Mapping the Increasing Use of LLMs in Scientific Papers

Weixin Liang[*], **Yaohui Zhang**[*], **Zhengxuan Wu**[*], **Haley Lepp**,
Stanford University

Wenlong Ji, **Xuandong Zhao**,
Stanford University, UC Santa Barbara

Hancheng Cao, **Sheng Liu**, **Siyu He**, **Zhi Huang**, **Diyi Yang**,
Stanford University

Christopher Potts[†], **Christopher D Manning**[†], **James Y. Zou**[†]
Stanford University

Abstract

Scientific publishing lays the foundation of science by disseminating research findings, fostering collaboration, encouraging reproducibility, and ensuring that scientific knowledge is accessible, verifiable, and built upon over time. Recently, there has been immense speculation about how many people are using large language models (LLMs) like ChatGPT in their academic writing, and to what extent this tool might have an effect on global scientific practices. However, we lack a precise measure of the proportion of academic writing substantially modified or produced by LLMs. To address this gap, we conduct the first systematic, large-scale analysis across 950,965 papers published between January 2020 and February 2024 on the *arXiv*, *bioRxiv*, and *Nature* portfolio journals, using a population-level statistical framework to measure the prevalence of LLM-modified content over time. Our statistical estimation operates on the corpus level and is more robust than inference on individual instances. Our findings reveal a steady increase in LLM usage, with the largest and fastest growth observed in Computer Science papers (up to 17.5%). In comparison, Mathematics papers and the Nature portfolio showed the least LLM modification (up to 6.3%). Moreover, at an aggregate level, our analysis reveals that higher levels of LLM-modification are associated with papers whose first authors post preprints more frequently, papers in more crowded research areas, and papers of shorter lengths. Our findings suggests that LLMs are being broadly used in scientific writings.

1 Introduction

Since the release of ChatGPT in late 2022, anecdotal examples of both published papers (Okunytė, 2023; Deguerin, 2024) and peer reviews (Oransky & Marcus, 2024) which appear to be ChatGPT-generated have inspired humor and concern.[1] While certain tells, such as "regenerate response" (Conroy, 2023b;a) and "as an AI language model" (Vincent, 2023), found in published papers indicate modified content, less obvious cases are nearly impossible to detect at the individual level (Else, 2023; Gao et al., 2022). Liang et al. (2024a) present a

[*]Co-first authors, Correspondence to: Weixin Liang <wxliang@stanford.edu>

[†]Co-supervised project, Correspondence to: James Zou <jamesz@stanford.edu>

[1]Increased attention to ChatGPT-use by multilingual scholars has also brought to the fore important conversations about entrenched linguistic discrimination in academic publishing (Lepp & Sarin, 2024; Khanna et al., 2022).

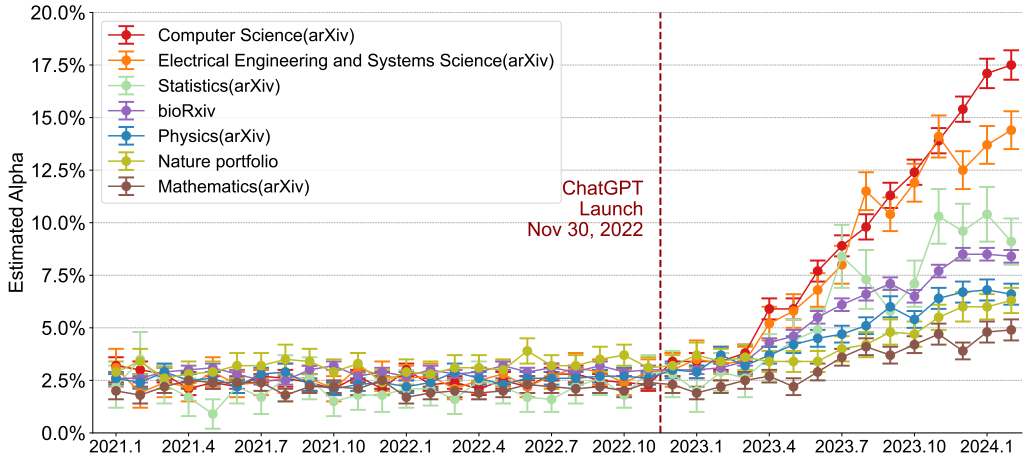

Figure 1: **Estimated Fraction of LLM-Modified Sentences across Academic Writing Venues over Time.** This figure displays the fraction (α) of sentences estimated to have been substantially modified by LLM in abstracts from various academic writing venues. The analysis includes five areas within *arXiv* (Computer Science, Electrical Engineering and Systems Science, Mathematics, Physics, Statistics), articles from *bioRxiv*, and a combined dataset from 15 journals within the *Nature* portfolio. 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. Error bars indicate 95% confidence intervals by bootstrap. Further analysis of paper introductions is presented in Supp Figure 6.

method for detecting the percentage of LLM-modified text in a corpus beyond such obvious cases. Applied to scientific publishing, the importance of this at-scale approach is two-fold: first, rather than looking at LLM-use as a type of rule-breaking on an individual level, we can begin to uncover structural circumstances which might motivate its use. Second, by examining LLM-use in academic publishing at-scale, we can capture epistemic and linguistic shifts, miniscule at the individual level, which become apparent with a birds-eye view.

Measuring the extent of LLM-use on scientific publishing has urgent applications. Concerns about accuracy, plagiarism, anonymity, and ownership have prompted some prominent scientific institutions to take a stance on the use of LLM-modified content in academic publications. The International Conference on Machine Learning (ICML) 2023, a major machine learning conference, has prohibited the inclusion of text generated by LLMs like ChatGPT in submitted papers, unless the generated text is used as part of the paper's experimental analysis (ICML, 2023). Similarly, the journal *Science* has announced an update to their editorial policies, specifying that text, figures, images, or graphics generated by ChatGPT or any other LLM tools cannot be used in published works (Thorp, 2023). Taking steps to measure the extent of LLM-use can offer a first-step in identifying risks to the scientific publishing ecosystem. Furthermore, exploring the circumstances in which LLM-use is high can offer publishers and academic institutions useful insight into author behavior. Sites of high LLM-use can act as indicators for structural challenges faced by scholars. These range from pressures to "publish or perish" which encourage rapid production of papers to concerns about linguistic discrimination that might lead authors to use LLMs as prose editors.

We conduct the first systematic, large-scale analysis to quantify the prevalence of LLM-modified content across multiple academic platforms, extending a recently proposed, state-of-the-art *distributional GPT quantification* framework (Liang et al., 2024a) for estimating the fraction of AI-modified content in a corpus. Throughout this paper, we use the term "LLM-modified" to refer to text content substantially updated by ChatGPT beyond basic

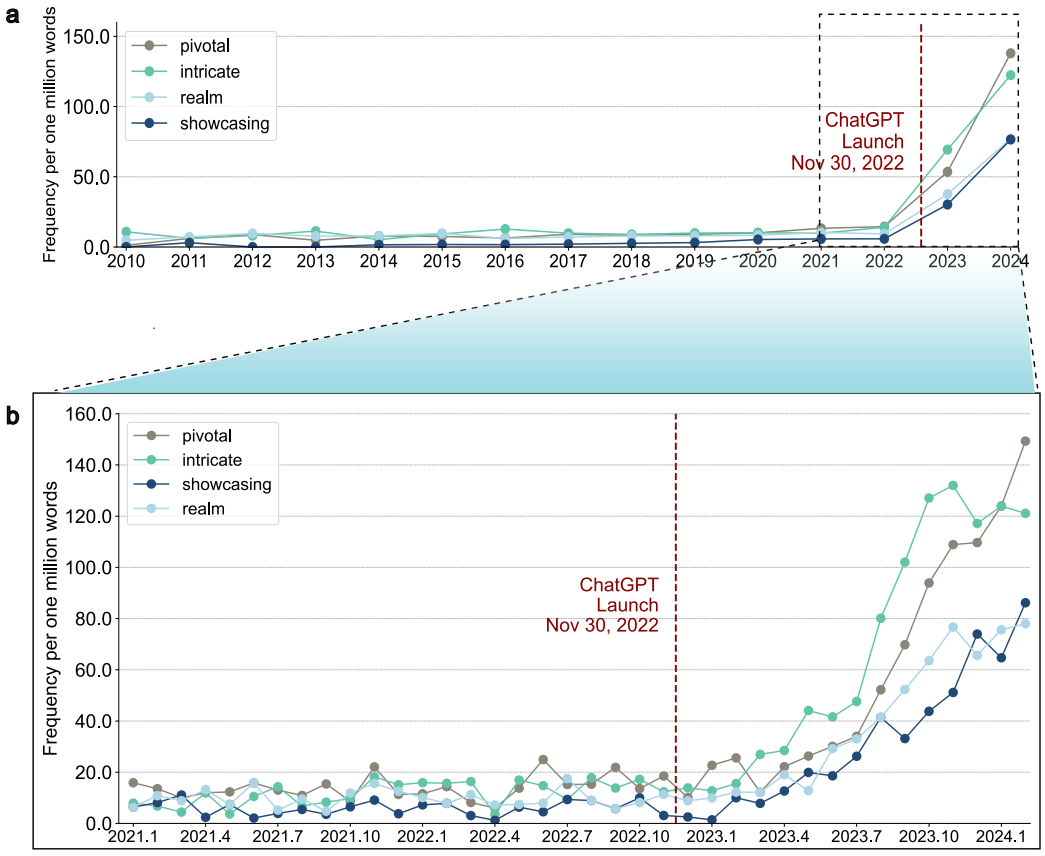

Figure 2: **Word Frequency Shift in arXiv Computer Science abstracts over 14 years (2010-2024).** The plot shows the frequency over time for the top 4 words most disproportionately used by LLM compared to humans, as measured by the log odds ratio. The words are: *realm*, *intricate*, *showcasing*, *pivotal*. 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.

spelling and grammatical edits. Modifications we capture in our analysis could include, for example, summaries of existing writing or the generation of prose based on outlines.

