LLM-Generated or Human-Written?
Comparing Review and Non-Review Papers on ArXiv

Anonymous ACL submission

Abstract

ArXiv recently prohibited the upload of unpublished review papers to its servers in the Computer Science domain, citing a high prevalence of LLM-generated content in these categories. However, this decision was not accompanied by quantitative evidence. In this work, we investigate this claim by measuring the proportion of LLM-generated content in review vs. non-review research papers in recent years. Using two high-quality detection methods, we find a substantial increase in LLM-generated content across both review and non-review papers, with a higher prevalence in review papers. However, when considering the number of LLM-generated papers published in each category, the estimates of non-review LLM-generated papers are almost six times higher. Furthermore, we find that this policy will affect papers in certain domains far more than others, with the CS subdiscipline Computers & Society potentially facing cuts of 50%. Our analysis provides an evidence-based framework for evaluating such policy decisions.[1]

1 Introduction

On October 31st, 2025, arXiv announced a policy change[2] prohibiting the upload of review,[3] survey, and position papers to their Computer Science (CS) servers (Boboris, 2025). According to the blog post, this decision was motivated by an observed increase in LLM-generated content in review papers in the CS subdomain over recent years. Since papers uploaded to arXiv undergo moderation by volunteer human experts, overburdened with other obligations, the administrators decided to restrict further approvals of unpublished review papers altogether. The blog post did not provide quantitative evidence or data to support this claim. In this paper, we address two central questions: *Do review papers contain higher proportions of LLM-generated content than regular research papers? What downstream effects would be created by this policy?*

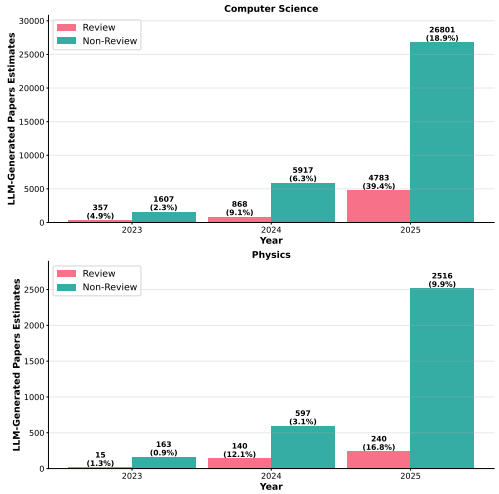


Figure 1: LLM-generated paper estimates using Pangram detection for review vs. non-review papers. While review papers show higher percentages of LLM-generated content (in parentheses), the absolute number of non-review papers detected as LLM-generated is substantially larger in all domains, suggesting that restricting review papers alone may not address the core concern about LLM-generated content on arXiv.

The proliferation of advanced AI writing assistants, particularly large language models (LLMs), has enabled researchers to incorporate these tools into their scientific workflows (Nejjar et al., 2025; Zhang et al., 2025; Eger et al., 2025; Si et al., 2025). In the extreme case, papers may be fully written by LLMs, including literature reviews and experimental design.[4] While we have preliminary evidence for fully LLM-generated papers receiving scores sufficient for workshop publication (Yamada et al., 2025), a central concern is whether LLM-generated papers maintain the same *standards* as

[1] We will release our code upon publication.

[2] While the blog post explicitly states that this is not a technical policy change, it is a de-facto change.

[3] Throughout this paper, we use the term *review papers* to refer to all three types of papers covered by arXiv's policy, which refers to "review, survey, and position" papers.

[4] Several instances of fully LLM-generated papers have been reported in the literature (Lu et al., 2024).

human-written papers. The ease with which papers can be generated, and the increasing rates of paper submissions, especially about AI topics, provoke concerns about overwhelm for arXiv moderators.

These are legitimate concerns, but the focus on review papers makes several assumptions that may or may not be accurate or helpful. First is the central assumption that review papers are more likely to be generated than non-review papers. Second is a set of epistemic assumptions about what constitutes a review paper and what purposes review papers serve in a research community, reflecting disciplinary biases (not shared across all CS subdisciplines) that elevate quantitative over qualitative methods and evidence.

In this paper, we quantitatively evaluate this new policy. We evaluate the ratios of review papers that are LLM-generated vs. non-review papers, the temporal trends of such ratios, and how these trends vary across fields. Our main analysis classification and detection pipeline uses only paper titles and abstracts (not full text), reflecting the data used by arXiv to make the initial decision. We then run an additional analysis on a subset of papers for which we run the analysis on the full text, where we validate the consistency of our results. As such, we employ a two-stage methodology. First, we develop an LLM-based classifier to distinguish review papers from regular research papers, achieving high accuracy (92.0% F1) on manually annotated data. Second, we apply two LLM-generation detection methods that estimate the fraction of LLM-generated content in a corpus to measure LLM prevalence. We analyze a large dataset of arXiv papers from different domains to measure these rates.

Our findings indicate that: (1) review papers contain higher rates of LLM-generated content than non-review papers in CS, with adjusted cohort-level estimates of 21.4% for review papers versus 14.0% for non-review papers; (2) LLM-generated content has increased dramatically in both categories following ChatGPT's release in November 2022, with review papers showing an increase from 12.9% in 2023 to 28.2% in 2025 and non-review papers from 6.2% to 18.9%; and (3) this pattern persists across all examined disciplines (Computer Science, Physics, Mathematics, Statistics) and shows an accelerating trend since 2023.

Crucially, when considering the number of papers published for each type, we estimate that in CS, there are almost six times more non-review papers generated by LLMs than review papers (Figure 1). This result throws into doubt arXiv's decision to restrict the upload of unpublished review papers to their CS servers. We also find that the policy will affect certain disciplines and topics differently; topics like education and safety have much higher rates of review papers, and Computers & Society papers could be censored by 50%, while Computer Vision will only lose 3% of its papers. We hope this work provides an evidence-based perspective on arXiv's policy decision while fostering broader discussion about the long-term implications of LLM-generated content in scientific publishing.

2 Background

A Brief History of ArXiv. Arxiv originally started as an automated email server before converting into a website. As its usage and submissions grew, arXiv faced challenges related to software scale and content moderation (Han, 2025).[5] While arXiv has historically relied on volunteer moderators to ensure basic quality standards, the scale and nature of submissions have evolved significantly, and as such, automatic classifiers were introduced to assist with moderation. The October 2025 policy restricting review papers in CS represents a notable departure from arXiv's traditional open policy.[6]

LLM-Generated Text Detection. The detection of LLM-generated text has become an active research area following the release of powerful language models (Brown et al., 2020; Achiam et al., 2023). Early approaches relied on statistical features such as perplexity and token probability distributions (Mitchell et al., 2023). More recent work has explored zero-shot detection methods (Mitchell et al., 2023), watermarking techniques (Kirchenbauer et al., 2023), and distributional estimation approaches that measure the fraction of AI content in document collections rather than classifying individual texts (Sadasivan et al., 2023). Beyond the design of accurate detection models, there is the question about the right task definition. For instance, Ghosal et al. (2023) argues that binary AI-versus-human classifications are insufficient, proposing instead to estimate the percentage of LLM-generated content within a document, with new definitions of such tasks and proposing novel methods (Wang et al., 2023, 2025; Thai et al., 2025).

[5] https://arxiv.org/stats/monthly_submissions
[6] https://info.arxiv.org/help/policies/

The *Alpha* estimator method (Liang et al., 2024a) represents a shift from instance-level to population-level detection. Rather than attempting to classify individual documents with high confidence, it uses maximum likelihood estimation to measure the fraction of LLM-generated content across an entire corpus. This approach is particularly well-suited for our research question, which focuses on comparing aggregate prevalence rates between paper categories.

LLMs in Academic Writing. Recent studies have documented the increasing use of LLM writing assistants and *agents* in academic contexts (Liang et al., 2024a). Concerns have been raised about the impact on scientific integrity, originality, and the peer review process (Lin et al., 2025; Latona et al., 2024). Several instances of fully or largely LLM-generated papers have been identified in published literature (Dupré, 2023), raising questions about quality control mechanisms.

