

# Human-LLM Coevolution: Evidence from Academic Writing

Mingmeng Geng<sup>1</sup> Roberto Trotta<sup>1,2</sup>

<sup>1</sup>International School for Advanced Studies (SISSA) <sup>2</sup>Imperial College London  
mgeng@sissa.it

## Abstract

With a statistical analysis of arXiv paper abstracts, we report a marked drop in the frequency of several words previously identified as overused by ChatGPT, such as “*delve*”, starting soon after they were pointed out in early 2024. The frequency of certain other words favored by ChatGPT, such as “*significant*”, has instead kept increasing. These phenomena suggest that some authors of academic papers have adapted their use of large language models (LLMs), for example, by selecting outputs or applying modifications to the LLM-generated content. Such coevolution and cooperation of humans and LLMs thus introduce additional challenges to the detection of machine-generated text in real-world scenarios. Estimating the impact of LLMs on academic writing by examining word frequency remains feasible, and more attention should be paid to words that were already frequently employed, including those that have decreased in frequency.

## 1 Introduction

After the launch of ChatGPT at the end of 2022, large language models (LLMs) began to be widely used and are now transforming many aspects of our work and life, with academic writing being one of them. The coevolution of AI and humans has also been recognized by researchers (Pedreschi et al., 2024).

For example, empirical studies from April 2024 observed that the frequency of certain words used in academic papers published in 2023 had changed and confirmed a strong correlation between these changes and the use of LLMs (Liang et al., 2024b; Geng and Trotta, 2024). Survey results also show that many researchers are utilizing LLMs in their work (Liao et al., 2024).

The detection of machine-generated text (MGT) has also attracted a lot of attention (Tang et al., 2024), but the performance of detectors has also

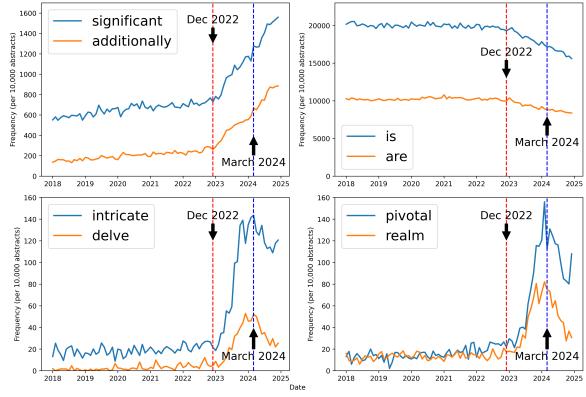


Figure 1: Frequency in arXiv abstracts of some of the words that were singled out as favored by ChatGPT around April 2024.

been questioned early on (Sadasivan et al., 2023; Weber-Wulff et al., 2023; Ghosal et al., 2023). Recent studies continue to show that some methods are not sufficiently robust (Zhang et al., 2024b; Wang et al., 2024; Creo and Pudasaini, 2025). The effectiveness of MGT detectors is also related to the model of LLMs and the type of text (Liu et al., 2024), and their accuracy may also be exaggerated (Doughman et al., 2024). The situations likely to arise in reality are more complicated and are not limited to a binary classification framework (Zhang et al., 2024a). Thus, examining and analyzing the ongoing evolution of word usage remains a useful and meaningful task.

Figure 1 illustrates the evolution in the frequency usage of some of the words that were singled out as favored by ChatGPT around April 2024. The frequency of “*significant*” and “*additionally*” continues to grow, while that of “*is*” and “*are*” continues its declining trend, as noted by Geng and Trotta (2024). Meanwhile, the frequency of some other words (e.g., “*intricate*” and “*delve*”) associated with LLMs begins to decrease after March and April 2024, which corresponds to the time when researchers identified these words in AI conference peer reviews (Liang et al., 2024a) and academic