A key characteristic of this framework is that it operates on the population level, without the need to perform inference on any individual instance. 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 significant temporal distribution shifts and other realistic distribution shifts.

We apply this framework to the abstracts and introductions (Figure 1 and Supp Figure 6) of academic papers across multiple academic disciplines,including *arXiv*, *bioRxiv*, and 15 journals within the Nature portfolio, such as *Nature*, *Nature Biomedical Engineering*, *Nature Human Behaviour*, and *Nature Communications*. Our study analyzes a total of 950,965 papers published between January 2020 and February 2024, comprising 773,147 papers from *arXiv*, 161,280 from *bioRxiv*, and 16,538 from the Nature portfolio journals. The papers from *arXiv* cover multiple academic fields, 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 was observed in Computer Science papers, with α reaching 17.5% for abstracts and 15.3% for introductions by February 2024.

In contrast, Mathematics papers and the *Nature* portfolio showed the least increase, with α reaching 4.9% and 6.3% for abstracts and 3.5% and 6.4% for introductions, respectively. Moreover, our analysis reveals at an aggregate level that higher levels of LLM-modification are associated with papers whose first authors post preprints more frequently and papers with shorter lengths. Results also demonstrate a closer relationship between papers with LLM-modifications, which could indicate higher use in more crowded fields of study (as measured by the distance to the nearest neighboring paper in the embedding space), or that generated-text is flattening writing diversity.

2 Related Work

GPT Detectors Various methods have been proposed for detecting LLM-modified text, including zero-shot approaches that rely on statistical signatures characteristic of machine-generated content (Lavergne et al., 2008; Badaskar et al., 2008; Beresneva, 2016; Solaiman et al., 2019; Mitchell et al., 2023; Yang et al., 2023a; Bao et al., 2023; Tulchinskii et al., 2023) and training-based methods that finetune language models for binary classification of human vs. LLM-modified text (Bhagat & Hovy, 2013; Zellers et al., 2019; Bakhtin et al., 2019; Uchendu et al., 2020; Chen et al., 2023; Yu et al., 2023; Li et al., 2023; Liu et al., 2023; Bhattacharjee et al., 2023; Hu et al., 2023a). 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 (Wolff, 2020), and bias against non-dominant language varieties (Liang et al., 2023). The effectiveness and reliability of publicly available LLM-modified text detectors have also been questioned (OpenAI, 2019; Jawahar et al., 2020; Fagni et al., 2021; Ippolito et al., 2023; Mitchell et al., 2023; Gehrmann et al., 2019; Heikkilä, 2022; Crothers et al., 2022; Solaiman et al., 2019; Kirchner et al., 2023; Kelly, 2023), with the theoretical possibility of accurate instance-level detection being debated (Weber-Wulff et al., 2023; Sadasivan et al., 2023; Chakraborty et al., 2023). In this study, we apply the recently proposed *distributional GPT quantification* framework (Liang et al., 2024a), 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. A more comprehensive discussion of related work can be found in Appendix H.

3 Background: the *distributional LLM quantification* framework

We adapt the *distributional LLM quantification* framework from Liang et al. (2024a) to quantify the use of AI-modified academic writing. The framework consists of the following steps:

1. **Problem formulation**: Let $\mathcal{P}$ and $\mathcal{Q}$ be 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 α is the fraction of AI-modified documents. The goal is to estimate α based on observed documents $\{X_i\}_{i=1}^N \sim \mathcal{D}_\alpha$.
2. **Parameterization**: To make α 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:
$$\hat{p}_t = \frac{1}{n_P}\sum_{j=1}^{n_P}\mathbb{1}\{t\in X_j^P\}, \quad \hat{q}_t = \frac{1}{n_Q}\sum_{j=1}^{n_Q}\mathbb{1}\{t\in X_j^Q\}.$$

4. **Inference**: The fraction α is estimated by maximizing the log-likelihood of the observed documents under the mixture distribution $\hat{\mathcal{D}}_{\alpha,T}(X) = (1-\alpha)\hat{\mathcal{P}}_T(X) + \alpha\hat{\mathcal{Q}}_T(X)$:

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

Liang et al. (2024a) 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 for 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 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 Appendix 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 (Lee et al., 2024).

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 (Supp Figures 7 and 8). 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.

4 Implementation and Validations

4.1 Data Collection and Sampling

We collect data from three sources: *arXiv*, *bioRxiv*, and 15 journals from the *Nature* portfolio. For each source, we randomly sample up to 2,000 papers per month from January 2020 to February 2024. The procedure for generating the LLM-generated corpus data is described in Section 3. We focused on the introduction sections for the main texts, as the introduction was the most consistently and commonly occurring section across diverse categories of papers. See Appendix D for comprehensive implementation details.

4.2 Data Split, Model Fitting, and Evaluation

For model training, we count word frequencies for scientific papers written before the release of ChatGPT and the LLM-modified corpora described in Section 3. 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 January 1, 2022, to November 29, 2022, a time period prior to the release of ChatGPT, as the validation data. We construct validation sets with LLM-modified content proportions (α) ranging from 0% to 25%, in 5% increments, and compared the model's estimated α with the ground truth α. Additionally, 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 α values (Supp Figures 7 and 8).

5 Main Results and Findings

5.1 Temporal Trends in AI-Modified Academic Writing

Setup We apply the model to estimate the fraction of LLM-modified content (α) 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.

Results Our findings reveal a steady increase in the fraction of AI-modified content (α) in both the abstracts (Figure 1) and the introductions (Supp Figure 6), with the largest and fastest growth observed in Computer Science papers. By February 2024, the estimated α for Computer Science had increased to 17.5% for abstracts and 15.5% for introductions. The second-fastest growth was observed in Electrical Engineering and Systems Science, with the estimated α reaching 14.4% for abstracts and 12.4% for introductions during the same period. In contrast, Mathematics papers and the *Nature* portfolio showed the least increase. By the end of the studied period, the estimated α for Mathematics had increased to 4.9% for abstracts and 3.9% for introductions, while the estimated α for the Nature portfolio had reached 6.3% for abstracts and 4.3% for introductions. We analyzed Nature portfolio journal papers using both submission and publication dates. The results were consistent, with Nature portfolio papers having among the lowest estimated alphas, even when plotted by submission date.

The November 2022 estimates serve as a pre-ChatGPT reference point for comparison, as ChatGPT was launched on November 30, 2022. The estimated α for Computer Science in November 2022 was 2.3%, while for Electrical Engineering and Systems Science, Mathematics, and the Nature portfolio, the estimates were 2.9%, 2.4%, and 3.1%, respectively. These values are consistent with the false positive rate reported in the earlier section (Section 4.2).

5.2 Relationship Between First-Author Preprint Posting Frequency and GPT Usage in Computer Science

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 across multiple fields in Computer Science. 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 (Figure 3). We used the 2023 author grouping for the 2024.1-2 data, as we don't have the complete 2024 author data yet.

By February 2024, abstracts of papers whose first authors had ≥ 3 preprints in 2023 showed an estimated 19.3% of sentences modified by AI, compared to 15.6% for papers whose first authors had ≤ 2 preprints (Figure 3a). We observe a similar trend in the introduction sections, with first authors posting more preprints having an estimated 16.9% LLM-modified sentences, compared to 13.7% for first authors posting fewer preprints (Figure 3b). 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 sub-categories: cs.CV (Computer Vision and Pattern Recognition), cs.LG (Machine Learning), and cs.CL (Computation and Language) (Supp Figure 13).

Our results suggest that researchers in Computer Science 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 Computer Science research communities incentivizes taking steps to accelerate the writing process. We do not evaluate whether these preprints were accepted for publication.