Some work has examined disciplinary differences in AI adoption, finding that fields with greater exposure to AI technology show earlier and higher adoption rates (Liao et al., 2025). However, to our knowledge, no prior work has systematically compared generated content between review papers and regular research papers or provided quantitative evidence for preprint server policies.

3 Methods

3.1 LLM-Generated Content Detection

We focus on estimating population-level trends (e.g., the proportion of LLM-generated content among all review papers posted on arXiv in 2025) and comparing these trends across groups (e.g., review vs. non-review, 2023 vs. 2025). While this is a challenging task, and we do not expect perfect accuracy, we can estimate and compare aggregate statistics over populations.[7]

We employ two complementary detection methods: (1) the Alpha estimator (Liang et al., 2024a), which provides a group-level estimate of the fraction of LLM-generated content based on the occurrence of adjectives used in a document, and (2) a commercial LLM-detector called Pangram (Emi and Spero, 2024),[8] a transformer-based (Vaswani et al., 2017) classifier trained to distinguish LLM-generated text from human-written content across multiple domains and language models. These methods are complementary: they were tuned on different datasets, they operate on different inputs, and they model the problem differently.

Evaluation. Our estimators were previously evaluated and shown to be highly accurate (below 2.4% error rates for Alpha (Liang et al., 2024a), and below 1% for Pangram (Emi and Spero, 2024)). While errors during the classification of papers post-LLM era are practically impossible for us to evaluate as we do not have access to the ground truth, we evaluate them on papers from the pre-LLM era to measure false positive errors. We test both Alpha and Pangram on CS papers from 2020-2022. We find that while Pangram did not predict any false positives, the Alpha method predicts a non-trivial amount of LLM-generated papers in the pre-LLM era: 13.1% and 1.9% for review and non-review papers, respectively. As such, we employ a correction to the Alpha estimates to correct for such bias. While the pre-LLM period may still include usage of writing assistance (e.g., grammar/style editing tools, by tools such as Grammarly [9]), such tools typically minimally edited texts locally, and do not imply LLM text generation; we therefore treat the pre-LLM signal primarily as detector bias and correct for it explicitly.

Alpha Estimates. We employ a distributional detection approach using the Alpha estimator method (Liang et al., 2024a), which uses maximum likelihood estimation (MLE) to measure the fraction α of LLM-generated or LLM-modified content across an entire dataset (group). For each group we compute Alpha, representing the estimated ratio of LLM-generated papers, and the 95% confidence interval of the estimation.

Correction. Due to non-trivial rates of false positives in the Alpha estimation in the pre-LLM era, we use these values to calibrate our post-LLM estimates via the Rogan-Gladen adjustment (Rogan and Gladen, 1978). This adjustment takes into account false positive and false negative rates to correct apparent (inaccurate) prevalence measurements. We use the detection rates of the pre-LLM estimates for Alpha per group: review and non-review papers, to correct every group estimate. Additional details on the Rogan-Gladen adjustment as well as our implementation and assumptions are provided in Appendix E. For readability, we center

[7] Assuming that any systematic biases in the classifier affect both groups similarly.

[8] https://www.pangram.com/

[9] https://www.grammarly.com/

the main analysis on corrected estimates because the uncorrected Alpha values are known to be biased upward in our validation.

Pangram Estimates. We use the Pangram API (Emi and Spero, 2024) and treat "Likely AI", "Possibly AI", and "Highly Likely AI" labels as LLM generated papers. "Unlikely AI" classifications are treated as human written. For each group, we compute the proportions of LLM-generated papers, and the standard error of the mean of the distribution.

3.2 Review Paper Classification

The arXiv policy relies on distinguishing between "review" vs. non-review papers. But what is a "review" paper? While this question might seem obvious, such taxonomies are neither universal nor neutral; different fields can have different epistemic cultures and "machineries of knowledge construction" (Cetina, 1999), and interdisciplinary work often resists such categorizations (Clement, 2016).

ArXiv's CS corpus spans communities with different epistemic cultures and publication norms, from established subfields like Programming Languages (CS.PL) to interdisciplinary areas like Computation & Language (CS.CL) and Computers & Society (CS.CY). The latter two include papers from the digital humanities (DH), computational social science, and fairness, accountability, and transparency research (FAccT), where community norms are known to clash (Laufer et al., 2022).

While we do not know exactly how arXiv moderators categorize review papers, we approximate what such a process might look like.[10]

What are Review Papers? *Review papers*, which we consider to include *survey papers*, often aim to synthesize and summarize a body of literature, identify trends over time, provide recommendations and guidelines for best practices, and introduce newcomers to a field. Structurally, review papers tend to emphasize literature coverage over novel experimental contributions, though high-quality reviews require substantial domain expertise. Some reviews offer novel insights through meta-analysis, and many different kinds of review papers exist, ranging from critical and mixed-method reviews to standardized Cochrane reports (Grant and Booth, 2009). *Position papers* (included in arXiv's ban) often function like editorials; they offer opinionated perspectives on current issues in a discipline and tend to rely more heavily on literature review than experimental results. However, these are generalizations based on traditional CS disciplines like machine learning, and we emphasize that this definitional task represents epistemic choices that are dependent on discipline.

Review vs. Non-Review Dataset. To test the feasibility of the task and to evaluate the performance of our review paper classifier, we create a manually annotated dataset sampled from the arXiv Computer Science category across different years. We hand-annotate a set of 200 papers as review or non-review papers. These papers were sampled such that 100 of them had keywords related to reviews[11] in their title or abstract, while the other 100 papers included "experimental" keywords.[12] Two of the paper's authors, both with NLP and CS expertise, discussed examples and definitions and then independently labeled each paper based on its title and abstract, achieving strong agreement (Cohen's κ = 0.85). We use 60 papers for evaluation, and 140 papers as a test set. We use one annotator's labels as the ground truth for evaluation purposes.

Evaluation. We evaluate the following LLMs for classifying review vs. non-review: Llama 3.3 (Dubey et al., 2024), Gemma 2 (Team et al., 2024), GPT-OSS (Agarwal et al., 2025), and GPT-4o-mini (Hurst et al., 2024) on our validation set, achieving 95.4, 95.6, 92.6, and 95.6 F1 scores, respectively. We select Gemma 2 as our classification model as it achieved the best performance on the validation set, and we can run it locally. We report the results on the held-out test set in Table 3, as well as the prompt used for all models in Appendix B.

3.3 Quantifying LLM-Generated Content

We apply the review paper classifier to a large corpus of arXiv papers, classifying them into review and non-review papers. We then apply the AI content detector to both groups[13] in each category and perform statistical significance tests: 95% confidence intervals over $30K$ bootstrapped samples for Alpha following Liang et al. (2024a), and standard error of the mean for Pangram. Our manual

[10] However, any such classifier will necessarily impose one epistemic culture's categories onto papers from others.

[11] 'review', 'survey', 'overview', 'primer', 'systematic review', and 'literature review'.

[12] 'results', 'experimental', 'empirical', 'evaluation', and 'dataset'.

[13] Using Rogan-Gladen adjustment to correct for the false positive of the Alpha method, and simply computing the ratio of LLM-generated papers using Pangram.

ground-truth annotation is CS-only by design, because the arXiv policy under study currently targets CS; we then apply the same title/abstract-based classifier across other domains for comparative context rather than policy-specific labeling claims about those domains. We use the entire dataset to compute the estimates using alpha, while using a sample of 100 papers from every month with Pangram, to save costs, for a total of 3,600 per domain for the post-LLM period.

3.4 Datasets

We use an official arXiv dataset released on Kaggle by Cornell University,[14] which contains metadata, abstracts and links to the PDF files for posted papers, in English. We focus on papers from Computer Science, Mathematics, Statistics, and Physics, published between 2020 and 2025, allowing us to evaluate and correct the estimates using the pre-LLM (2020-2022) era and to perform our main analysis on the post-LLM (2023-2025) era.

