Towards Linguistic Shifts in AI Research Writing in the LLM Era

Anonymous ACL submission

Abstract

We analyze 51,877 papers from ten AI conferences (2018–2025) to examine linguistic shifts in the era of LLMs. Recent AI papers exhibit increased AI content, with greater specific non-technical words, longer and more uniform-length sentences, greater syntactic complexity, fewer function words, and preferences for active voice and nominalization. Stylometric analysis shows that AI writing tools exhibit a measurable influence on the writing styles of post-LLM papers.

1 Introduction

"Large language models (LLMs) have demonstrated unprecedented capabilities across a wide range of tasks, advancing at a rapidly evolving pace." Sound familiar? That might be the sound of a thousand prompts echoing through recent AI publications. Since 2023, researchers have gained access to powerful LLMs that can draft, revise, and polish scientific prose. While these tools promise to enhance productivity, their influence on academic writing style remains less addressed. Are AI tools systematically reshaping how researchers write, to what extent, and with what concerns?

This short paper explores linguistic shifts in AI research writing across the pre- and post-LLM eras. We analyze 51,877 papers from ten leading AI conferences spanning Natural Language Processing, Computer Vision, and Machine Learning between 2018 and 2025. We begin by validating that AI content in these papers became more prevalent starting in 2023 and peaking in 2025. We then identify linguistic shifts from the pre-LLM to the post-LLM era across three dimensions: word choice, syntactic patterns, and stylometric analysis.

Our experiments show significant changes in writing style, including increased use of certain non-technical words, longer sentences with more uniform length, greater syntactic complexity, decreased use of function words, and a tendency toward active voice with nominalization. Through stylometric analysis, we show that AI writing tools exhibit a measurable influence on the writing styles of post-LLM papers.

2 Data Collection and Statistics

To investigate linguistic shifts of research writing in the LLM era, we compiled a corpus of papers published between 2018 and 2025. We selected ten leading AI conferences across three categories: Natural Language Processing (ACL, EMNLP, NAACL, COLING), Computer Vision (CVPR, ECCV, ICCV), and Machine Learning (NeurIPS, ICLR, ICML). We collected paper PDFs and metadata through PaperCopilot (https://papercopilot.com/), including publication year, authors, affiliations, countries, presentation formats (e.g., spotlight, oral, poster at NeurIPS), and sections (main or findings, available for ACL* venues). All PDF files were converted to plain text format using PyMuPDF (https://pypi.org/project/PyMuPDF/), with images removed. We further processed the paper content by removing tables, equations, the appendix, and references using a script. After excluding incomplete or formatting-corrupted files (fewer than 100 characters after processing), our final corpus comprises 51,877 papers distributed across eight years and ten conferences. The detailed data statistics are shown in Appendix A.

For all analyses in Sections 4 and beyond, we partitioned the papers into two groups by temporal period: the pre-LLM era (2018–2022), with 26,419 papers (50.9%), and the post-LLM era (2024–2025), with 25,458 papers (49.1%). This partition corresponds to the period before and after the public release of ChatGPT and the widespread availability of other LLMs to researchers, enabling a comparative analysis of shifts in writing patterns.

3 A Pilot Experiment on AI Prevalence

If we hypothesize that the widespread adoption of AI writing tools may have influenced scientific writing patterns, we should observe at least an increasing prevalence of AI-generated content over the years in the corpus. To validate this, we employed the Winston AI API (https://gowinston.ai/ai-content-detection-api/) to detect AI content in the papers. Due to budget constraints, we sample 10 papers per year per conference and use only the abstract as the analysis target, which yields 800 samples for this pilot experiment.

For each sample, the Winston AI detection API returns an overall AI score computed as a weighted (by sentence length) average of sentence-level AI scores. Each sentence's AI score represents the estimated probability that the sentence was AI-generated, ranging from 0 to 100%, where 0 indicates human writing, and 100 indicates fully AI-generated content.

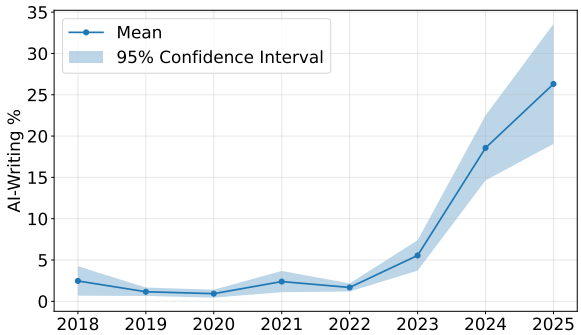

Figure 1: Trend of averaged AI score (%).