5.3 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

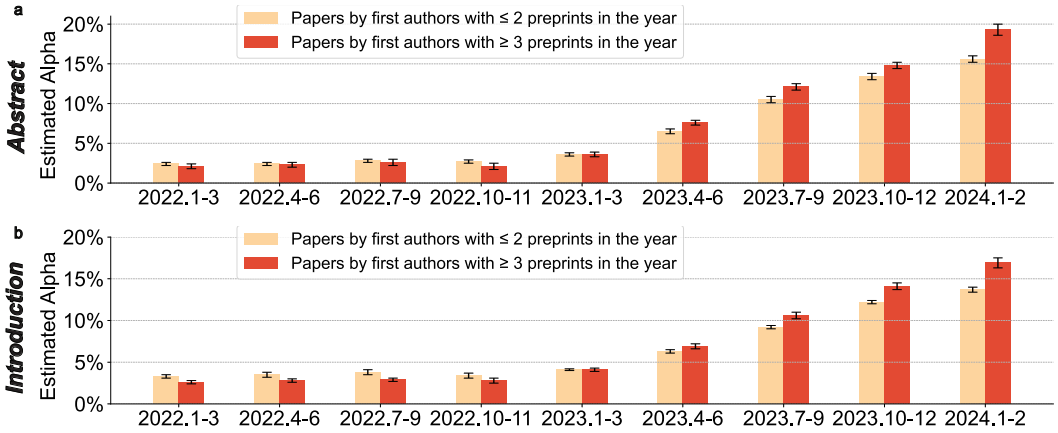

Figure 3: **Papers authored by first authors who post preprints more frequently tend to have a higher fraction of LLM-modified content in Computer Science.** Papers in *arXiv* Computer Science are stratified into two groups based on the preprint posting frequency of their first author, as measured by the number of first-authored preprints in the year. Error bars indicate 95% confidence intervals by bootstrap.

from the *arXiv* Computer Science papers using OpenAI's text-embedding-ada-002 model, creating a vector representation for each abstract. We then calculate the distance between each paper's vector and its nearest neighbor 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 Figure 4. After the release of ChatGPT, papers most similar to their closest peer consistently showed higher LLM usage compared to those least similar. By February 2024, the abstracts of papers more similar to their closest peer had an estimated 22.2% of sentences modified by LLMs, compared to 14.7% 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 neighbor distance within each of the three *arXiv* Computer Science sub-categories: 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 sub-category (Supp Figure 14).

There are several ways to interpret these findings. First, LLM-use in writing could cause the similarity in writing or content. Community pressures may even motivate scholars to try to sound more similar – to assimilate to the "style" of text generated by an LLM. Alternatively, LLMs may be more commonly used in research areas where papers tend to be more similar to each other. This could be due to the competitive nature of these crowded subfields, which may pressure researchers to write faster and produce similar findings. Future interdisciplinary research should explore these hypotheses.

5.4 Relationship Between Paper Length and AI Usage

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

Figure 5 shows the temporal trends of LLM usage for these two groups. After the release of ChatGPT, shorter papers consistently showed higher AI usage compared to longer papers. By February 2024, the abstracts of shorter papers had an estimated 17.7% of sentences modified by LLMs, compared to 13.6% for longer papers (Figure 5a). A similar trend was observed in the introduction sections (Figure 5b). To account for potential confounding effects of research fields, we conducted an additional robustness check. The finding holds

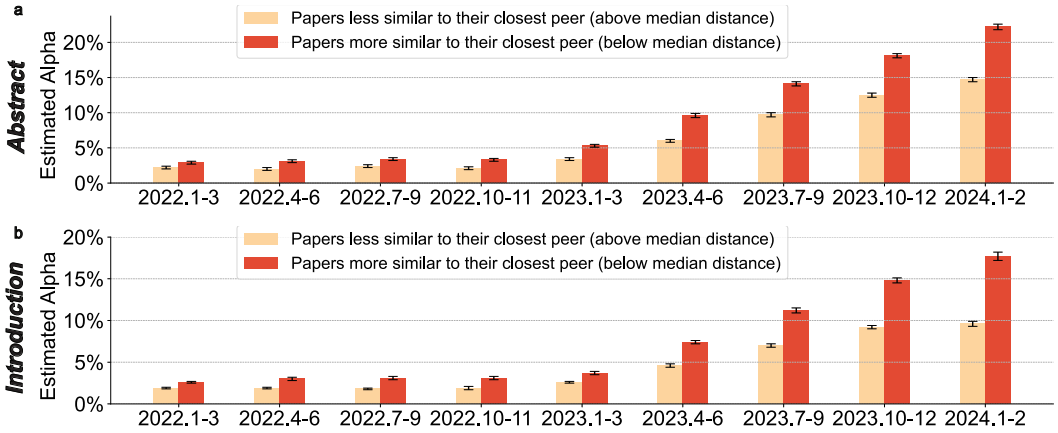

Figure 4: **Papers in more crowded research areas tend to have a higher fraction of LLM-modified content.** Papers in *arXiv* Computer Science are divided into two groups based on their abstract's embedding distance to their closest peer: papers more similar to their closest peer (below median distance) and papers less similar to their closest peer (above median distance). Error bars indicate 95% confidence intervals by bootstrap.

for both cs.CV (Computer Vision and Pattern Recognition) and cs.LG (Machine Learning) (Supp Figure 15). However, for cs.CL (Computation and Language), we found no significant difference in LLM usage between shorter and longer papers, possibly due to the limited sample size, as we only parsed a subset of the PDFs and calculated their full length.

As Computer Science conference papers typically have a fixed page limit, longer papers likely 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.

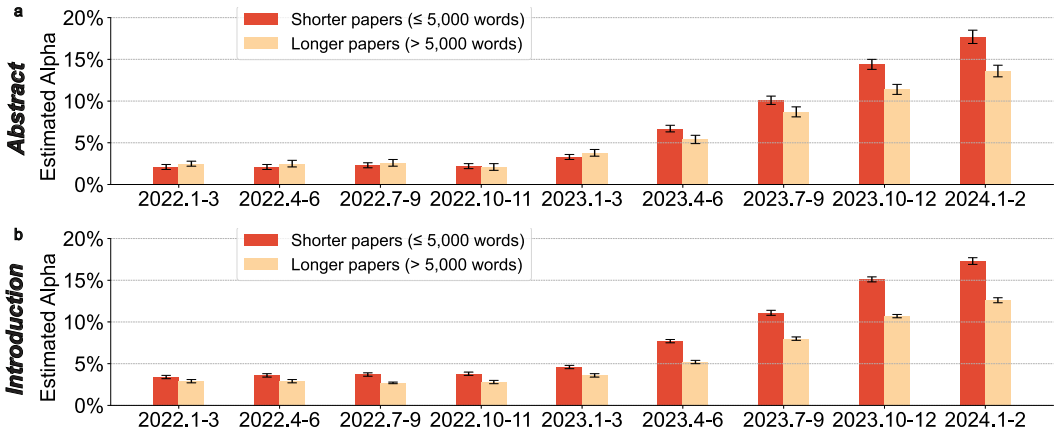

Figure 5: **Shorter papers tend to have a higher fraction of LLM-modified content.** *arXiv* Computer Science papers are stratified by their full text word count, including appendices, into two bins: below or above 5,000 words (the rounded median). Error bars indicate 95% confidence intervals by bootstrap.

6 Discussion

Our findings show a sharp increase in the estimated fraction of LLM-modified content in academic writing beginning about five months after the release of ChatGPT, with the fastest growth observed in Computer Science papers. This trend may be partially explained by Computer Science researchers' familiarity with and access to LLMs. Additionally, the fast-paced nature of LLM research and the associated pressure to publish quickly may incentivize the use of LLM writing assistance (Foster et al., 2015).

We expose several other factors associated with higher LLM usage in academic writing. First, authors who post preprints 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 to those in less crowded areas. Third, shorter papers consistently showed higher LLM-modification compared to longer papers, which may indicate that researchers working under time constraints are more likely to rely on AI for writing assistance. While these results suggest a potential link between competitive research environments and the pressure to publish quickly, further studies are needed to empirically validate this hypothesis.

In a research environment which values English-language publishing, scholars who are not heritage speakers of English may find it productive to have an AI model polish their writing (Lee et al., 2024). Additionally, LLMs offer the possibility of immediate feedback on initial drafts compared to traditional peer review processes, which can be time-consuming (Liang et al., 2024b). However, if the majority of modification comes from an LLM owned by a private company, there could be risks to the security and independence of scientific practice. We hope our results inspire further studies of widespread LLM-modified text and conversations about how to promote transparent, epistemically diverse, accurate, and independent scientific publishing.

Limitations While our study focused on ChatGPT, which accounts for more than three-quarters of worldwide internet traffic in the category (Van Rossum, 2024), we acknowledge that there are other large language models used for assisting academic writing. One potential confounder of our study is the increased prevalence of research on LLMs after the launch of ChatGPT. This shift in research focus could potentially affect the accuracy of our method in detecting LLM-modified content. However, our validation has shown that our framework is robust under temporal distribution shifts of research topics. Still, future studies could further validate and analyze the robustness of our method with more systematic control of the study content. Furthermore, while Liang et al. (2023) demonstrate that GPT-detection methods can falsely identify the writing of language learners as LLM-generated, our results showed that consistently low false positives estimates of α in 2022, which contains a significant fraction of texts written by multilingual scholars. We recognize that significant author population changes (MacroPolo, 2024) or other language-use shifts could still impact the accuracy of our estimates. 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. Investigation of the causal factors shaping this usage is an important direction for future research.

References

Scott Aaronson. Simons Institute Talk on Watermarking of Large Language Models, 2023. URL https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17.

Mikhail J. Atallah, Victor Raskin, Michael Crogan, Christian F. Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. In *Information Hiding*, 2001. URL https://doi.org/10.1007/3-540-45496-9_14.