Our first dataset, *arxiv-domains*, consists of four popular domains (CS, Mathematics, Statistics, and Physics), with up to $36,000$ papers from arXiv (500 per month, for every month between 2020-2025 $= 500*12*6$), including metadata: titles, abstracts, publication dates, and categories. Overall, we collect $124,461$ papers for *arxiv-domains*. In addition, we collect a dataset of CS subcategories,[15] which we name *cs-subcategories*, sampling up to 500 papers per month per subcategory, up to 36,000 per subcategory. Overall, we collect $138,244$ papers for *cs-subcategories*. We provide more detailed statistics in Appendix C. Unless noted otherwise, we use titles and abstracts for classification and detection, and we discuss this decision in §5. We also experiment with analyzing the full texts in §4.5.

4 Results

4.1 Rates Vary Dramatically By Discipline

Figures 2-3, as well as 5-6 in the Appendix, show that review papers tend to have higher rates of LLM-generated content than non-review papers in CS. This trend holds across domains and when using both the Alpha estimates and Pangram. We elaborate on these results in Appendix A.

4.2 Upward Trends in Generated Papers

We then examine how these patterns have shifted over time. Figure 2 shows the LLM-generated content estimates by year. We observe a dramatic increase in LLM-generated content since the release of ChatGPT between 2023-2025, consistent with the findings of Liang et al. (2024b).

In CS, the Alpha in review papers increases from 12.9% in 2023 to 28.2% in 2025. For non-review papers, the increase was from 6.2% in 2023 to 18.9% in 2025. The gap between review and non-review papers has remained substantial and significant in all years. The Pangram results show similar, but more dramatic trends. In CS, the percentage of AI detected papers grew from 7.4% in 2023 to 43.3% in 2025 in review papers, compared with 2.7% to 23.3% in non-review papers.

4.3 Computer Science Subcategory Analysis

We conduct a fine-grained analysis examining AI adoption patterns across different CS subcategories. Overall, we find large variations in review paper and AI-generated paper rates. Table 1 shows that Computers & Society has a far higher rate of review papers than other subcategories (49.2%, compared to the next highest category, Software Engineering, with 23.4%). We find a high Pearson correlation ($\rho = 0.886$) between the review and LLM-generated paper rates.

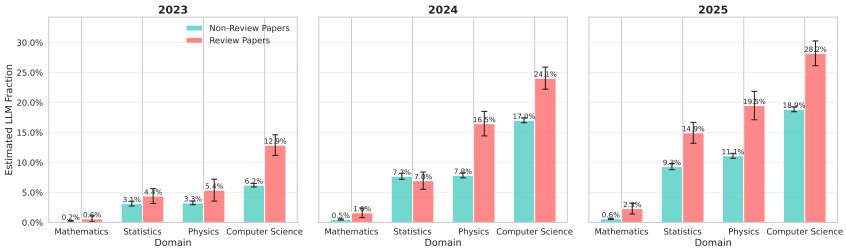
Figure 3 presents the Alpha and Pangram estimates for review and non-review papers across major CS subcategories. Overall, the predicted rates for review papers are always higher or equivalent to the predicted rates for non-review papers. The analysis reveals some variance within CS, with some subcategories showing much higher AI adoption rates than others. For example, Computation and Language (the NLP subcategory) shows the lowest use of AI in review papers (22.06% and 10.88% using the Alpha and Pangram estimates, respectively). On the other hand, Cryptography and Security shows the highest adoption rates for review papers, with estimates of 31.35% and 24.68% using Alpha and Pangram.[16]

Figure 6 in the Appendix further breaks down these patterns over time. These subcategory-level findings have important policy implications.

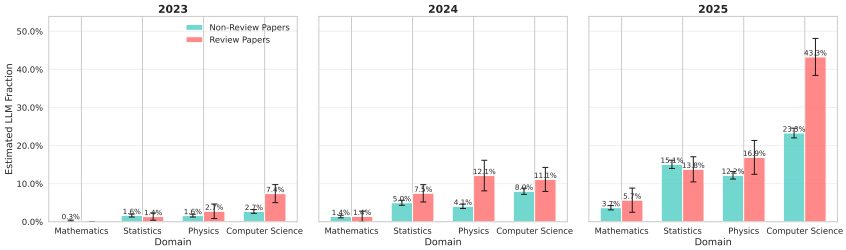
[14] https://www.kaggle.com/datasets/Cornell-University/arxiv, sampled on December 15th, 2025.

[15] The ten most common CS subcategories: *Artificial Intelligence*, *Computation and Language*, *Cryptography and Security*, *Computer Vision and Pattern Recognition*, *Computers and Society*, *Human-Computer Interaction*, *Information Retrieval*, *Machine Learning*, *Robotics*, and *Software Engineering*

[16] This result is unintuitive, and we leave it as an open question for future work.



(a) Alpha estimates.



(b) Pangram ratios.

Figure 2: LLM-generated papers estimates using (a) Alpha estimates and (b) Pangram detection ratios on the *arxiv-domains* dataset. Both methods reveal temporal patterns in LLM-adoption across domains, and years.

Category	Review Rate	AI Rate
Computers & Society	49.2%	20.2%
Software Engineering	23.4%	10.6%
HCI	23.3%	11.3%
AI	15.8%	13.0%
Cryptography & Security	15.2%	13.8%
Information Retrieval	13.4%	7.8%
Computation & Language	9.4%	8.8%
Machine Learning	5.7%	9.9%
Robotics	3.7%	6.7%
Computer Vision	2.4%	7.0%

Table 1: Review vs. AI-generated rates across CS in the *cs-subcategories* dataset. The Pearson correlation between the review and AI columns is high ($\rho = 0.886$).

Topic	Review Rate	AI Rate
Safety Research	61.4 %	13.9 %
Health Informatics	44.0 %	20.1 %
Sociology & Political Sci.	32.4 %	10.7 %
Management Info. Systems	32.1 %	17.7 %
Education	31.0 %	13.3 %
CV & Pattern Recog.	5.3 %	5.2 %
Ocean Engineering	4.5 %	8.2 %
Control & Systems Engin.	4.0 %	8.6 %
Aerospace Engineering	4.0 %	6.0 %
Biomedical Engineering	3.2 %	7.1%

Table 2: The ten OpenAlex topics with the highest and lowest predicted review paper rates in *cs-subcategories*.

ArXiv's blanket restriction on CS review papers treats all CS subcategories uniformly, despite dramatic differences in AI adoption rates and community standards within CS.

4.4 Who is Affected by the Policy Change?

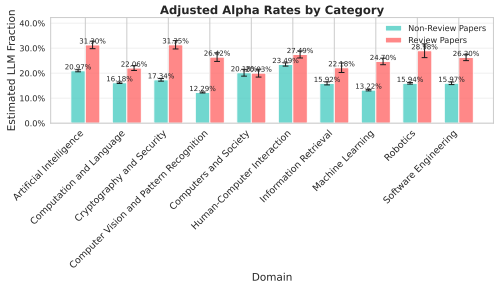
Using the OpenAlex API,[17] we explore what kinds of papers and authors are more affected by the arXiv policy change. For each paper in the *cs-subcategories* dataset (~3.6k papers per CS subcategory), we use the paper's DOI to collect the paper's topics, keywords, citation count, and authors, as well as each author's affiliation, h-index, number of works, and citations count, and each institution's country and h-index. We identify 74% of the papers, and of these, 89% had identifiable topics. See Appendix F for more details.

We find that junior authors ($h <= 5$) are more likely to write abstracts that are predicted to be generated by Pangram, and we find that authors of review papers are more likely to be from the Global North and affiliated with "elite" ($h > 500$) institutions. Finally, papers associated with topics like "safety research," "health informatics," and "education" are more likely to be review papers.

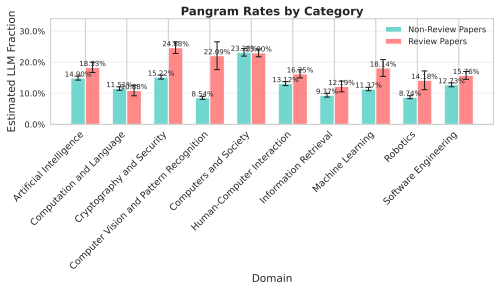
4.5 Do our results change for full texts?