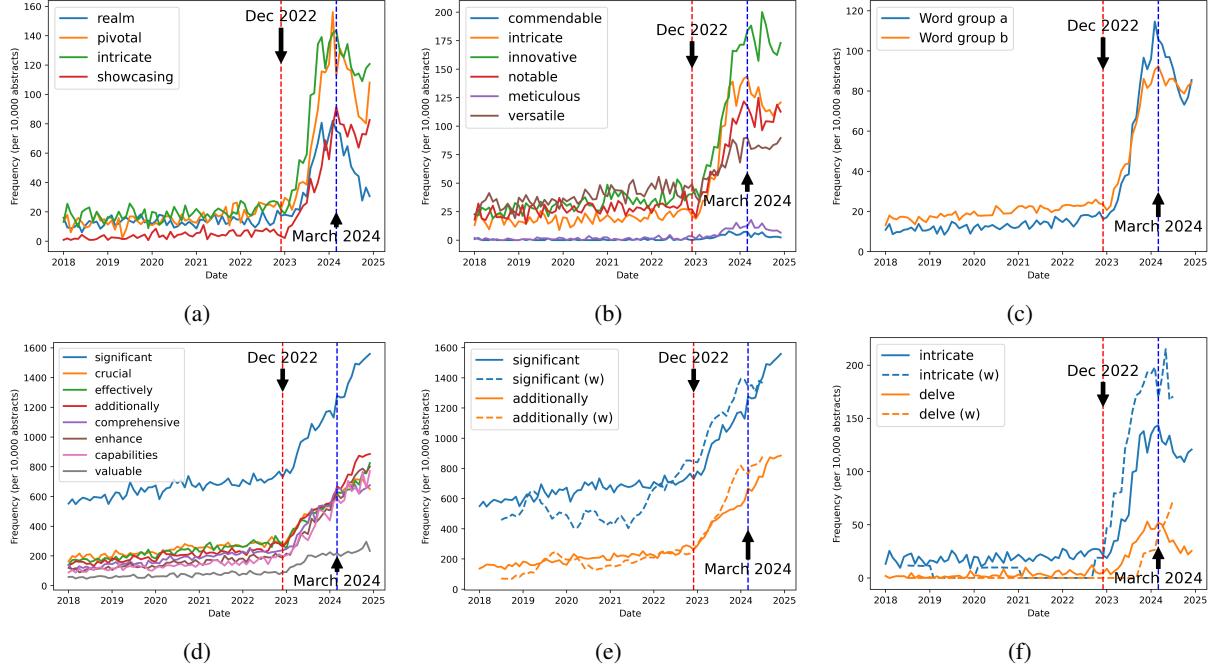


Figure 2: Frequency of words in arXiv abstracts previously identified as indicative of LLM usage. All word frequencies are normalized based on 10,000 abstracts. Word groups a and b correspond to the average frequencies of the words in 2a and 2b. The data for withdrawn papers represents a 12-month rolling average, labeled by “w”.

papers (Liang et al., 2024b).

Changes in the words used in academic writing, as discussed above, serve as an excellent example of AI and human coevolution. Researchers are constantly proposing new detection techniques, but the language and expressions of LLM users are also likely evolving due to their use of LLMs (Geng et al., 2024). Given the lack of a precise definition, LLM-generated text might even be undetectable in individual instances. Therefore, statistically measuring the impact of LLMs over a large corpus of texts is a more practical option.

This paper focuses on the following key points:

- The different fates of word frequencies after changes have been pointed out and scribed to LLMs usage.
- The challenges of MGT detectors.
- The long-term impact of LLMs in academic writing.

## 2 Data

**arXiv paper metadata** Metadata of arXiv papers updated weekly on Kaggle<sup>1</sup>. Our paper used version 214 of this dataset. Between January 2018 and December 2024, February 2018 and October

2024 recorded the lowest and highest numbers of papers, at 10,593 and 24,226, respectively. During this period, the total number of papers is 1,294,653.

**Withdrawn arXiv papers data** Withdrawn arXiv dataset (Rao et al., 2024), containing over 14,000 arXiv withdrawn papers up to September 2024.

## 3 Word Frequency Analysis

The analysis presented in Figure 2 is based on the abstracts of all arXiv papers submitted between 2018 and 2024. The frequency of words is calculated on a monthly basis and normalized per 10,000 abstracts.

Figure 2a shows the frequency of the 4 words highlighted by Liang et al. (2024b) and Figure 2b presents the frequency of the 6 words emphasized by Liang et al. (2024a). The former paper analyzes academic papers, while the latter focuses on AI conference peer reviews, and the average frequency of these words is shown in Figure 2c. The trend is clear: starting from April 2024, the frequency of these well-known LLM-style words began to decrease. Some other words show patterns of consistent growth or a rise followed by a decline, as illustrated in Figure 7a of the Appendix.