Sameer Badaskar, Sachin Agarwal, and Shilpa Arora. Identifying real or fake articles: Towards better language modeling. In *International Joint Conference on Natural Language Processing (IJCNLP)*, 2008. URL https://aclanthology.org/I08-2115/.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. Real or fake? Learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*, 2019. URL https://arxiv.org/abs/1906.03351.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*, 2023. URL https://arxiv.org/abs/2310.05130.

Daria Beresneva. Computer-generated text detection using machine learning: A systematic review. In *International Conference on Applications of Natural Language to Data Bases*, 2016. URL https://doi.org/10.1007/978-3-319-41754-7_43.

Rahul Bhagat and Eduard H. Hovy. Squibs: What is a paraphrase? *Computational Linguistics*, 2013. URL https://aclanthology.org/J13-3001/.

Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. ConDA: Contrastive domain adaptation for AI-generated text detection. In *International Joint Conference on Natural Language Processing (IJCNLP)*, 2023. URL https://aclanthology.org/2023.ijcnlp-main.40/.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. On the possibilities of AI-generated text detection. *arXiv preprint arXiv:2304.04736*, 2023. URL https://arxiv.org/abs/2304.04736.

Yutian Chen, Hao Kang, Vivian Zhai, Liang Li, Rita Singh, and Bhiksha Ramakrishnan. GPT-Sentinel: Distinguishing human and ChatGPT generated content. *arXiv preprint arXiv:2305.07969*, 2023. URL https://arxiv.org/abs/2305.07969.

Yuei-Lin Chiang, Lu-Ping Chang, Wen-Tai Hsieh, and Wen-Chih Chen. Natural language watermarking using semantic substitution for chinese text. In *International Workshop on Digital Watermarking*, 2003. URL https://doi.org/10.1007/978-3-540-24624-4_10.

Gemma Conroy. How ChatGPT and other AI tools could disrupt scientific publishing. *Nature*, October 2023a. URL https://www.nature.com/articles/d41586-023-03144-w.

Gemma Conroy. Scientific sleuths spot dishonest ChatGPT use in papers. *Nature*, September 2023b. URL https://www.nature.com/articles/d41586-023-02477-w.

Evan Crothers, Nathalie Japkowicz, and Herna Viktor. Machine generated text: A comprehensive survey of threat models and detection methods. *arXiv preprint arXiv:2210.07321*, 2022. URL https://arxiv.org/abs/2210.07321.

Mack Deguerin. AI-generated nonsense is leaking into scientific journals. *Popular Science*, March 2024. URL https://www.popsci.com/technology/ai-generated-text-scientific-journals/.

Fabrizio Dell'Acqua, Edward McFowland, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Management*, 2023. URL https://www.hbs.edu/ris/Publication%20Files/24-013_d9b45b68-9e74-42d6-a1c6-c72fb70c7282.pdf.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's in my big data? In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2310.20707.

Holly Else. Abstracts written by ChatGPT fool scientists. *Nature*, Jan 2023. URL https://www.nature.com/articles/d41586-023-00056-7.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. TweepFake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415, 2021. URL https://arxiv.org/abs/2008.00036.

Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. Three bricks to consolidate watermarks for large language models. *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2023. URL https://arxiv.org/abs/2308.00113.

Jacob G Foster, Andrey Rzhetsky, and James A Evans. Tradition and innovation in scientists' research strategies. *American sociological review*, 80(5):875–908, 2015. URL https://arxiv.org/abs/1302.6906.

Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv 2022.12.23.521610*, 2022. URL https://doi.org/10.1101/2022.12.23.521610.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. URL https://arxiv.org/abs/2101.00027.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. GLTR: Statistical detection and visualization of generated text. In *Association for Computational Linguistics (ACL): System Demonstrations*, 2019. URL https://arxiv.org/abs/1906.04043.

Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and A. S. Bedi. Towards possibilities & impossibilities of AI-generated text detection: A survey. *arXiv preprint arXiv:2310.15264*, 2023. URL https://arxiv.org/abs/2310.15264.

Sourojit Ghosh and Aylin Caliskan. 'Person'== Light-skinned, western man, and sexualization of women of color: Stereotypes in stable diffusion. *arXiv preprint arXiv:2310.19981*, 2023. URL https://arxiv.org/abs/2310.19981.

Melissa Heikkilä. How to spot AI-generated text. *MIT Technology Review*, Dec 2022. URL https://www.technologyreview.com/2022/12/19/1065596/how-to-spot-ai-generated-text/.

Xiaobing Hu, Pin-Yu Chen, and Tsung-Yi Ho. RADAR: Robust AI-text detection via adversarial learning. *arXiv preprint arXiv:2307.03838*, 2023a. URL https://arxiv.org/abs/2307.03838.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. *arXiv preprint arXiv:2310.10669*, 2023b. URL https://arxiv.org/abs/2310.10669.

ICML. Clarification on large language model policy LLM. https://icml.cc/Conferences/2023/llm-policy, 2023.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*, 2023. URL https://arxiv.org/abs/1911.00650.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. Automatic detection of machine generated text: A critical survey. *arXiv preprint arXiv:2011.01314*, 2020. URL https://arxiv.org/abs/2011.01314.

Samantha Murphy Kelly. ChatGPT creator pulls AI detection tool due to 'low rate of accuracy'. *CNN Business*, Jul 2023. URL https://www.cnn.com/2023/07/25/tech/openai-ai-detection-tool/index.html.

Saurabh Khanna, Jon Ball, Juan Pablo Alperin, and John Willinsky. Recalibrating the scope of scholarly publishing: A modest step in a vast decolonization process. *Quantitative Science Studies*, 3(4):912–930, 12 2022. ISSN 2641-3337. doi: 10.1162/qss_a_00228. URL https://doi.org/10.1162/qss_a_00228.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning (ICML)*, 2023. URL https://arxiv.org/abs/2301.10226.

Jan Hendrik Kirchner, Lama Ahmad, Scott Aaronson, and Jan Leike. New AI classifier for indicating AI-written text, 2023. URL https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023. URL https://arxiv.org/abs/2307.15593.

Thomas Lavergne, Tanguy Urvoy, and François Yvon. Detecting fake content with relative entropy scoring. *Pan*, 2008. URL https://ceur-ws.org/Vol-377/paper4.pdf.

Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A Alghamdi, et al. A design space for intelligent and interactive writing assistants. In *The CHI Conference on Human Factors in Computing Systems (CHI)*, 2024. URL https://arxiv.org/abs/2403.14117.

Haley Lepp and Parth Sarin. A global AI community requires language-diverse publishing. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://globalaicultures.github.io/pdf/11_a_global_ai_community_requires.pdf.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. Deepfake text detection in the wild. *arXiv preprint arXiv:2305.13242*, 2023. URL https://arxiv.org/abs/2305.13242.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Y. Zou. GPT detectors are biased against non-native English writers. *arXiv preprint arXiv:2304.02819*, 2023. URL https://arxiv.org/abs/2304.02819.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 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 (ICML)*, 2024a. URL https://arxiv.org/abs/2403.07183.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 2024b. URL https://ai.nejm.org/doi/abs/10.1056/AIoa2400196.

Ryan Liu and Nihar B Shah. ReviewerGPT? An exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*, 2023. URL https://arxiv.org/abs/2306.00622.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Yu Lan, and Chao Shen. CoCo: Coherence-enhanced machine-generated text detection under data limitation with contrastive learning. *arXiv preprint arXiv:2212.10341*, 2023. URL https://arxiv.org/abs/2212.10341.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

MacroPolo. The global AI talent tracker, 2024. URL https://macropolo.org/digital-projects/the-global-ai-talent-tracker/.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning (ICML)*, 2023. URL https://arxiv.org/abs/2301.11305.

Paulina Okunytė. Google search exposes academics using ChatGPT in research papers. *Cybernews*, 2023. URL https://cybernews.com/news/academic-cheating-chatgpt-openai/.

OpenAI. GPT-2: 1.5b release. https://openai.com/research/gpt-2-1-5b-release, 2019. Accessed: 2019-11-05.

Ivan Oransky and Adam Marcus. Papers and peer reviews with evidence of ChatGPT writing. *Retraction Watch*, 2024. URL https://retractionwatch.com/papers-and-peer-reviews-with-evidence-of-chatgpt-writing/.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. In *The Journal of Machine Learning Research (JMLR)*, 2020. URL https://arxiv.org/abs/1910.10683.

Vinu Sankar Sadasivan, Aounon Kumar, S. Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023. URL https://arxiv.org/abs/2303.11156.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning (ICML)*, 2023. URL https://arxiv.org/abs/2303.17548.

Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. Red teaming language model detectors with language models. *arXiv preprint arXiv:2305.19713*, 2023. URL https://arxiv.org/abs/2305.19713.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023. URL https://arxiv.org/abs/2305.17493.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

H. Holden Thorp. ChatGPT is fun, but not an author. *Science*, 379(6630):313–313, 2023. doi: 10.1126/science.adg7879. URL https://www.science.org/doi/abs/10.1126/science.adg7879.

Mercan Topkara, Giuseppe Riccardi, Dilek Z. Hakkani-Tür, and Mikhail J. Atallah. Natural language watermarking: challenges in building a practical system. In *Electronic Imaging*, 2006a. URL https://disi.unitn.it/~riccardi/papers2/SPIE06.pdf.