So far, we have focused on classifying papers into LLM-generated vs. human-generated (and review vs. non-review) based on the title and abstracts alone. While this decision may seem naive at first, it is a deliberate one. First, fully LLM-generated papers also include the abstract, and as such, while we may overestimate the usage of LLMs in paper writing, we are less likely to miss any. Second, and perhaps most important for the context of this work, is that arXiv is basing its initial automated decisions for the types of papers (review vs. non-

[17] OpenAlex is an online library: https://openalex.org/.



(a) Alpha estimator method



(b) Pangram detection method

Figure 3: LLM-generated papers estimates by subcategories and paper type in *cs-subcategories* using (a) Alpha estimates and (b) Pangram detection ratios. The review versus non-review paper gap (or lack thereof) varies considerably across subcategories, suggesting that field-specific norms influence AI adoption patterns.

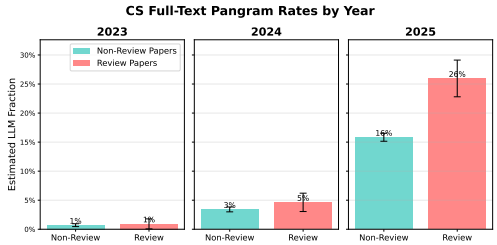


Figure 4: LLM-generated papers estimates by year in CS (*arxiv-domains*) using Pangram, when considering the entire text. While the percentages drop by about half compared to the abstract estimates, the yearly upward trend, as well as the review vs. non-review gap persist.

review) on the title and abstract alone.[18] As such, the implications based on our results reflect the new arXiv policy.

To validate that this simplification produces reliable results, we extend our analysis and quantify LLM-generated content using the entire paper, instead of solely the title and abstract. In this experiment, we use the post-LLM era CS papers that Pangram classified as LLM-generated. We successfully downloaded and extracted the full text from 430 of the 437 papers from that period. Using Pangram, we find that 57.9% of the full texts were predicted as LLM-generated. Split by paper type, we find that 216 out of 365 non-review papers (59.2%) are classified as LLM-generated, compared to 33 out of 65 of review papers (50.8%) (again, of the papers already classified by Pangram as being generated). These results suggest that review papers tend to use LLMs for generating abstracts over the entire paper more often then non-review papers. Using the full texts, we also confirm the trends reported in §4.2 and find that overall, review papers are more likely to be predicted as LLM-generated, and these rates are increasing over time (see Figure 4). For instance, 26% of the CS papers in 2025 were classified as review, vs. 16% of the non-review papers. These results provide additional support for questioning the new arXiv policy.

5 Discussion

Our findings provide quantitative evidence that review papers on arXiv contain higher proportions of LLM-generated content than non-review papers, *seemings* to support arXiv's stated rationale for their policy change. However, further analysis of the results reveals crucial nuances and unintended effects that warrant reconsideration of this policy.

Magnitude of the Difference. Over the past three years, we estimate the use of LLMs in CS papers to be 21.4% in review papers compared to 14.0% in non-review papers. Considering the per-year analysis, this gap increases considerably (e.g., a 6.7% difference in 2023 to 9.3% in 2025 using the Alpha estimate, and 4.7% in 2023 to 20.0% using Pangram). Crucially, however, **when considering the number of papers from each type, the worry about LLM usage in reviews fades completely.** We estimate the number of non-review papers to be 141K compared to 12K review papers in CS in 2025. Thus, even though the difference in review percentages is substantial in 2025 (43.3% vs. 23.3% using Pangram), **the number of review papers generated by LLMs is estimated to be 4,783 compared to 26,801 for non-review papers** (Figures 1 & 11, and Tables 6-9 in Appendix D.). As such, the arXiv ban on unpublished review papers makes little sense, if the goal is to save moderators time and energy, as the overall burden from review papers is tiny in comparison to the very large number of non-review papers using AI in paper writing.

[18] In conversations with a contact internal to arXiv, we were told that the initial filter mechanism is done based on the title and abstract alone. Flagged papers are then passed to a human moderator for a final decision.

Disciplinary Variations Matter. The cross-domain analysis (Figures 2 and 5) reveals that AI adoption patterns vary considerably across domains. CS shows the highest prevalence, unsurprising given the domain's focus on AI technology and early access to these tools. However, the review versus non-review paper gap persists across most examined domains, suggesting this is a general phenomenon rather than one specific to CS.

The temporal patterns are particularly revealing. The sharp increase following ChatGPT's release in late 2022 is evident across all domains, but the rate of adoption differs. CS shows earlier and faster adoption, while traditional domains like Mathematics show more gradual increases. This suggests that domain-specific norms and tool familiarity might influence AI adoption rates.

The validation of these subcategory patterns through Pangram analysis adds an important dimension to understanding AI adoption across CS communities. Crucially, the interdisciplinary subcategory Computers & Society displays much higher rates of review papers, meaning that **under the new policy, nearly 50% of Computers & Society papers would be censored by arXiv, in contrast to only 3% of Computer Vision papers.** Our analysis of topics from OpenAlex reemphasizes this point, as we found that **papers about education and safety (timely and important topics as society grapples with the effects of AI) will be censored at much higher rates than papers focused on engineering topics**.

Implications for Policy. Our results raise serious questions about arXiv's decision to restrict review papers. Considering the raw counts, we would be better off banning non-review papers if our goal is to save moderators time. And if the concern is LLM-generated content quality, should the policy focus solely on review papers when non-review papers also show substantial (albeit lower) AI content? More broadly, given the sharp increases in AI usage across fields, we need a general policy on the usage of AI in academic writing, beyond CS.

An alternative interpretation is that arXiv's policy targets not the use of LLMs in itself but the *type* of content that CS arXiv moderators perceive as "easiest" to generate with current AI systems—synthesis and summarization. While we do not investigate such claims in this paper, as research improves in developing automated scientists (Lu et al., 2024; Yamada et al., 2025), the ease of fully generating AI "regular" papers decreases substantially, making this assumption questionable.

The Future of Scientific Publishing. Our cross-disciplinary findings suggest that LLM adoption in academic writing is not a temporary phenomenon but an ongoing trend that varies across fields. As LLM writing tools improve and become more integrated into more research workflows, we may see convergence in adoption rates across disciplines. On the other hand, as academic communities develop more established norms (ethical or otherwise), we may see further divergence. Finally, these patterns may evolve as both LLMs and detection models improve and as authors develop more sophisticated usage strategies.

Publishers and preprint servers will need to develop clear, well-thought-out, and reasonable policies that account for disciplinary differences while maintaining scientific integrity and considering the development and incorporation of AI-assisted tools. The heterogeneity revealed by our subcategory analyses suggests that blanket restrictions may be overly broad, while the substantial LLM content in non-review papers indicates that concerns should extend beyond review papers.

6 Conclusions

Our study evaluates the evidence behind arXiv's 2025 restriction on review papers in the CS domain. Across two independent LLM detectors (Alpha and Pangram) and an accurate review vs. non-review classifier, we find that review papers consistently contain higher proportions of LLM-generated content. However, the much larger volume of non-review papers means that they account for the vast majority of LLM-generated manuscripts on arXiv, challenging the premise of a review-only ban.

Temporal and disciplinary analyses reveal a sharp post-ChatGPT rise in AI-assisted writing that spans CS, Physics, Mathematics, and Statistics, with heterogeneous uptake across CS subfields. These patterns suggest that LLM-assisted writing is a broad, accelerating shift rather than a CS or review-specific anomaly. Blanket restrictions on review papers risk missing most LLM-generated content while disproportionately affecting communities where synthesis and qualitative work are core scholarship. We advocate for evidence-based moderation that combines transparent monitoring with clearer guidance on acceptable LLM assistance.

Limitations

Our study has several important limitations that should be considered when interpreting the results.

Detection Method Limitations. The Alpha estimator, while powerful for group-level estimation, relies on distributional assumptions about LLM-generated text. The method may be sensitive to changes in the LLM used for generation, the data used to estimate alpha, or post-editing by human authors. Our pre-LLM baseline validation suggests the Alpha estimator has differential false positive rates for review and regular papers, which we address through type-specific Rogan-Gladen adjustment. However, this assumes the false positive pattern remains stable over time and does not interact with actual AI use. On the other hand, we detected no false positives with the Pangram detector.