A study published in December 2024 also observed a decline in the use of certain words, such as “delve”, in some selected arXiv papers (Leiter et al., 2024). While they suggested that this was

<sup>1</sup><https://www.kaggle.com/datasets/Cornell-University/arxiv/data>

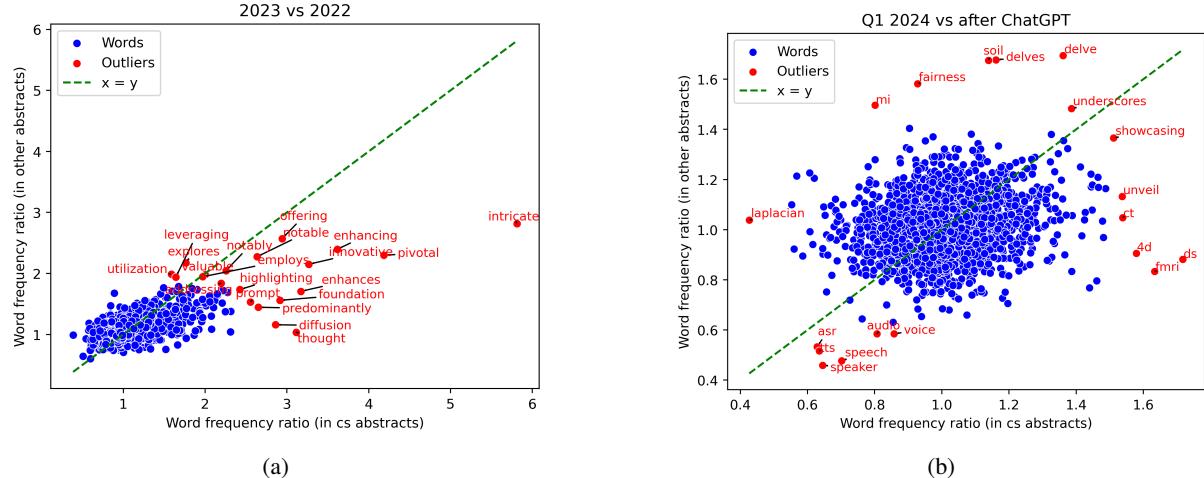


Figure 3: Comparing the ratio of word frequency between Computer Science abstracts and other disciplines. Only words that appear at least 20 times on average per 10,000 abstracts are plotted.

likely due to the release of GPT-4o in May 2024, we suggest that the main reason is that LLMs may have given these words a bad reputation. Many researchers noticed such kind of words in March and April and quickly changed their arXiv abstracts. If new LLMs were the cause, the drop in word frequency would have been delayed.

In addition, the frequency of words like “significant”, specifically pointed out by Geng and Trotta (2024), continues to grow. This may be because these terms are relatively common and frequently used, their presence alone would not easily lead one to suspect the text as the product of LLMs. Besides, as presented in Table 1, this article has attracted less attention than the former, for example, in terms of Google Scholar citation counts. Therefore, fewer researchers should have noticed the relationship between these words and LLMs.

We compared the results with the abstracts of withdrawn papers, as illustrated in Figures 2e and 2f. Given the small number of withdrawn papers, the 12-month rolling averages of their word frequency are used in the graphs. The frequency of some words, such as “intricate”, is higher in the withdrawn papers, but the difference is not very large, as is also the case in Figures 7b, 7c and 7d of the Appendix.

To better compare the changes in word frequency, we define  $R_{ij}(T_1, T_2)$  (the ratio of word  $i$  in the abstracts of category  $j$  between periods  $T_1$  and  $T_2$ ) as follows:  $R_{ij}(T_1, T_2) = \frac{f_{ij}(T_1)}{f_{ij}(T_2)}$ , where  $f_{ij}(T)$  is the frequency of word  $i$  in the abstracts of category  $j$  during the time period  $T$ .

We also categorized the abstracts into two groups based on the first category of the papers: com-

puter science (*cs*) and others. Figure 3a represents the ratio  $R$  between 2023 and 2022, where some words, like “diffusion”, are related to the research topics, but some other words have also become much more common in different fields. The ratio  $R$  in Figure 3b is calculated using the word frequency in the first quarter of 2024 divided by the word frequency from January 2023 to December 2024. Some words like “delve” and “showcasing” actually reached their peak usage from January to March 2024, and such words are very few. Figure 6 provides more detailed examples. Words that appear more often in *cs* paper abstracts have also clearly increased in other disciplines.