Umut Topkara, Mercan Topkara, and Mikhail J. Atallah. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions. In *Workshop on Multimedia & Security*, 2006b. URL https://doi.org/10.1145/1161366.1161397.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, S. Barannikov, Irina Piontkovskaya, Sergey I. Nikolenko, and Evgeny Burnaev. Intrinsic dimension estimation for robust detection of AI-generated texts. *arXiv preprint arXiv:2306.04723*, 2023. URL https://arxiv.org/abs/2306.04723.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. Authorship attribution for neural text generation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020. URL https://aclanthology.org/2020.emnlp-main.673/.

Dann. Van Rossum. Generative AI top 150: The world's most used AI tools. https://www.flexos.work/learn/generative-ai-top-150, February 2024.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*, 2023. URL https://arxiv.org/abs/2306.07899.

James Vincent. 'As an AI language model': The phrase that shows how AI is pollulating the web. *The Verge*, Apr 2023. URL https://www.theverge.com/2023/4/25/23697218/ai-generated-spam-fake-user-reviews-as-an-ai-language-model.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 2023. URL https://doi.org/10.1007/s40979-023-00146-z.

Max Wolff. Attacking neural text detectors. *arXiv preprint arXiv:2002.11768*, abs/2002.11768, 2020. URL https://arxiv.org/abs/2002.11768.

Yihan Wu, Zhengmian Hu, Hongyang Zhang, and Heng Huang. DiPmark: A stealthy, efficient and resilient watermark for large language models. *arXiv preprint arXiv:2310.07710*, 2023. URL https://arxiv.org/abs/2310.07710.

Xianjun Yang, Wei Cheng, Linda Petzold, William Yang Wang, and Haifeng Chen. DNA-GPT: Divergent n-gram analysis for training-free detection of gpt-generated text. *arXiv preprint arXiv:2305.17359*, 2023a. URL https://arxiv.org/abs/2305.17359.

Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Ruth Petzold, William Yang Wang, and Wei Cheng. A survey on detection of LLMs-generated content. *arXiv preprint arXiv:2310.15654*, 2023b. URL https://arxiv.org/abs/2310.15654.

Kiyoon Yoo, Wonhyuk Ahn, Jiho Jang, and No Jun Kwak. Robust multi-bit natural language watermarking through invariant features. In *Association for Computational Linguistics (ACL)*, 2023. URL https://arxiv.org/abs/2305.01904.

Xiao Yu, Yuang Qi, Kejiang Chen, Guoqiang Chen, Xi Yang, Pengyuan Zhu, Weiming Zhang, and Neng H. Yu. GPT Paternity Test: GPT generated text detection with gpt genetic inheritance. *arXiv preprint arXiv:2305.12519*, 2023. URL https://arxiv.org/abs/2305.12519.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. URL https://arxiv.org/abs/1905.12616.

Yi-Fan Zhang, Zhang Zhang, Liang Wang, Tien-Ping Tan, and Rong Jin. Assaying on the robustness of zero-shot machine-generated text detectors. *arXiv preprint arXiv:2312.12918*, 2023. URL https://arxiv.org/abs/2312.12918.

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. Protecting language generation models via invisible watermarking. In *International Conference on Machine Learning (ICML)*, 2023. URL https://arxiv.org/abs/2302.03162.

Xuandong Zhao, Prabhanjan Vijendra Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. In *International Conference on Learning Representations (ICLR)*, 2024a. URL https://arxiv.org/abs/2306.17439.

Xuandong Zhao, Lei Li, and Yu-Xiang Wang. Permute-and-Flip: An optimally robust and watermarkable decoder for LLMs. *arXiv preprint arXiv:2402.05864*, 2024b. URL https://arxiv.org/abs/2024.05864.

A Estimated Fraction of LLM-Modified Sentences in *Introductions*

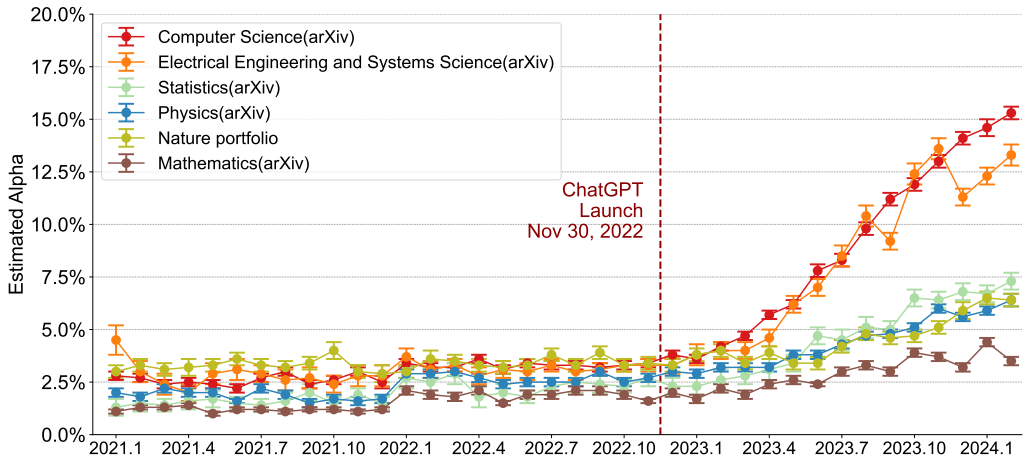

Figure 6: **Estimated Fraction of LLM-Modified Sentences in *Introductions* Across Academic Writing Venues Over Time.** We focused on the introduction sections for the main texts, as the introduction was the most consistently and commonly occurring section across different categories of papers. This figure presents the estimated fraction (α) of sentences in introductions which are LLM-modified, across the same venues as Figure 1. We found that the results are consistent with those observed in abstracts (Figure 1). We did not include *bioRxiv* introductions as there is no bulk download of PDFs available. Error bars indicate 95% confidence intervals by bootstrap.

B Fine-grained Validation of Model Performance

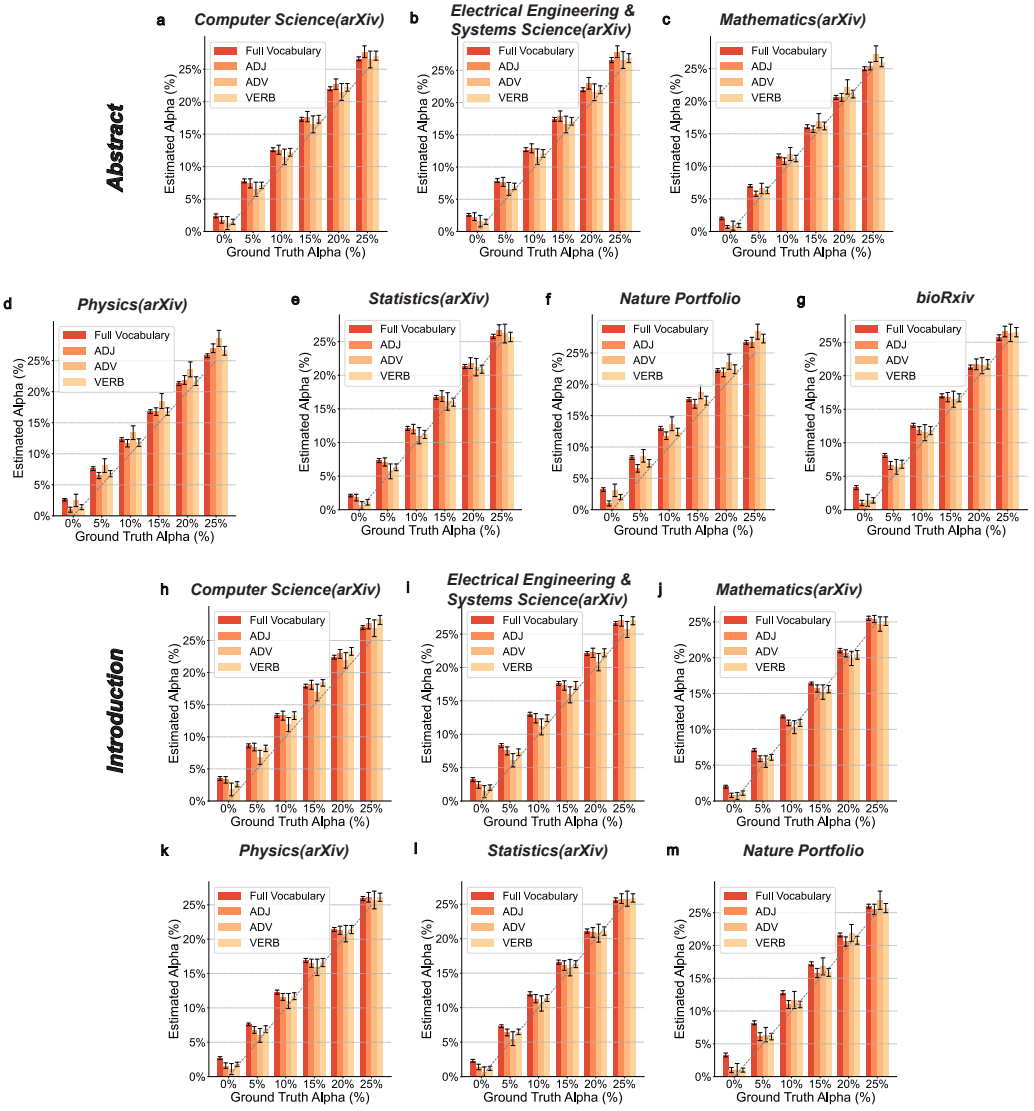

Figure 7: **Fine-grained Validation of Model Performance Under Temporal Distribution Shift.** We evaluate the accuracy of our models in estimating the fraction of LLM-modified content (α) under a challenging temporal data split, where the validation data (sampled from 2022-01-01 to 2022-11-29) are temporally separated from the training data (collected up to 2020-12-31) by at least a year. The X-axis indicates the ground truth α, while the Y-axis indicates the model's estimated α. In all cases, the estimation error for α is less than 3.5%. The first 7 panels (a–g) are the validation on abstracts for each academic writing venue, while the later 6 panels (h–m) are the validation on introductions. We did not include *bioRxiv* introductions due to the unavailability of bulk PDF downloads. Error bars indicate 95% confidence intervals by bootstrap.