Classification Accuracy. Our review vs. non-review paper classifier achieves a high classification score (92.0% F1), but misclassifications could still bias our results. We validated the classifier on held-out data, but some borderline cases (e.g., tutorial papers, hybrid papers with both novel contributions and community norms) may be ambiguous. In general, our annotations reflect an attempt at a "typical" CS view of what a "review" paper might represent, as this is likely the stance of many arXiv moderators, rather than trying to represent the many epistemic views of different subfields.

Interdisciplinary Papers. Papers in interdisciplinary areas of CS often blend methods and place greater emphasis on position-taking and literature synthesis, even in empirical work. A single paper might present an argument about how the field should develop, review existing literature to support that argument, and include empirical analysis to substantiate its claims. These functions are not mutually exclusive; they can coexist within a paper, which may be published in general venues rather than isolated in, e.g., "position paper tracks." While this is a limitation for our work—we had to rely on a simple definition of review papers for our annotations—it is an even larger limitation for the arXiv policy that will need to moderate, across a large number of papers, which papers are reviews.

Data Sampling A key limitation of our work is that our data is restricted to papers *posted* to arXiv rather than *submitted* to arXiv. This data is not publicly available, and our attempts to request this data were not successful. We therefore cannot provide quantitative evidence of the number of papers or distribution of papers that are submitted to arXiv but then flagged by moderators to never be posted. This likely means that our results in this paper are lower bounds for the rates of generated papers, as we believe some level of moderation has been in place to detect these. But we have no knowledge of any preceding policy about review papers until the current proposal.

Ethical Considerations. Labeling content as "LLM-generated" can be inaccurate and carries potential stigma. We emphasize that AI assistance in writing is not inherently problematic—many legitimate uses exist, including improving clarity for non-native English speakers. A limitation of current LLM-detection methods is that they cannot distinguish between different types of LLM usage in writing. For instance, we may want to treat differently papers that were fully written by LLMs versus those where LLMs were used to assist with writing. Our goal is to provide empirical evidence for policy discussions, not to judge individual authors or papers.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Kat Boboris. 2025. Attention authors: Updated practice for review articles and position papers in arxiv cs category. https://blog.arxiv.org/2025/10/31/attention-authors-updated-practice-for-review-articles-and-position-papers-in-arxiv-cs-category/. Accessed: 2025-12-05.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Karin Knorr Cetina. 1999. *Epistemic cultures: How the sciences make knowledge*. harvard university press.

Tanya Clement. 2016. Where is methodology in digital humanities? *Debates in the digital humanities*, 2.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Maggie Harrison Dupré. 2023. Researchers reveal that paper about academic cheating was generated using chatgpt. Accessed: 2025-01-05.

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, and 1 others. 2025. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *arXiv preprint arXiv:2502.05151*.

Bradley Emi and Max Spero. 2024. Technical report on the pangram ai-generated text classifier. *arXiv preprint arXiv:2402.14873*.

Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit Bedi. 2023. A survey on the possibilities & impossibilities of ai-generated text detection. *Transactions on Machine Learning Research*.

Maria J Grant and Andrew Booth. 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health information & libraries journal*, 26(2):91–108.

Sheon Han. 2025. Inside arxiv–the most transformative platform in all of science. Accessed January 5, 2026.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR.

Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West. 2024. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *Preprint*, arXiv:2405.02150.

Benjamin Laufer, Sameer Jain, A. Feder Cooper, Jon Kleinberg, and Hoda Heidari. 2022. Four years of facct: A reflexive, mixed-methods analysis of research contributions, shortcomings, and future prospects. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 401–426, New York, NY, USA. Association for Computing Machinery.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, and 1 others. 2024a. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. In *International Conference on Machine Learning*, pages 29575–29620. PMLR.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, and 1 others. 2024b. Mapping the increasing use of llms in scientific papers. In *First Conference on Language Modeling*.

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X Zhang. 2025. LLMs as research tools: A large scale survey of researchers' usage and perceptions. In *Second Conference on Language Modeling*.

Dingkang Lin, Naixuan Zhao, Dan Tian, and Jiang Li. 2025. Chatgpt as linguistic equalizer? quantifying llm-driven lexical shifts in academic writing. *Preprint*, arXiv:2504.12317.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*, pages 24950–24962. PMLR.

Mohamed Nejjar, Luca Zacharias, Fabian Stiehle, and Ingo Weber. 2025. Llms for science: Usage for code generation and data analysis. *Journal of Software: Evolution and Process*, 37(1):e2723.

Walter J Rogan and Beth Gladen. 1978. Estimating prevalence from the results of a screening test. *American journal of epidemiology*, 107(1):71–76.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Katherine Thai, Bradley Emi, Elyas Masrour, and Mohit Iyyer. 2025. Editlens: Quantifying the extent of ai editing in text. *arXiv preprint arXiv:2510.03154*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. Seqxgpt: Sentence-level ai-generated text detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1144–1156.

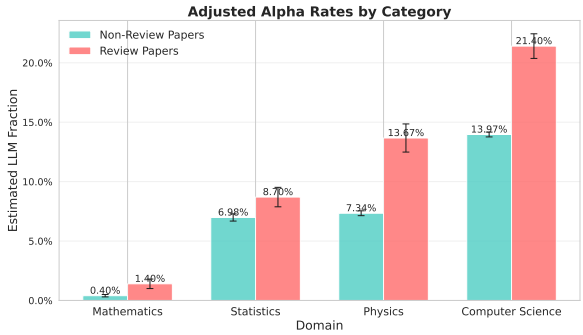
Yitong Wang, Zhongping Zhang, Margherita Piana, Zheng Zhou, Peter Gerstoft, and Bryan A Plummer. 2025. Real, fake, or manipulated? detecting machine-influenced text. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15022–15037.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*.

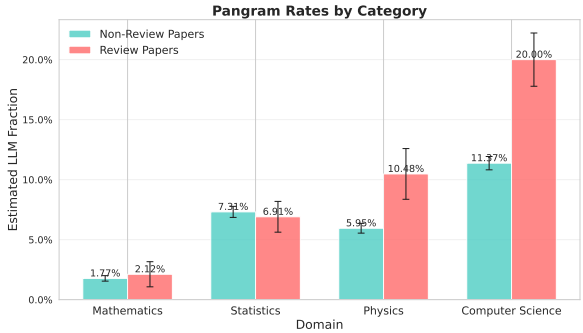
Yanbo Zhang, Sumeer A Khan, Adnan Mahmud, Huck Yang, Alexander Lavin, Michael Levin, Jeremy Frey, Jared Dunnmon, James Evans, Alan Bundy, and 1 others. 2025. Advancing the scientific method with large language models: From hypothesis to discovery. *arXiv preprint arXiv:2505.16477*.

Model	Precision	Recall	F1
Llama 3.3	92.6	87.7	90.1
Gemma 2	92.9	91.2	92.0
GPT-OSS	93.9	80.7	86.8
GPT-4o-mini	92.9	91.2	92.0

Table 3: Review vs. Non-Review paper classification performance of different models on the test set.



(a) Alpha estimates.



(b) Pangram ratios.

Figure 5: LLM-generated papers estimates by domain and paper type on the *arxiv-domains* dataset using (a) Alpha estimates and (b) Pangram detection ratios. Review papers consistently show higher LLM-generated ratios for CS and Physics, and mixed results for Statistics and Math depending on the method. Results are averaged across the post-LLM years (2023-2025).

A Rates Vary Dramatically By Discipline: Ellaboration

Computer Science. We find that review papers have significantly higher LLM-generated content than non-review papers in CS. In the post-LLM period (2023-2025), the Alpha estimator reveals that review papers have $\alpha = 21.40\%$, compared to $\alpha = 13.97\%$ for non-review papers. Similarly, the Pangram results show a similar trend: 20.00% review papers compare to 11.37% non-review papers. Figure 5 showcase the year-aggregated results, whereas Figure 2b shows the results split by year. However, we also note that while the predicted rates are higher for review papers, overall there are far more non-review papers, and so the total number of non-review papers that are predicted to be generated is much higher than for review papers (Figure 1).