More researchers have now noticed issues with word usage and diversity in LLM-generated content (Kobak et al., 2024; Reviriego et al., 2024; Guo et al., 2024). Based on the above results, people are likely still using LLMs, but they may avoid some words that are typical of LLM output. Therefore, detecting LLM-generated content in real-world scenarios may become more difficult.

## 4 Challenges in Machine-Generated Text Detection

The first 1000 arXiv papers submitted each year from 2018 to 2025 were utilized for this part of the analysis. We also used the following two simple prompts to examine the differences between original arXiv abstracts and those revised by GPT-4o-mini:

- (P1) *Revise the following sentences: ...*
- (P2) *Don't use the following words in your responses: 'realm', 'pivotal', 'intricate', 'show-*

*casing’.* Revise the following sentences: ...

The results in Figure 4 reinforce the point that the frequency of certain words increases after LLMs revision. Using prompt P2, aimed at suppressing them, reduces the frequency of such words, although it does not eliminate them completely.

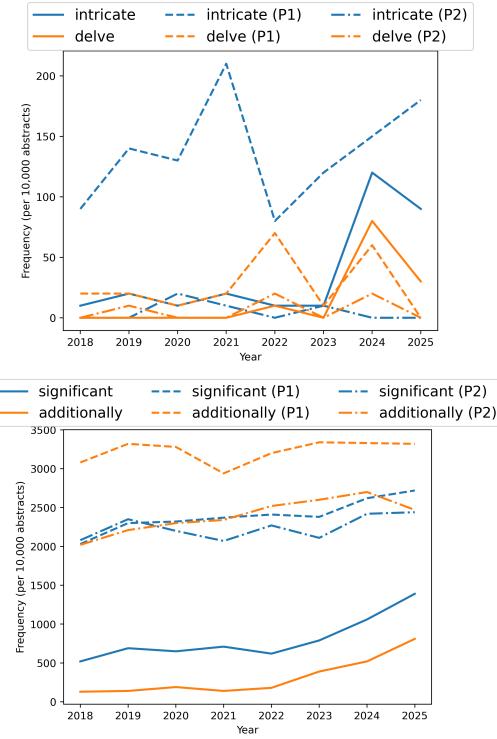
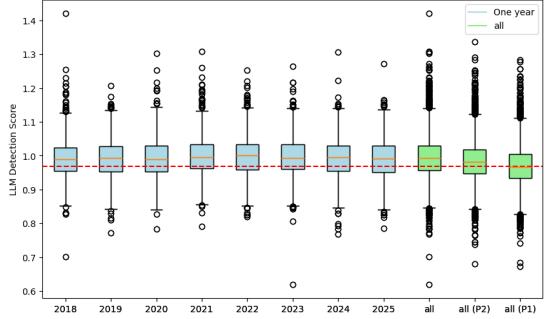


Figure 4: Comparison of word frequencies before and after LLM processing (with prompt P1 or P2).

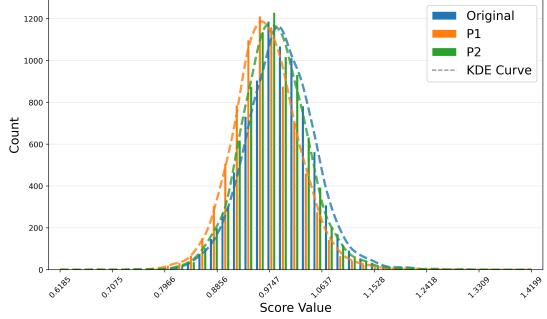
Figure 5 presents the detection results based on Binoculars (Hans et al., 2024), one of the state-of-art MGT detectors, where a lower score indicates a greater probability that the text is machine-generated. Unlike the results obtained with our frequency analysis, Binoculars on average does not return any difference in score for the real abstracts of papers as a function of time. Moreover, the change in the detection score between the original abstracts and the texts processed by LLMs (true positives) is not significant. Furthermore, the prompts used for processing can influence the results of MGT detectors. These results raise doubts about the accuracy of the detectors, given that they are analyzing texts that have been fully processed by LLMs.