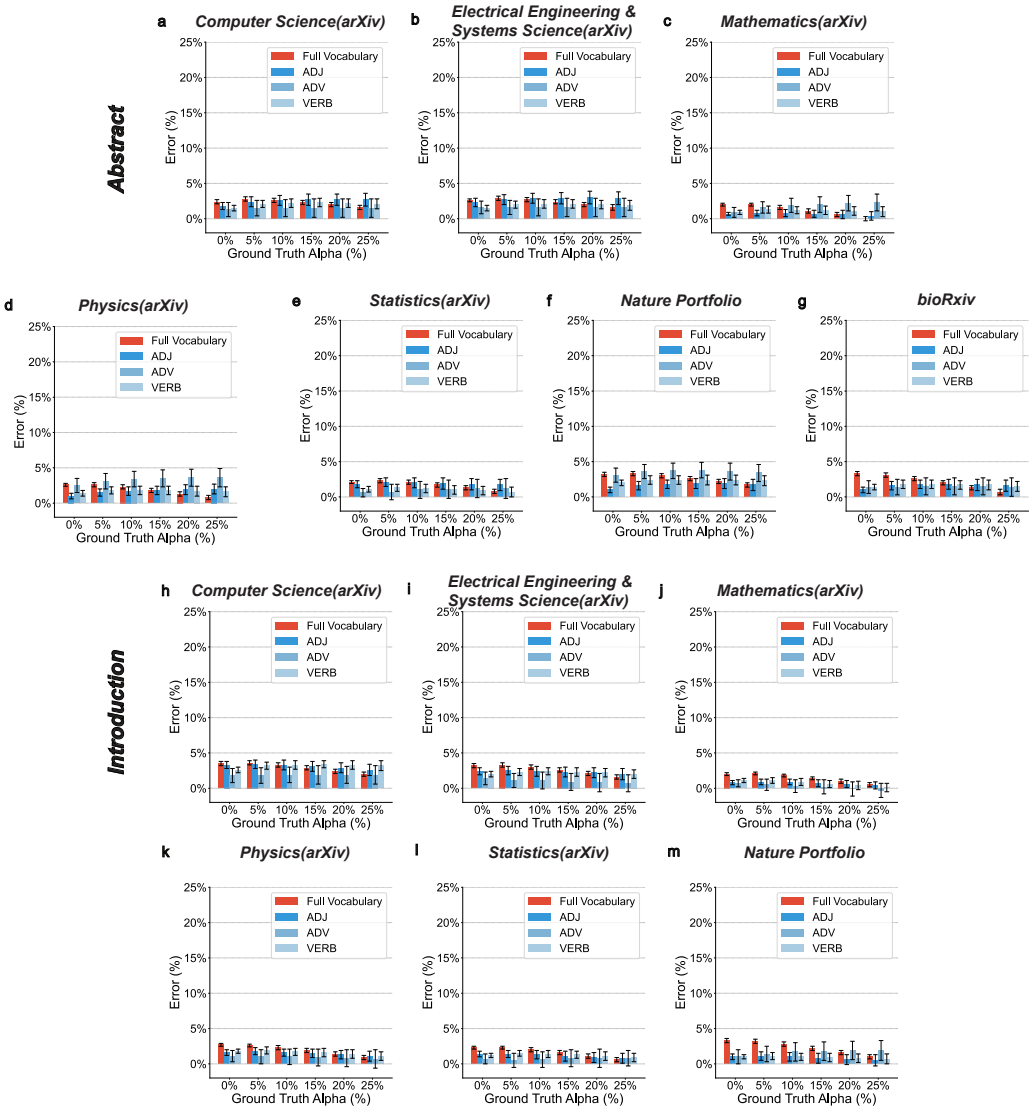

Figure 8: **Fine-grained Validation of Model Performance Under Temporal Distribution Shift.** We evaluate the accuracy of our models in estimating the fraction of LLM-modified content (α) under a challenging temporal data split, where the validation data (sampled from 2022-01-01 to 2022-11-29) are temporally separated from the training data (collected up to 2020-12-31) by at least a year. The X-axis indicates the ground truth α, while the Y-axis indicates represents the error (difference between model's estimated α and ground truth α). In all cases, the estimation error for α is less than 3.5%. The first 7 panels (a–g) are the validation on abstracts for each academic writing venue, while the later 6 panels (h–m) are the validation on introductions. We did not include *bioRxiv* introductions due to the unavailability of bulk PDF downloads. Error bars indicate 95% confidence intervals by bootstrap.

C LLM prompts used in the study

```
The aim here is to reverse-engineer the author's writing process by
    taking a piece of text from a paper and compressing it into a more
    concise form. This process simulates how an author might distill
    their thoughts and key points into a structured, yet not overly
    condensed form.

Now as a first step, first summarize the goal of the text, e.g., is it
    introduction, or method, results? and then given a complete piece of
    text from a paper, reverse-engineer it into a list of bullet points.
```

Figure 9: Example prompt for summarizing a paragraph from a human-authored paper into a skeleton: This process simulates how an author might first only write the main ideas and core information into a concise outline. The goal is to capture the essence of the paragraph in a structured and succinct manner, serving as a foundation for the previous prompt.

```
Following the initial step of reverse-engineering the author's writing
    process by compressing a text segment from a paper, you now enter the
     second phase. Here, your objective is to expand upon the concise
    version previously crafted. This stage simulates how an author
    elaborates on the distilled thoughts and key points, enriching them
    into a detailed, structured narrative.

Given the concise output from the previous step, your task is to develop
    it into a fully fleshed-out text.
```

Figure 10: Example prompt for expanding the skeleton into a full text: The aim here is to simulate the process of using the structured outline as a basis to generate comprehensive and coherent text. This step mirrors the way an author might flesh out the outline into detailed paragraphs, effectively transforming the condensed ideas into a fully articulated section of a paper. The format and depth of the expansion can vary, reflecting the diverse styles and requirements of different academic publications.

```
Your task is to proofread the provided sentence for grammatical accuracy.
     Ensure that the corrections introduce minimal distortion to the
    original content.
```

Figure 11: Example prompt for proofreading.

D Additional Information on Implementation and Validations

Supplementary Information about Data We collected data for this study from three publicly accessible sources: official 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, Statistics), we randomly sampled 2,000 papers per month from January 2020 to February 2024. Similarly, from *bioRxiv*, we randomly sampled 2,000 papers for each month within the same timeframe. 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 February 2024. The procedure for generating the AI corpus data for a given time period is described in aforementioned Section 3.

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 & Evolution, Nature Structural & Molecular Biology, Nature Cell Biology, Nature Human Behaviour, Nature Immunology, Nature Microbiology, Nature Biomedical Engineering, Communications Earth & Environment, Communications Biology, Communications Physics, Communications Chemistry, Communications Materials,* and *Communications Medicine*.

Additional Information on Large Language Models In this study, we utilized the gpt-3.5-turbo-0125 model, which was trained on data up to September 2021, to generate the training data for our analysis. The LLM was employed solely for the purpose of creating the training dataset and was not used in any other aspect of the study.

We chose to focus on ChatGPT due to its dominant position in the generative AI market. According to a comprehensive analysis conducted by FlexOS in early 2024, ChatGPT accounts for an overwhelming 76% of global internet traffic in the category, followed by Bing AI at 16%, Bard at 7%, and Claude at 1% (Van Rossum, 2024). This market share underscores ChatGPT's widespread adoption and makes it a highly relevant subject for our investigation. Furthermore, recent studies have also shown that ChatGPT demonstrates substantially better understanding of scientific papers than other LLMs (Liang et al., 2024b; Liu & Shah, 2023).

We chose to use GPT-3.5 for generating the training data due to its free availability, which lowers the barrier to entry for users and thereby captures a wider range of potential LLM usage patterns. This accessibility makes our study more representative of the broad phenomenon of LLM-assisted writing. Furthermore, the previous work by Liang et al. (2024a) has demonstrated the framework's robustness and generalizability to other LLMs. Their findings suggest that the framework can effectively handle significant content shifts and temporal distribution shifts.