Domain Variations. To understand whether LLM-generated content prevalence varies across scientific disciplines, we analyze papers from four major arXiv categories. Figure 5 presents the Alpha and Pangram estimates for the post-LLM period.

We observe notable domain differences. CS papers show the highest overall LLM-generated content, likely reflecting the field's early adoption of AI tools and researchers' familiarity with LLM technology. The review vs. regular paper gap persists across all disciplines[19] but varies in magnitude, with some fields showing more pronounced differences than others. This suggests that while the general trend holds across fields, the extent of AI adoption in writing depends on domain norms.

B Review Paper Classification Details

We report the results on the test data ($n = 140$) for the four models we consider in Table 3. Llama 3.3, Gemma 2, GPT-OSS, and GPT-4o-mini achieves F1 scores of 90.1, 92.0, 86.8, and 92.0 respectively. We select the model to use for all our experiments based on the validation set (§3.2) which was Gemma 2. While GPT-4o-mini performed on par (both on the validation set and the test set), Gemma 2 has the advantage that it is an open-weights model, and we are able to run it locally.

LLM Prompt The following prompt was used to classify papers as review/survey papers versus other types of papers:

[19] Aside from Statistics and Mathematics using Pangram, where we do not observe a statistically significant difference between the paper types.

You are an expert academic paper classifier. Your task is to classify papers into one of two categories:
1. Review/Survey: Papers that primarily review, survey, synthesize existing research, or present positions/perspectives on a field. These papers:

- Systematically review existing literature
- Provide comprehensive overviews of a research area
- Summarize and compare multiple existing approaches
- Present position papers or perspectives on research directions
- Often have titles containing words like "survey", "review", "overview", "primer", "position", "perspective"

2. Other: All other papers including:

- Original research papers presenting new methods, experiments, or results
- Technical reports
- Case studies
- Papers presenting new datasets or benchmarks
- Papers focused primarily on novel empirical results

You must respond with ONLY a JSON object in this exact format:
`{"classification": "review", "reasoning": "brief explanation"}`
OR
`{"classification": "other", "reasoning": "brief explanation"}`
The classification field must be either "review" or "other" (lowercase). Keep the reasoning brief (1-2 sentences).

For each paper, the classifier was provided with the paper's title and abstract in the user prompt. We did not optimize the prompt, as it worked well out of the box.

C Dataset Statistics by Year and Category

We provide statistics of the two datasets we collected. Tables 4 and 5 show the number of papers collected for the *arxiv-domains* and *cs-subcategories* datasets, respectively, broken down by year and subcategory. For each category, we sample up to 500 papers per month from 2020-2025.

D Estimated review vs. non-review papers generated by LLMs

Using our review vs. non-review classifier and the two LLM-paper detection methods we estimate the rates of LLM-generated papers in 2023-2025. By scaling the LLM-detection ratios with the total number of papers per year, we estimate the number of total LLM-generated papers per domain. We present the results for every domain in Tables 6-9

E Rogan-Gladen Adjustment

The Rogan-Gladen adjustment (Rogan and Gladen, 1978) corrects for biased prevalence estimates.

Given:

- $\hat{p}$ = observed (apparent) prevalence from the test
- Se = sensitivity (true positive rate) = $1 - \text{FNR}$
- Sp = specificity (true negative rate) = $1 - \text{FPR}$

Using simple arithmetic, it can be shown that the prevalence is:

$$\hat{p}_{\text{adj}} = \frac{\hat{p} + Sp - 1}{Se + Sp - 1} = \frac{\hat{p} - \text{FPR}}{1 - \text{FPR} - \text{FNR}} \quad (1)$$

In our application, we estimate the false positive rate (FPR) from pre-LLM era papers (2020–2022), when large language models like ChatGPT did not exist. We assume the false negative rate (FNR) is negligible or zero, simplifying Equation 1 to:

$$\hat{p}_{\text{adj}} = \frac{\hat{p} - \text{FPR}}{1 - \text{FPR}} \quad (2)$$

This adjusted estimate accounts for the baseline false positive rate of our detector, allowing us to better isolate the true prevalence of LLM-generated content in the post-LLM era.

F OpenAlex Data & Results

We use the OpenAlex API[20] to enrich the *cs-subcategories* dataset with metadata about the papers, their authors, and the institutional affiliations of the authors. OpenAlex is a nonprofit online library that contains "over 250M scholarly works from 250k sources, with extra coverage of humanities, non-English languages, and the Global South."[21] For each CS paper which also has Pangram results (~3.6k papers per CS subcategory), we

[20] https://docs.openalex.org/
[21] https://openalex.org/

Category	2020	2021	2022	2023	2024	2025	Total
Computer Science	6,000	6,000	6,000	6,000	6,000	6,000	36,000
Mathematics	6,000	6,000	6,000	6,000	6,000	6,000	36,000
Physics	6,000	6,000	6,000	6,000	6,000	5,963	35,963
Statistics	2,409	2,500	2,521	2,654	3,034	3,380	16,498
Total	20,409	20,500	20,521	20,654	21,034	21,343	124,461

Table 4: Number of papers in the *arxiv-domains* data, broken by category and year.

Subcategory	2020	2021	2022	2023	2024	2025	Total
Artificial Intelligence	727	928	822	1,309	2,029	3,758	9,573
Computation and Language	2,698	3,283	3,379	3,951	5,496	5,516	24,323
Cryptography and Security	1,103	1,198	1,252	1,449	1,752	2,226	8,980
Computer Vision and Pattern Recognition	5,414	5,689	5,778	5,954	6,000	6,000	34,835
Computers and Society	584	619	424	590	812	914	3,943
Human-Computer Interaction	578	742	772	1,162	1,911	2,237	7,402
Information Retrieval	337	494	539	743	944	1,098	4,155
Machine Learning	812	3,521	3,779	3,914	4,693	5,736	22,455
Robotics	1,175	1,633	1,915	2,648	3,534	3,762	14,667
Software Engineering	777	1,198	980	1,308	1,575	2,073	7,911
Total	14,205	19,305	19,640	23,028	28,746	33,320	138,244

Table 5: Number of papers in the *cs-subcategories* data, broken by category and year.

Subset / Year	Review Papers		Non-Review Papers	
	n.	%	n.	%
All				
2023	7,276	9.5%	69,182	90.5%
2024	9,558	9.2%	94,343	90.8%
2025	12,133	7.9%	141,775	92.1%
LLM (Alpha)				
2023	937	17.9%	4,301	82.1%
2024	2,298	12.5%	16,060	87.5%
2025	3,420	11.3%	26,736	88.7%
LLM (Pangram)				
2023	357	18.2%	1,607	81.8%
2024	868	12.8%	5,917	87.2%
2025	4,783	15.1%	26,801	84.9%

Table 6: Comparison of **Computer Science** paper estimates by estimation method and year on the *arxiv-domains* dataset. For each subset, we show the estimated review vs. regular papers for each year (2023-2025).