## 5 Conclusions and Discussions

Humans and LLMs are coevolving and we can already conclude that, for this reason, the impact



(a) The last 3 columns all include abstracts of 8 years.



(b) KDE means kernel density estimation.

Figure 5: MGT detection results for real and LLM-processing abstracts (with prompt P1 or P2). A lower score indicates a greater probability that the text is machine-generated.

of LLMs on academic writing will fully assert itself over the long term. According to recent studies, people who frequently use ChatGPT for writing tasks can accurately distinguish AI-generated text (Russell et al., 2025), which implies they are also able to foil MGT detectors.

Grammarly can sometimes achieve effects similar to those of ChatGPT (Rudnicka, 2023), and the mix of human-written text and machine-generated text should be very common in academic writing. Detecting LLM-generated content with accuracy is becoming more difficult, perhaps impossible on a text-by-text basis.

Our findings suggest that some researchers may intentionally avoid using LLM-characteristic terms, but they are not as sensitive to some relatively common words. The gradual decrease in the occurrence of “is” and “are” in arXiv abstracts is an excellent example of such a trend, which we ascribe to a more subtle – and continually increasing – LLM influence.

Therefore, using the frequency of more common words (Geng and Trotta, 2024) to measure the impact of LLMs on a vast number of publications will be more reliable, although this approach is less suitable for the precise detection of short texts.

## References

- Aldan Creo and Shushanta Pudasaini. 2025. Silverspeak: Evading ai-generated text detectors using homoglyphs. In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, pages 1–46.
- Jad Doughman, Osama Mohammed Afzal, Hawau Olamide Toyin, Shady Shehata, Preslav Nakov, and Zeerak Talat. 2024. Exploring the limitations of detecting machine-generated text. *arXiv preprint arXiv:2406.11073*.
- Mingmeng Geng, Caixi Chen, Yanru Wu, Dongping Chen, Yao Wan, and Pan Zhou. 2024. The impact of large language models in academia: from writing to speaking. *arXiv preprint arXiv:2409.13686*.
- Mingmeng Geng and Roberto Trotta. 2024. Is chatgpt transforming academics' writing style? *arXiv preprint arXiv:2404.08627*.
- Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amit Singh Bedi. 2023. Towards possibilities & impossibilities of ai-generated text detection: A survey. *arXiv preprint arXiv:2310.15264*.
- Andrew Gray. 2024. Chatgpt" contamination": estimating the prevalence of llms in the scholarly literature. *arXiv preprint arXiv:2403.16887*.
- Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2024. Benchmarking linguistic diversity of large language models. *arXiv preprint arXiv:2412.10271*.
- Abhimanyu Hans, Avi Schwarzschild, Valeria Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*.
- Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2024. Delving into chatgpt usage in academic writing through excess vocabulary. *arXiv preprint arXiv:2406.07016*.
- Christoph Leiter, Jonas Belouadi, Yanran Chen, Ran Zhang, Daniil Larionov, Aida Kostikova, and Steffen Eger. 2024. Nllg quarterly arxiv report 09/24: What are the most influential current ai papers? *arXiv preprint arXiv:2412.12121*.
- Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. 2024a. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*.
- Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. 2024b. Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*.
- Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X Zhang. 2024. Llms as research tools: A large scale survey of researchers' usage and perceptions. *arXiv preprint arXiv:2411.05025*.
- Jialin Liu and Yi Bu. 2024. Towards the relationship between aigc in manuscript writing and author profiles: evidence from preprints in llms. *arXiv preprint arXiv:2404.15799*.
- Yule Liu, Zhiyuan Zhong, Yifan Liao, Zhen Sun, Jingyi Zheng, Jiaheng Wei, Qingyuan Gong, Fenghua Tong, Yang Chen, Yang Zhang, et al. 2024. On the generalization ability of machine-generated text detectors. *arXiv preprint arXiv:2412.17242*.
- Dino Pedreschi, Luca Pappalardo, Emanuele Ferragina, Ricardo Baeza-Yates, Albert-László Barabási, Frank Dignum, Virginia Dignum, Tina Eliassi-Rad, Fosca Giannotti, János Kertész, et al. 2024. Human-ai co-evolution. *Artificial Intelligence*, page 104244.
- Delip Rao, Jonathan Young, Thomas Dietterich, and Chris Callison-Burch. 2024. Withdrarxiv: A large-scale dataset for retraction study. *arXiv preprint arXiv:2412.03775*.
- Pedro Reviriego, Javier Conde, Elena Merino-Gómez, Gonzalo Martínez, and José Alberto Hernández. 2024. Playing with words: Comparing the vocabulary and lexical diversity of chatgpt and humans. *Machine Learning with Applications*, 18:100602.
- Karolina Rudnicka. 2023. Can grammarly and chatgpt accelerate language change? ai-powered technologies and their impact on the english language: wordiness vs. conciseness. *Procesamiento de Lenguaje Natural*, 71.
- Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. People who frequently use chatgpt for writing tasks are accurate and robust detectors of ai-generated text. *arXiv preprint arXiv:2501.15654*.
- 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*.
- Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. 2024. The science of detecting llm-generated text. *Communications of the ACM*, 67(4):50–59.
- Yichen Wang, Shangbin Feng, Abe Bohan Hou, Xiao Pu, Chao Shen, Xiaoming Liu, Yulia Tsvetkov, and Tianxing He. 2024. Stumbling blocks: Stress testing the robustness of machine-generated text detectors under attacks. *arXiv preprint arXiv:2402.11638*.
- Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltynek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, 19(1):26.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, et al. 2024a. Llm-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 409–436.