Regarding the parameter settings for the LLM, we set the decoding temperature to 1.0 and the maximum decoding length to 2048 tokens during our experiments. The Top P hyperparameter, which controls the cumulative probability threshold for token selection, was set to 1.0. Both the frequency penalty and presence penalty, which can be used to discourage the repetition of previously generated tokens, were set to 0.0. Additionally, we did not configure any specific stop sequences during the decoding process.

E Word Frequency Shift in arXiv Computer Science introductions

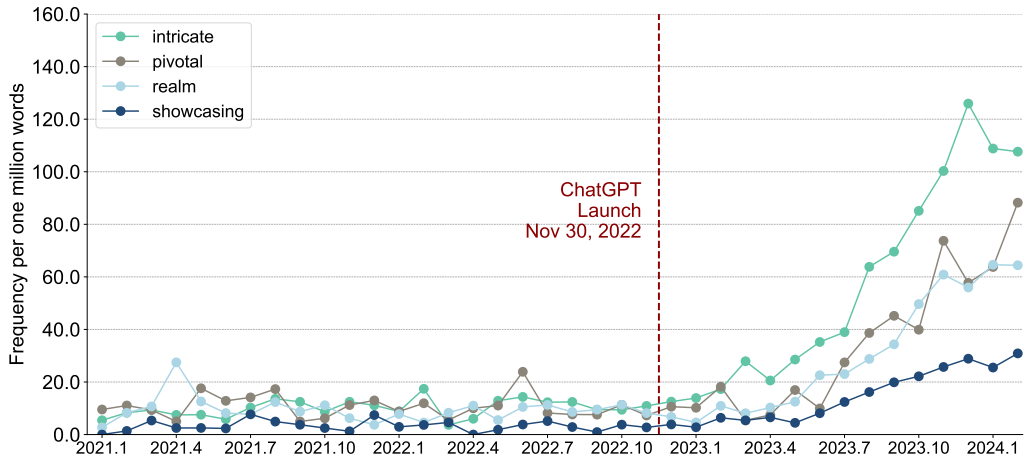

Figure 12: **Word Frequency Shift in sampled *arXiv* Computer Science introductions in the past two years.** The plot shows the frequency over time for the same 4 words as demonstrated in Figure 2. The words are: *realm*, *intricate*, *showcasing*, *pivotal*. The trend is similar for two figures. Data from 2010-2020 is not included in this analysis due to the computational complexity of parsing the full text from a large number of arXiv papers.

F Fine-grained Main Findings

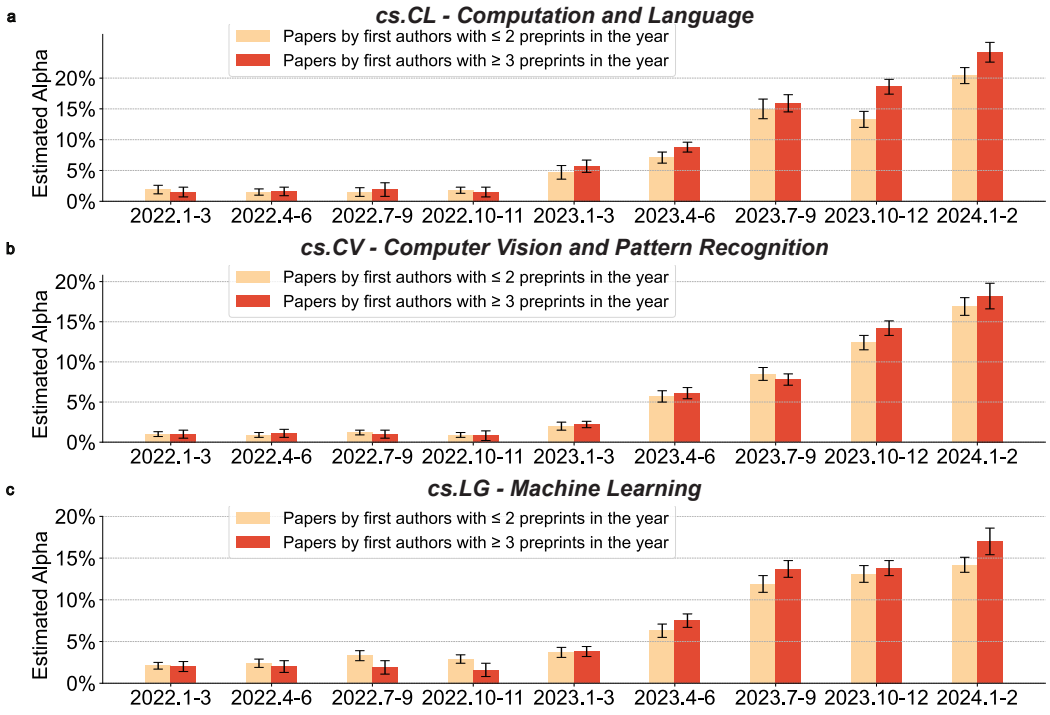

Figure 13: **The relationship between first-author preprint posting frequency and LLM usage holds across *arXiv* Computer Science sub-categories.** Papers in each *arXiv* Computer Science sub-category (cs.CV, cs.LG, and cs.CL) are stratified into two groups based on the preprint posting frequency of their first author, as measured by the number of first-authored preprints in the year: those with ≤ 2 preprints and those with ≥ 3 preprints. Error bars indicate 95% confidence intervals by bootstrap.

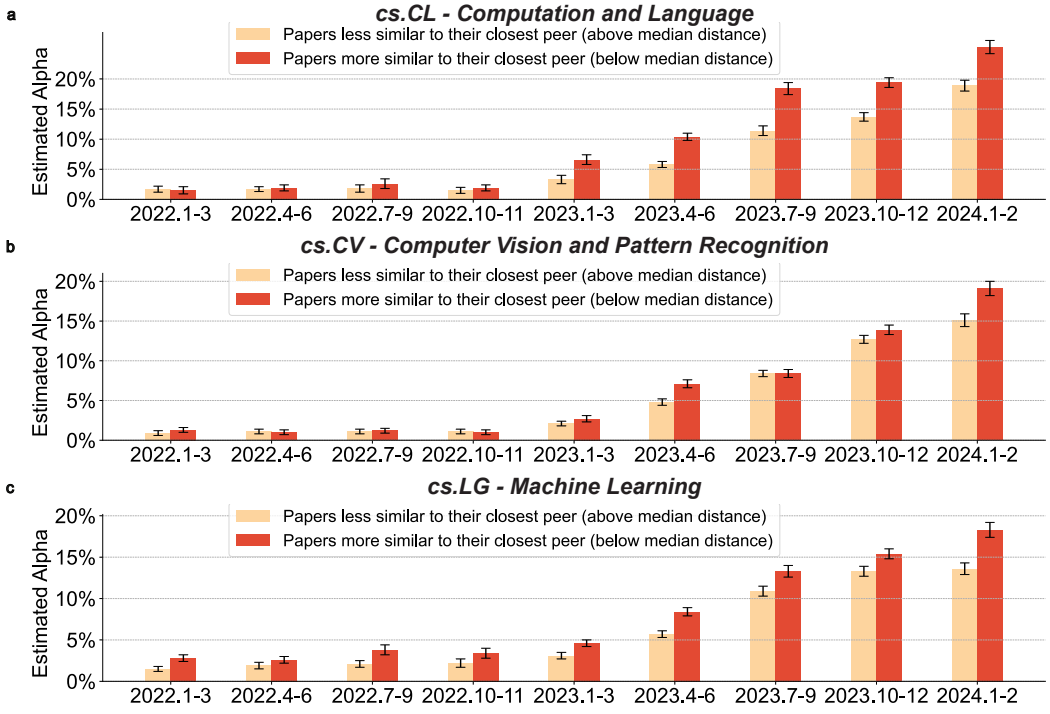

Figure 14: **The relationship between paper similarity and LLM usage holds across *arXiv* Computer Science sub-categories.** Papers in each *arXiv* Computer Science sub-category (cs.CV, cs.LG, and cs.CL) are divided into two groups based on their abstract's embedding distance to their closest peer within the respective sub-category: papers more similar to their closest peer (below median distance) and papers less similar to their closest peer (above median distance). Error bars indicate 95% confidence intervals by bootstrap.