Figure 1 shows the trend of AI scores averaged over papers per year. We observe relatively stable scores near 0% from 2018 to 2022 (with minor deviations due to detection noise), followed by a slight increase in 2023 and dramatic increases in 2024 and 2025, peaking around 30% in 2025. Interestingly, we find that although NLP papers show the earliest AI adoption in 2024, ML papers exhibit the highest AI usage levels by 2025, while CV papers show the lowest AI scores (See Appendix B). Additionally, we observe a significant difference in AI usage across paper sections in ACL* venues, with Findings papers showing higher AI scores (39.9%) than Main conference papers (15.1%). We also found a significant difference in AI usage across author countries. However, we did not observe significant differences between LLM-related papers and papers on other topics.

4 Linguistic Shifts in the LLM Era

Human writing style and word preferences are typically stable over short time frames, and sudden shifts can be noteworthy. This section discusses linguistic shifts in scientific writing in the era of LLMs in three dimensions: word salience, sentence and part-of-speech (POS) structures, and stylometry.

4.1 Word Salience

The most direct way to examine linguistic shifts is to look at the word choice. To identify such, we can simply compute the difference between the TF-IDF scores for each word in the two partitions (papers grouped in pre-LLM and post-LLM eras): $\text{TF-IDF}_{\text{post-LLM}} - \text{TF-IDF}_{\text{pre-LLM}}$. A larger positive value indicates words that became more salient in the post-LLM era.

By examining the top-k words ranked by this difference, we can identify both technical terms that emerged with new research directions (e.g.,"LLM," "prompting") and, more importantly, shifts in non-technical vocabulary that may reflect changes in writing style. By ranking, the most representative unigrams in the era of LLMs are:

- **Technical**: *llm, gpt, transformer, prompt, inference, pretraining, fine-tuning, generation, reasoning*
- **Non-technical**: *however, therefore, moreover, thus, consequently, additionally, furthermore, notably, significantly, specifically, pose, accordingly, ultimately, approximately, relatively, primarily, generally, similarly, substantial, robust, overall*

Figure 2 additionally visualizes the trends of frequency of certain non-technical words. Several words exhibit sharp increases beginning in 2024, such as *notably*, *subsequently*, *predominantly*, and *primarily*.

4.2 Sentence and POS Structures

Table 1 shows the sentence-level statistics (upper-half) and Part-of-Speech statistics (lower-half). Standard deviations of the number of words per sentence capture variability in sentence length. The short/long sentence ratio is the proportion of sentences falling into the short (<15 words) or long (>25 words) categories. Overall, the papers published in the post-LLM era tend to have longer

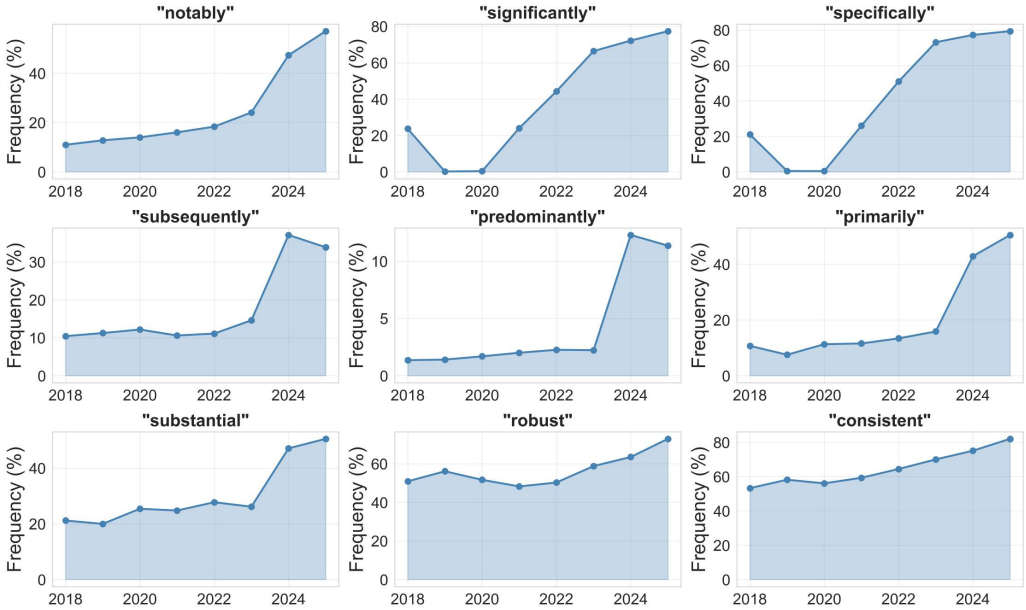

Figure 2: Word trend over years.

sentences and more uniform sentence lengths. The usage of commas, semicolons, and conjunctions per sentence is also increased, indicating greater use of clauses and more complex sentence structures.