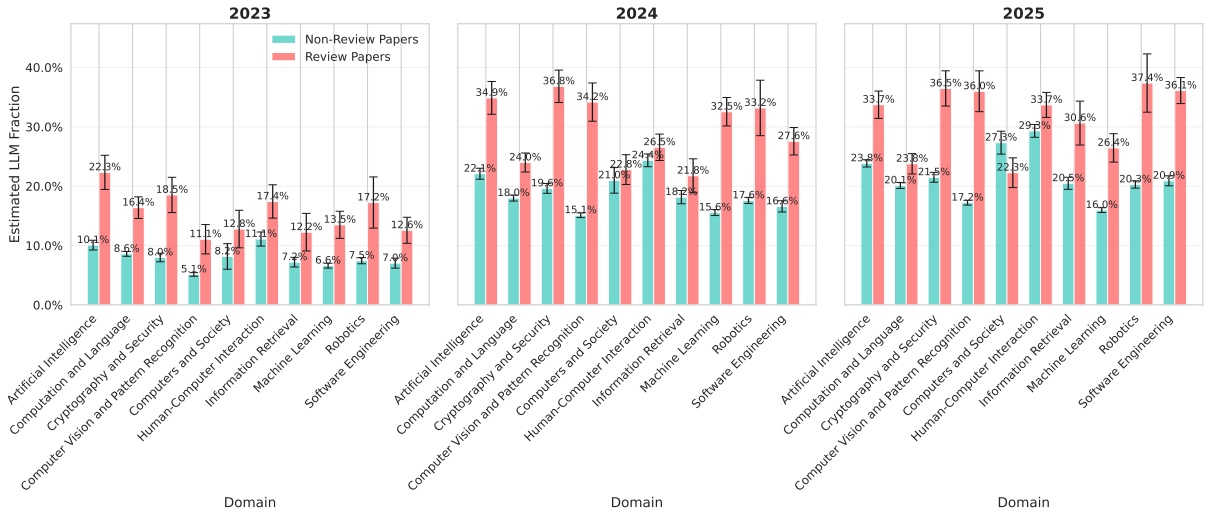
Subset / Year	Review Papers		Non-Review Papers	
	n.	%	n.	%
All				
2023	2,038	5.7%	33,830	94.3%
2024	2,524	5.9%	40,509	94.1%
2025	2,889	5.0%	55,292	95.0%
LLM (Alpha)				
2023	12	15.2%	67	84.8%
2024	40	16.5%	202	83.5%
2025	66	16.6%	331	83.4%
LLM (Pangram)				
2023	0	0.0%	91	100.0%
2024	35	10.6%	294	89.4%
2025	163	9.3%	1,586	90.7%

Table 7: Comparison of **Mathematics** paper estimates by estimation method and year on the *arxiv-domains* dataset. For each subset, we show the estimated review vs. regular papers for each year (2023-2025).

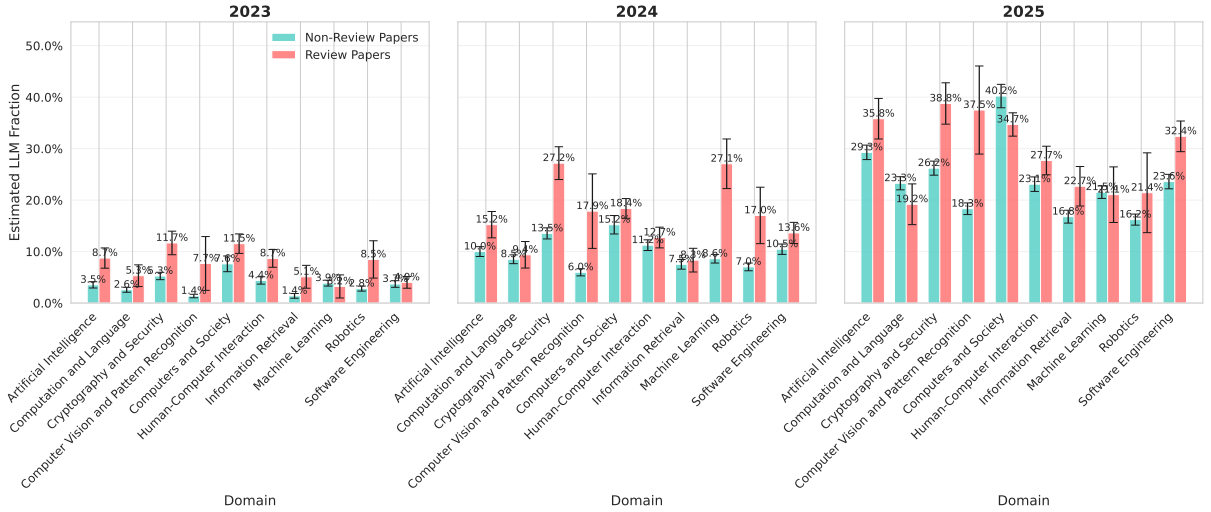
use the paper's DOI to collect the paper's topics, keywords, citation count, and authors, as well as each author's affiliation, h-index, number of works, and citation count, and each institution's country and h-index. We were able to match 74% of the papers in our sample to OpenAlex records, with increasing match rates over time (see Figure 7).

Figure 8 shows the predicted review rates by institution prestige and region. We observe that review papers are more likely to have been written by authors with higher prestige institutions ($h > 500$) and based in the Global North (determined by a manual list[22]). If an author has multiple affiliations listed, only the first one is considered. If a paper has multiple authors, each author's affiliation will

[22] https://meta.wikimedia.org/wiki/List_of_countries_by_regional_classification



(a) Alpha estimator method



(b) Pangram detection method

Figure 6: LLM-generated papers estimates by Computer Science subcategories in the *cs-subcategories* dataset, paper type, and year using (a) Alpha estimates and (b) Pangram detection ratios.

include this paper in its counts.

Figure 9 shows that junior authors are more likely to have written papers that are flagged as AI-generated. To bin the authors by seniority, we rely on their h-index from OpenAlex. We use h-index instead of number of works because the works count can be inflated by prolific co-authorship or certain publication patterns, while h-index is more robust to this inflation. Examining this figure, one might worry that the lower rates of generated papers for senior authors could be caused by senior authors writing more papers before these tools were available; however, we are not including all papers by these authors in this analysis, but rather only the papers in our sample, which are all post-2019. The AI detection score is a continuous value from 0 to 1 representing the probability that a paper was AI-generated, as returned by the Pangram API, while the AI flag rate is a binary yes/no flag derived from the score (> 0.5).

Finally, in Figure 10, we show the same topics and rates as in Table 2 but in a scatter plot that emphasizes outliers (like the high-review topics shown before) and shows that the rates of review papers do *not* rise proportionally to the rates of generated papers, despite their overall correlation.

G Additional Details

G.1 Data Annotation

Two authors of this paper performed the data annotation described in 3.2, by the paper's definition of review papers, and used binary labels (review vs. non-review) to annotate papers based on the title

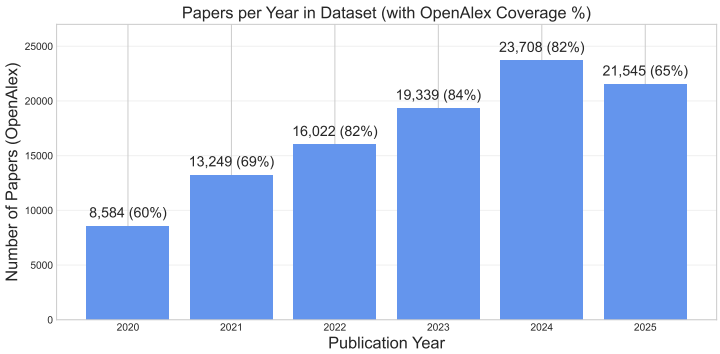


Figure 7: The number of papers per year that we were able to match to OpenAlex records. At the top of each column, we show the total number of papers successfully identified as well as the percent of the papers in the original dataset that were identified (the number of papers in the original dataset also increases over time).

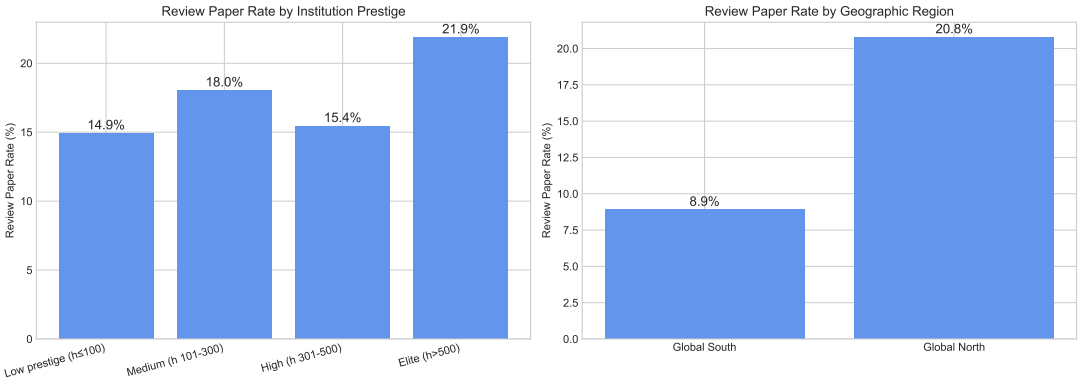


Figure 8: Types of institutions and their rates of papers that are predicted to be review papers.

and abstract alone.