Yuehan Zhang, Yongqiang Ma, Jiawei Liu, Xiaozhong Liu, Xiaofeng Wang, and Wei Lu. 2024b. Detection vs. anti-detection: Is text generated by ai detectable? In *International Conference on Information*, pages 209–222. Springer.

## A Appendix

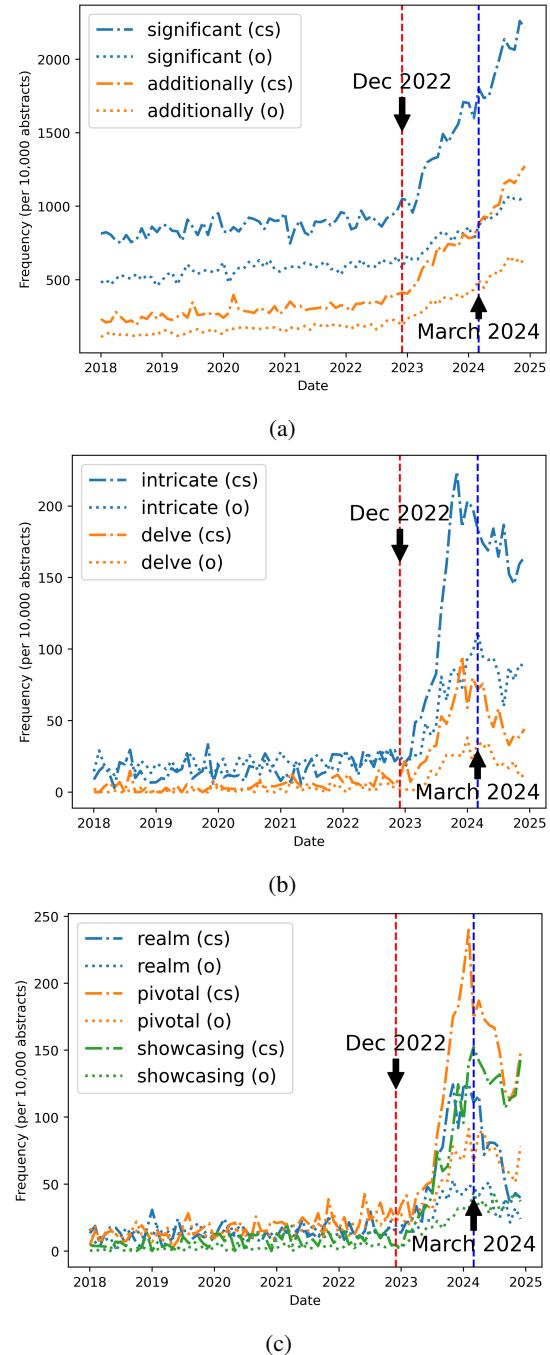


Figure 6: Frequency of some words in arXiv abstracts.