For part-of-speech (POS) distributions, we use NLTK's POS tagger (https://www.nltk.org/), with densities calculated as the number of occurrences of specific POS tokens per 100 words. We additionally computed the passive voice ratio by detecting auxiliary verbs followed by past participles (VBN tags) in the two groups and observed a 28.17% decrease in post-LLM papers. In general, papers in the era of LLMs use more active voice and display a nominalization trend (18.91% increase of noun/verb ratio). Function words showed consistent declines (pronouns (-13%), auxiliaries (-18%), and modals (-17%)), indicating a more rigid "academic" writing style that is frequently suggested by AI writing tools. All differences achieved statistical significance with t-test ($p < 0.01$ for Avg # words and Std. # words and $p < 0.0001$ for others).

4.3 Stylometric Analysis for AI Influence

While the previous analysis identified stylistic shifts between pre-LLM and post-LLM papers, it remains unclear whether these changes reflect AI tool usage or other recently developed human preferences (e.g., the increased use of "notably"). To address this, we employ stylometric analysis, which quantifies writing style based on statistics and is often used to attribute authorship to disputed documents.

We use Burrows' Delta (Burrows, 2002), a stylometric approach based on the discrepancies of word frequencies. Given a document x and a collection of documents with known author y, Delta computes the distance between the document x and the collection c_y as:

$$\Delta(x, c_y) = \frac{1}{|V|} \sum_{w \in V} \left| z_x(w) - z_{c_y}(w) \right|, \quad (1)$$

where V denotes the vocabulary size and $z_x(w)$ and $z_{c_y}(w)$ are the z-score of normalized frequencies of word w in document x and collection c_y, respectively. A lower Delta value indicates greater stylistic similarity.

We constructed two reference authorship groups: (1) 300 papers from the pre-LLM era (human-written baseline) and (2) the *same* 300 papers rephrased by GPT-4o-mini (AI-rewritten baseline). We then sampled 300 papers from the post-LLM era as disputed documents.

Since post-LLM papers naturally differ from pre-LLM papers due to evolving research topics, comparing post-LLM papers against both human-written and AI-rewritten versions of the *same* pre-LLM papers isolates stylistic shifts impacted by AI. If post-LLM papers contain AI-generated content, their style should be more similar to the AI-rewritten baseline than to the human-written baseline.

Metric / Feature	Pre-LLM	Post-LLM	% Change
Sentence Structure			
Avg # words per sentence	23.699	24.101	↑ 1.7%
Std. of # words per sentence	11.956	11.285	↓ 5.6%
Short/long sentence ratio	0.494	0.428	↓ 13.3%
Avg # commas per sentence	0.971	1.161	↑ 19.6%
Avg # semicolons per sentence	0.016	0.016	↑ 4.2%
Part-of-Speech Tag Density (Per 100 Words)			
Determiner	9.420	7.970	↓ 15.39%
Verb	11.869	10.751	↓ 9.42%
Noun	33.953	36.518	↑ 7.55%
Noun/Verb ratio	2.955	3.514	↑ 18.91%
Preposition	12.013	11.050	↓ 8.02%
Pronoun	2.399	2.096	↓ 12.62%
Auxiliary	0.597	0.489	↓ 18.11%
Modal Verb	0.588	0.487	↓ 17.21%
Lexical	58.987	60.484	↑ 2.54%
Conjunction	2.381	2.495	↑ 4.80%

Table 1: Sentence and Part-of-Speech Structure Comparison (Pre-LLM era vs. Post-LLM era).

Figure 3 shows the distribution (histogram, with x-axis being the delta values) of the delta between each post-LLM paper and the two authorship groups, respectively. We can observe that the distribution of delta between the post-LLM paper and the AI-rewritten baseline shifts leftwards, indicating that post-LLM papers are stylistically closer to AI-generated text.

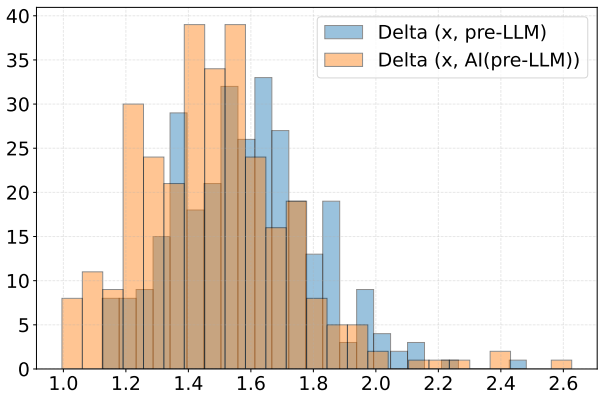

Figure 3: Delta distributions of post-LLM papers between the pre-LLM papers and AI-rewritten pre-LLM papers.