G.2 AI Assistant Usage

We used Claude Code to assist in writing the code for this paper, which we manually validated and edited. We also used AI assistants to help with editing the paper.

Subset / Year	Review Papers		Non-Review Papers	
	n.	%	n.	%
All				
2023	1,124	5.8%	18,368	94.2%
2024	1,156	5.6%	19,314	94.4%
2025	1,425	5.3%	25,479	94.7%
LLM (Alpha)				
2023	60	9.1%	599	90.9%
2024	190	11.1%	1,516	88.9%
2025	277	8.9%	2,831	91.1%
LLM (Pangram)				
2023	15	8.4%	163	91.6%
2024	140	19.0%	597	81.0%
2025	240	8.7%	2,516	91.3%

Table 8: Comparison of **Physics** paper estimates by estimation method and year on the *arxiv-domains* dataset. For each subset, we show the estimated review vs. regular papers for each year (2023-2025).

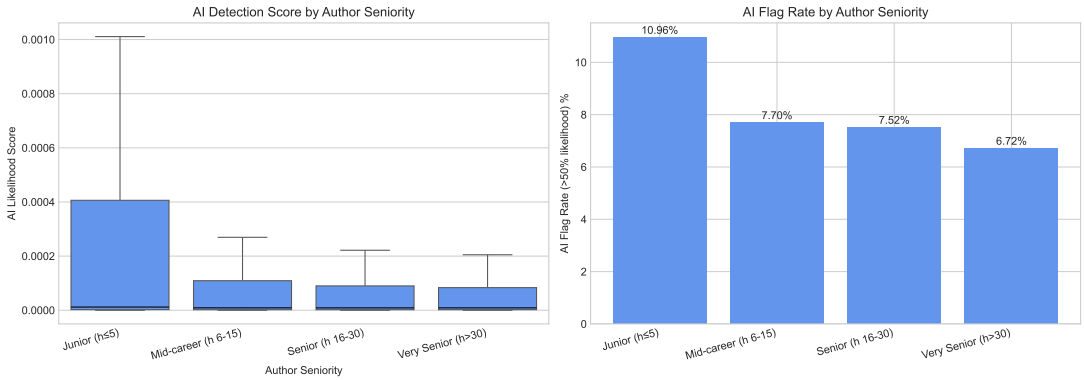


Figure 9: Author seniority (measured by h-index) and rates of predicting that those authors' papers are generated. Only papers included in our post-2019 sample are included.

Subset / Year	Review Papers		Non-Review Papers	
	n.	%	n.	%
All				
2023	401	12.5%	2,801	87.5%
2024	407	10.6%	3,419	89.4%
2025	568	9.6%	5,358	90.4%
LLM (Alpha)				
2023	17	16.3%	87	83.7%
2024	28	9.6%	263	90.4%
2025	84	14.4%	499	85.6%
LLM (Pangram)				
2023	2	4.5%	42	95.5%
2024	24	17.6%	112	82.4%
2025	57	8.3%	630	91.7%

Table 9: Comparison of **Statistics** paper estimates by estimation method and year on the *arxiv-domains* dataset. For each subset, we show the estimated review vs. regular papers for each year (2023-2025).

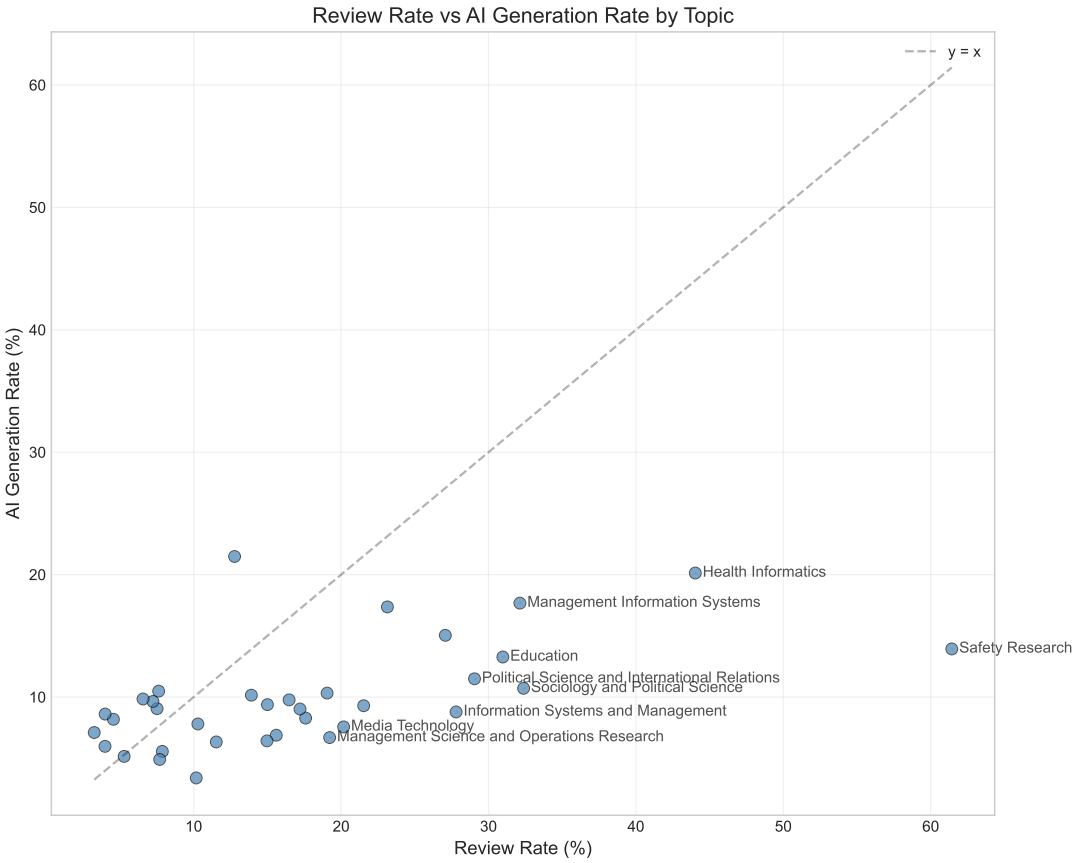


Figure 10: Paper topics from OpenAlex according to their predicted review and generation rates.

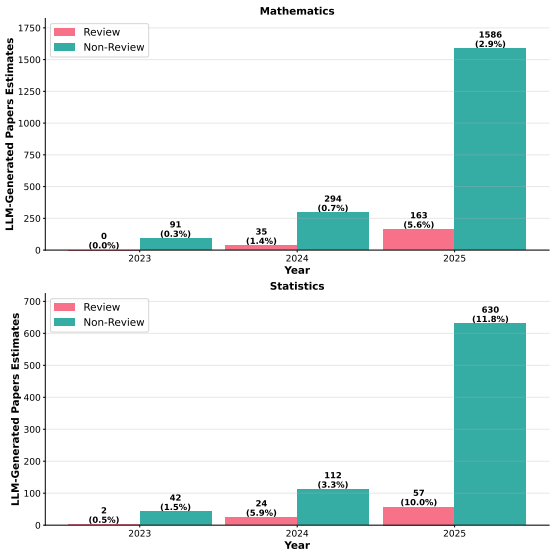


Figure 11: LLM-generated paper estimates using Pangram detection of Mathematics and Statistics for review vs. non-review papers on the *arxiv-domains* dataset. Notably, while review papers show higher percentages of LLM-generated content (in parentheses), the absolute number of non-review papers detected as LLM-generated is substantially larger in all domains, suggesting that restricting review papers alone may not address the core concern about LLM-generated content in academic publishing.