the code for data analysis and conducted the analysis. W.L., Y.Z., Z.W., H.L., W.J., X.Z. and H.C. wrote the paper, with substantial input from all authors. All authors contributed to the review and editing of the paper. D.Y., C.P., C.D.M. and J.Z. provided the overall direction and planning of the project.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41562-025-02273-8.

**Correspondence and requests for materials** should be addressed to Weixin Liang or James Zou.

**Peer review information** *Nature Human Behaviour* thanks Casey Greene, Phillip Howard and Ruixiang Tang for their contribution to the peer review of this work.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature portfolio

Corresponding author(s): Weixin Liang, James Zou

Last updated by author(s): May 26, 2025

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| □ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☒ | □ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | □ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | □ | A description of all covariates tested |
| ☒ | □ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| □ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | □ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | □ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | □ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | □ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

Data collection

The arXiv Dataset can be accessed through https://www.kaggle.com/datasets/Cornell-University/arxiv , we use https://github.com/allenai/s2orc-doc2json to parse the latex sources of arXiv papers. The bioRxiv Dataset is collected through https://github.com/nicholasmfraser/biorxiv which is a R client for interacting with the bioRxiv API. The Nature portfolio Dataset is collected using a Python script we wrote ourselves. Data collection is performed using Python version 3.8.19 and R version 4.4.1 .

Data analysis

Data is analyzed with customized code in Python 3.8.19. We implemented the framework proposed in the paper to analyze the data. The code can be accessed at https://github.com/Weixin-Liang/Mapping-the-Increasing-Use-of-LLMs-in-Scientific-Papers . We use Fast-DetectGPT (https://github.com/baoguangsheng/fast-detect-gpt), Deepfake Text Detect (https://github.com/yafuly/MAGE) and RADAR (https://github.com/IBM/RADAR) for baseline comparison.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The datasets analysed in the current study are public at the following links: for arXiv, https://www.kaggle.com/datasets/Cornell-University/arxiv; for bioRxiv, https://github.com/nicholasmfraser/rbiorxiv; for Nature portfolio, https://www.nature.com/nature-portfolio.

## Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

| | |
|---|---|
| Reporting on sex and gender | Not applicable. |
| Reporting on race, ethnicity, or other socially relevant groupings | Not applicable. |
| Population characteristics | Not applicable. |
| Recruitment | Not applicable. |
| Ethics oversight | Not applicable. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Study description | The data in the study are quantitative as we analyze the exact usage of LLM in academic writing. |
| Research sample | Preprints and published papers from January 2020 to September 2024 on arXiv, bioRxiv, and Nature portfolio journals. |
| Sampling strategy | For each dataset, we randomly sample a subset of papers per month from January 2020 to September 2024. The size of each subset is the minimum of 2000 and the total number of papers in that month. For arXiv Computer Science and Physics, the ratio of sample size to the total number of papers is approximately 25%; for arXiv Mathematics and bioRxiv, the ratio is about 50%; for datasets with smaller number of papers like Nature portfolio, arXiv Electrical Engineering and Systems Science, and arXiv Statistics, the ratio is 100%. The confidence interval produced by our framework also indicates the sufficiency of our sample size, with smaller error bars suggesting that the sample size is relatively sufficient. The sample is representative as the sampling approach uses random selection within each month, which helps ensure representativeness across time periods. The 2000-paper monthly cap balances analytical depth with inference time for large datasets. In addition, the confidence intervals validate sample sufficiency across all datasets. |
| Data collection | The arXiv Dataset can be accessed through https://www.kaggle.com/datasets/Cornell-University/arxiv , we use https://github.com/allenai/s2orc-doc2json to parse the latex sources of arXiv papers. The bioRxiv Dataset is collected through https://github.com/nicholasmfraser/rbiorxiv which is a R client for interacting with the bioRxiv API. The Nature portfolio Dataset is collected using a Python script we wrote ourselves. The researcher was blind to the study hypothesis during data collection. |
| Timing | For each dataset, we use data from January 2020 to September 2024. The data was collected in December 2024. |
| Data exclusions | No data were excluded from the analyses. |
| Non-participation | No participants were involved in the study. |

| | |
|---|---|
| Randomization | Our study analyzed existing academic papers rather than conducting a controlled experiment, so experimental group allocation was not applicable. We randomly sampled papers from academic databases (arXiv, bioRxiv, Nature Portfolio) between January 2020 and September 2024 to estimate temporal trends in LLM usage across different research fields. All subsequent comparisons were made among papers categorized by research field, author preprint frequency, and geographic region, etc. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study | |
|---|---|---|
| ☒ | ☐ | Antibodies |
| ☒ | ☐ | Eukaryotic cell lines |
| ☒ | ☐ | Palaeontology and archaeology |
| ☒ | ☐ | Animals and other organisms |
| ☒ | ☐ | Clinical data |
| ☐ | ☒ | Dual use research of concern |
| ☒ | ☐ | Plants |

## Methods

| n/a | Involved in the study | |
|---|---|---|
| ☒ | ☐ | ChIP-seq |
| ☒ | ☐ | Flow cytometry |
| ☒ | ☐ | MRI-based neuroimaging |

# Dual use research of concern

Policy information about dual use research of concern

## Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

| No | Yes | |
|---|---|---|
| ☒ | ☐ | Public health |
| ☒ | ☐ | National security |
| ☒ | ☐ | Crops and/or livestock |
| ☒ | ☐ | Ecosystems |
| ☒ | ☐ | Any other significant area |

## Experiments of concern

Does the work involve any of these experiments of concern:

| No | Yes | |
|---|---|---|
| ☒ | ☐ | Demonstrate how to render a vaccine ineffective |
| ☒ | ☐ | Confer resistance to therapeutically useful antibiotics or antiviral agents |
| ☒ | ☐ | Enhance the virulence of a pathogen or render a nonpathogen virulent |
| ☒ | ☐ | Increase transmissibility of a pathogen |
| ☒ | ☐ | Alter the host range of a pathogen |
| ☒ | ☐ | Enable evasion of diagnostic/detection modalities |
| ☒ | ☐ | Enable the weaponization of a biological agent or toxin |
| ☒ | ☐ | Any other potentially harmful combination of experiments and agents |

## Plants

| | |
|---|---|
| Seed stocks | Not applicable. |
| Novel plant genotypes | Not applicable. |
| Authentication | Not applicable. |