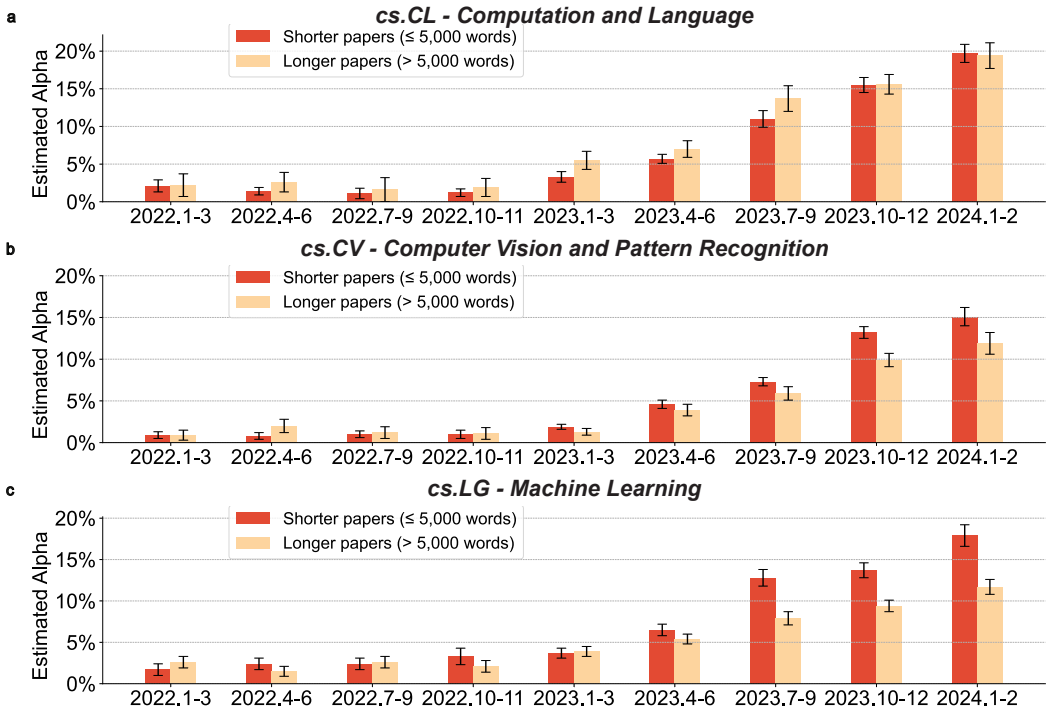

Figure 15: **The relationship between paper length and LLM usage holds for cs.CV and cs.LG, but not for cs.CL.** Papers in each *arXiv* Computer Science sub-category (cs.CV, cs.LG, and cs.CL) are stratified by their full text word count, including appendices, into two bins: below or above 5,000 words (the rounded median). For cs.CL, no significant difference in LLM usage was found between shorter and longer papers, possibly due to the limited sample size, as only a subset of the PDFs were parsed to calculate the full length. Error bars indicate 95% confidence intervals by bootstrap.

G Proofreading Results on arXiv data

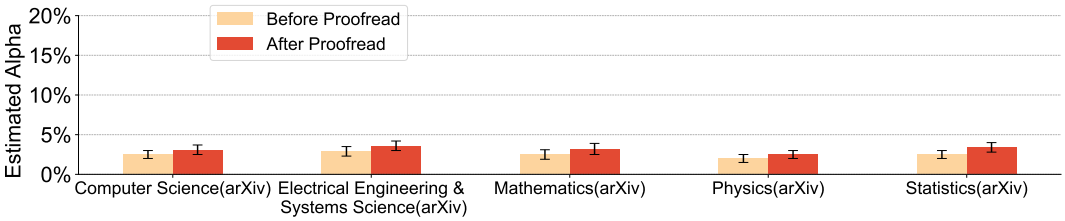

Figure 16: **Robustness of estimations to proofreading.** The plot demonstrates a slight increase in the fraction of LLM-modified content after using Large Language Models (LLMs) for "proofreading" across different *arXiv* main categories. This observation validates our method's robustness to minor LLM-generated text edits, such as those introduced by simple proofreading. The analysis was conducted on 1,000 abstracts from each *arXiv* main category, randomly sampled from the period between January 1, 2022, and November 29, 2022. Error bars indicate 95% confidence intervals by bootstrap.

H Extended Related Work

Zero-shot LLM detection. A major category of LLM text detection uses statistical signatures that are characteristic of machine-generated text, and the scope is to detect the text within individual documents. Initially, techniques to distinguish AI-modified text from human-written text employed various metrics, such as entropy (Lavergne et al., 2008), the frequency of rare n-grams (Badaskar et al., 2008), perplexity (Beresneva, 2016), and log-probability scores (Solaiman et al., 2019), which are derived from language models. More recently, DetectGPT (Mitchell et al., 2023) found that AI-modified text is likely to be found in areas with negative log probability curvature. DNA-GPT (Yang et al., 2023a) improves performance by examining the divergence in n-gram patterns. Fast-DetectGPT (Bao et al., 2023) enhances efficiency by utilizing conditional probability curvature over raw probability. Tulchinskii et al. (2023) studied the intrinsic dimensionality of generated text to perform the detection. We refer to recent surveys by Yang et al. (2023b); Ghosal et al. (2023) for additional details and more related works. However, zero-shot detection requires direct access to LLM internals to enable effective detection. Closed-source commercial LLMs, like GPT-4, necessitate using proxy LLMs, which compromises the robustness of zero-shot detection methods across various scenarios (Sadasivan et al., 2023; Shi et al., 2023; Yang et al., 2023b; Zhang et al., 2023).

Training-based LLM detection. Another category is training-based detection, which involves training classification models on datasets that consist of both human and AI-modified texts for the binary classification task of detection. Early efforts applied classification algorithms to identify AI text across various domains, such as peer review submissions (Bhagat & Hovy, 2013), media publications (Zellers et al., 2019), and other contexts (Bakhtin et al., 2019; Uchendu et al., 2020). Recently, researchers have finetuned pretrained language model backbones for this binary classification. GPT-Sentinel (Chen et al., 2023) uses the constructed dataset OpenGPTText to train RoBERTa (Liu et al., 2019) and T5 (Raffel et al., 2020) classifiers. GPT-Pat (Yu et al., 2023) trains a Siamese neural network to compute the semantic similarity of AI text and human text. Li et al. (2023) build a wild testbed by gathering texts from various human writings and texts generated by different LLMs. Using techniques such as contrastive and adversarial learning can enhance classifier robustness (Liu et al., 2023; Bhattacharjee et al., 2023; Hu et al., 2023a). We refer to recent surveys Yang et al. (2023b); Ghosal et al. (2023) for additional methods and details. However, these publicly available tools for detecting AI-modified content have sparked a debate about their effectiveness and reliability (OpenAI, 2019; Jawahar et al., 2020; Fagni et al., 2021; Ippolito et al., 2023; Mitchell et al., 2023; Gehrmann et al., 2019; Heikkilä, 2022; Crothers et al., 2022; Solaiman et al., 2019). OpenAI's decision to discontinue its AI-modified text classifier in 2023 due to "low rate of accuracy" further highlighted this discussion (Kirchner et al., 2023; Kelly, 2023).

Training-based detection methods face challenges such as overfitting to training data and language models, making them vulnerable to adversarial attacks (Wolff, 2020) and biased against non-dominant language varieties (Liang et al., 2023). The theoretical possibility of achieving accurate *instance*-level detection has also been questioned (Weber-Wulff et al., 2023; Sadasivan et al., 2023; Chakraborty et al., 2023).

LLM watermarking. Text watermarking introduces a method to detect AI-modified text by embedding an imperceptible signal, known as a watermark, directly into the text. This watermark can be retrieved by a detector that shares the model owner's secret key. Early watermarking techniques included synonym substitution (Chiang et al., 2003; Topkara et al., 2006b) and syntactic restructuring (Atallah et al., 2001; Topkara et al., 2006a). Modern watermarking strategies involve integrating watermarks into the decoding process of language models (Aaronson, 2023; Kirchenbauer et al., 2023; Zhao et al., 2023). Researchers have developed various techniques, such as the Gumbel watermark (Aaronson, 2023), which uses traceable pseudo-random softmax sampling, and the red-green list approach (Kirchenbauer et al., 2023; Zhao et al., 2024a), which splits the vocabulary based on hash values of previous n-grams. Some methods focus on preserving the original token probability distributions (Hu et al., 2023b; Kuditipudi et al., 2023; Wu et al., 2023), while others aim to improve detectability and perplexity (Zhao et al., 2024b) or incorporate multi-bit watermarks (Yoo

et al., 2023; Fernandez et al., 2023). However, one major concern with watermarking is that it requires the involvement of the model or service owner, such as OpenAI, to implant the watermark during the text generation process. In contrast, the framework by Liang et al. (2024a) operates independently of the model or service owner's intervention, allowing for the monitoring of AI-modified content without requiring their active participation or adoption.

Implications for LLM Pretraining Data Quality The increasing prevalence of AI-modified content in academic papers, particularly on platforms like *arXiv*, has important implications for the quality of LLM pretraining data. *arXiv* has become a significant source of training data for LLMs, contributing approximately 2.5% of the data for models like Llama (Touvron et al., 2023), 12% for RedPajama (Elazar et al., 2023), and 8.96% for the Pile (Gao et al., 2020). Our findings suggest that a growing proportion of this pretraining data may contain LLM-modified content. Preliminary research indicates that the inclusion of LLM-modified content (Veselovsky et al., 2023) in LLM training can lead to several pitfalls, such as the reinforcement of stereotypes and biases against anyone who is not a middle-aged "European/North American man" (Ghosh & Caliskan, 2023; Santurkar et al., 2023), the flattening of variation in language and content (Dell'Acqua et al., 2023), and the potential failure of models to accurately capture the true distribution of the original content, which may result in model collapse (Shumailov et al., 2023). Santurkar et al. (2023) demonstrate that this phenomenon amplifies the effect of LLMs providing content that is unrepresentative of most of the world. As such, our results underscore the importance of robust data curation and filtering strategies even in seemingly unpolluted datasets.