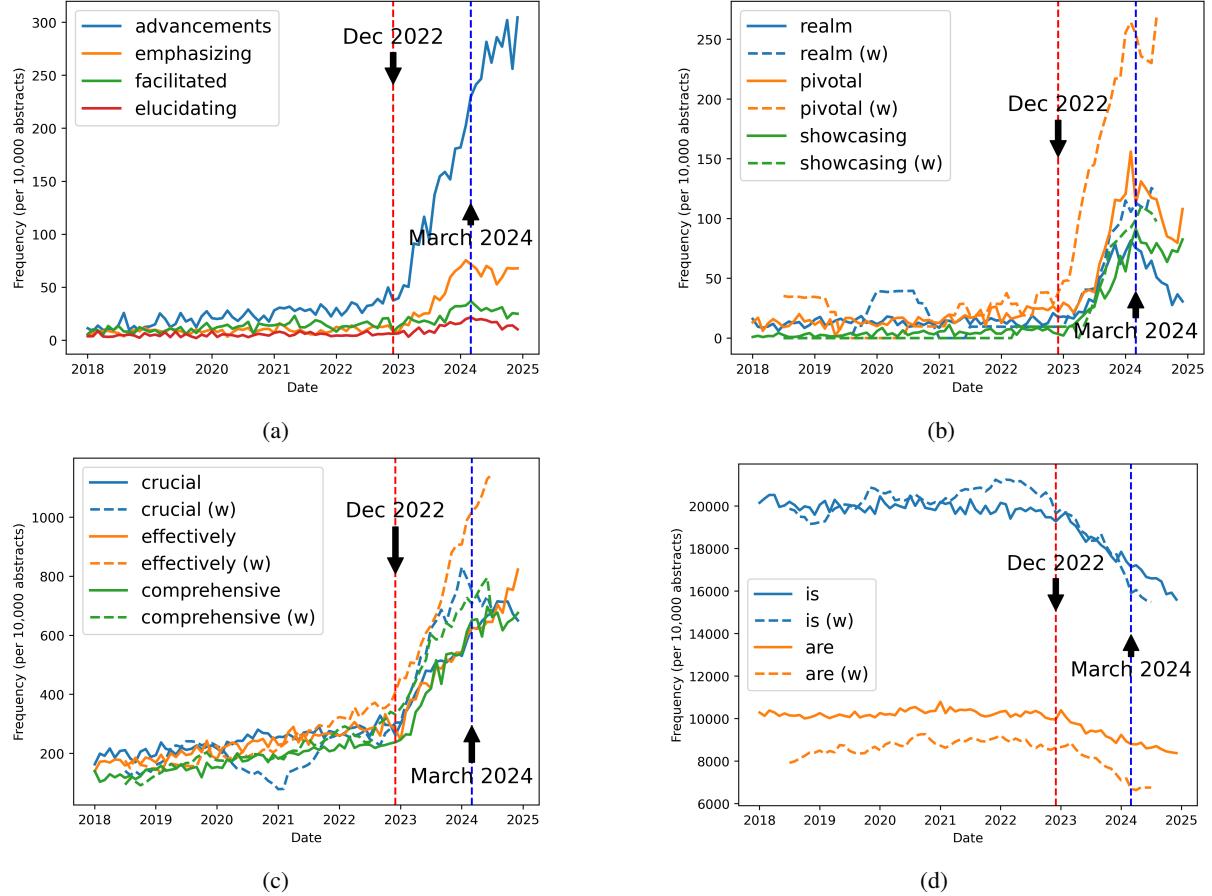


Figure 7: Frequency of some words in arXiv abstracts.

Paper	Citations	Highlighted words
(Liang et al., 2024a)	87	<b>commendable, innovative, meticulous, intricate, notable, versatile.</b>
(Liang et al., 2024b)	58	<b>pivotal, intricate, realm, showcasing.</b>
(Gray, 2024)	41	words listed based on Liang et al. (2024a)
(Geng and Trotta, 2024)	11	<b>significant, crucial, effectively, additionally, comprehensive, enhance, capabilities, valuable.</b>
(Liu and Bu, 2024)	4	

Table 1: Papers on word frequency analysis published in March and April 2024 (submitted to arXiv). The Google citation counts are as of January 16, 2022.