5 Related Work

Large language models such as OpenAI GPT, Gemini, and Anthropic Claude have become common tools for drafting and revising academic text (OpenAI, 2023; Anthropic, 2024; Anil et al., 2024); however, the dynamic influence of AI in academic writing style remains unclear. Linguistic style analysis leverages statistical methods to examine how language conveys underlying meaning, focusing on elements such as word choice, sentence structure, and rhetoric, with applications including stylometry for authorship attribution (Burrows, 2002; Stamatatos, 2009; Jockers, 2013; Bevendorff et al., 2019; Barlas and Stamatatos, 2020).

Recent work has begun examining AI's impact on spoken and written language at scale, documenting shifts in lexical diversity, syntactic complexity, and stylistic uniformity across scientific and online corpora (Gero et al., 2024; Zhou et al., 2024; Bao et al., 2025).

6 Conclusion and Discussion

Through linguistic analysis, we find measurable shifts in writing style and AI influence in post-LLM papers. These findings may also raise concerns about the evolution of scholarly communication. Like urban sparrows developing shorter wings within a few generations due to traffic, academic writing may be rapidly transforming through the collective adoption of AI-preferred patterns. Moreover, when researchers receive AI suggestions for alternative phrasing, to what extent do they defer to AI judgment or gradually adopt them as if learning from oracles? Understanding these dynamics is crucial as we navigate an era in which the line between human and machine authorship is increasingly blurred.

Limitations

We intentionally focused on leading AI conferences, as researchers in these communities are among the earliest and most intensive adopters of AI writing tools, making them interesting subjects for linguistic shifts. While our analysis reveals strong correlations between the post-LLM era and specific linguistic patterns, proving AI writing tools as the sole causal factor is beyond the scope of this paper.

Ethics Statement

This work analyzes publicly available academic papers without human subjects involvement. We do not identify individual papers or authors as using AI tools, nor do we examine differences in AI tool usage across demographic groups, author characteristics, or institutional affiliations.

References

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805. Only available as preprint.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Yifan Bao et al. 2025. Large language models and shifts in academic writing style. *arXiv preprint arXiv:2505.12218*.

Gerasimos Barlas and Efstathios Stamatatos. 2020. A comparative study on authorship attribution in greek social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2059–2066, Marseille, France. European Language Resources Association.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2019. Generalizing unmasking for short texts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 654–659, Minneapolis, Minnesota. Association for Computational Linguistics.

John Burrows. 2002. 'delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17:267–287.

Katy Gero et al. 2024. The impact of large language models on human writing. *arXiv preprint arXiv:2409.01754*.

Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, Urbana, IL.

OpenAI. 2023. Gpt-4 technical report. Technical report, OpenAI.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Xinyi Zhou et al. 2024. Detecting the influence of ai on scientific writing. *arXiv preprint arXiv:2409.09613*.

A Data Statistics

Table 2 presents the details of the corpus by year and conference.

B Data Statistics

Figure 4 illustrates the AI score trend in three disciplines.

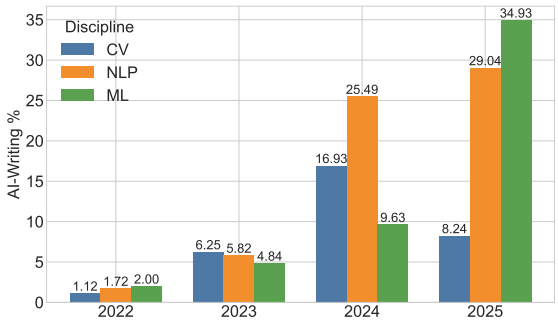

Figure 4: AI scores in different AI-related disciplines.

C Computation Cost

Winston AI: $70
OpenAI API: $25

Conf.	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**	**2024**	**2025**	**Total**
ACL	–	–	–	687	999	1,546	1,278	–	4,510
COLING	–	–	632	–	629	–	1,551	–	2,812
CVPR	979	1,293	1,318	1,587	1,500	1,441	1,424	1,500	10,042
ECCV	683	–	1,358	–	1,011	–	1,053	–	4,105
EMNLP	–	–	–	1,307	1,118	1,715	2,368	–	6,508
ICCV	–	986	–	1,056	–	1,598	–	–	3,640
ICLR	992	501	1,638	1,463	1,067	1,380	547	1,971	9,559
ICML	592	772	1,079	–	1,232	1,357	1,552	1,516	8,100
NAACL	–	–	–	533	702	–	907	1,268	3,410
NeurIPS	1,007	1,426	1,897	1,526	2,669	1,574	1,187	–	11,286
Total	**4,253**	**4,978**	**7,922**	**8,159**	**10,927**	**10,611**	**11,867**	**6,255**	**51,877**

Table 2: Data Statistics.