Does AI-assisted scientific writing affect readability?

Daniel Duhnev[1], Marcelo Jimenez[1], and Lydia Krifka-Dobes[1]

Universitat Pompeu Fabra (UPF), Barcelona, Spain {daniel.duhnev, marcelo.jimenez, lydia.krifka-dobes}@upf.edu

Abstract. The growing use of Large Language Models (LLMs) in academic writing raises questions about their influence on the clarity and accessibility of scientific communication. This study examines whether the adoption of AI-assisted writing tools correlates with measurable changes in readability across peer-reviewed scientific literature. Using the OpenAlex API, we compiled a dataset of 33,000 abstracts published between 2015 and 2025 across three domains - Artificial Intelligence, Medicine, and Business. Readability was assessed using three established formulas - Flesch–Kincaid Grade Level (FKGL), Automated Readability Index (ARI), and Dale–Chall. Results reveal distinct patterns. The AI domain shows a consistent rise in complexity that sharply accelerates after 2022, whereas the Medicine and Business domains show high year-to-year volatility before converging in a less pronounced upward spike in 2023-2025. These findings suggest that the increased integration of AI tools in academic writing may be associated with more complex sentence structures and denser vocabulary, potentially hindering accessibility rather than improving it. It also shows that writers in AI likely adopted their tools earlier, and to a greater extent, than other scientists. Future work should employ statistical modeling to confirm the significance of these trends and explore other textual features such as text coherence.

Keywords: Meta-Research · Scientific Writing · Artificial Intelligence · LLM · Readability · Textual Analysis · Observational Study.

1 Introduction

The proliferation of advanced AI tools, particularly Large Language Models (LLMs) like ChatGPT, has fundamentally altered the landscape of scientific writing [1]. Researchers are increasingly using these tools for drafting, editing, and refining their manuscripts. While the potential benefits for productivity are widely discussed, the specific impact of this AI assistance on the final textual product remains an open and critical question.

These very recent changes in the way we write are being investigated as they develop. Experimental studies comparing human-written abstracts to AI-generated ones have found that AI-generated text can be "vaguer" or more "formulaic" [2], and while sometimes scoring as more readable, can be of inferior

scientific quality [3]. While these experiments are informative, the net, real-world effect of this adoption on published literature is only just beginning to be measured.

The question of this real-world effect of the use of AI in composing text is becoming a opic in research. The baseline trend, established long before the rise of LLMs, was already one of slowly decreasing readability in science [4]. However, recent observational studies analyzing preprint servers have provided the first evidence of a new, sharp inflection. Most notably, Alsudais (2024) analyzed arXiv abstracts and found a significant decrease in readability (i.e., increased complexity) following the 2022 release of ChatGPT [6].

This paper builds on these findings. It remains unknown if this trend of declining readability holds for the formal, peer-reviewed literature, or how it varies by discipline. We address this gap by performing a large-scale analysis of abstracts from 2015 to 2025. We focus on three distinct fields: **AI**, **Medicine**, and **Business**. This investigation across three distinct fields and over the span of 11 years, in the historic time when LLMs started to be used, allows us to investigate the precise timing and magnitude of any trend changes.

We measure readability using three standardized tests (Flesch-Kincaid, ARI, and Dale-Chall) that target different features of text. Our central hypothesis is twofold: (1) We hypothesize that, confirming the trend in preprints [6], the readability trends will show a statistically significant inflection point towards decreased readability around the year 2022. (2) We further hypothesize that this inflection will be earliest and steepest in the AI domain, as the authors of these articles are more likely early adopters of the new ways that AI can be used to write text. Such findings would suggest that current AI assistance correlates with an obstruction of readability, reinforcing initial concerns about its impact on scientific communication.

2 Methodology

Our methodological approach consists of two main parts. First, we detail our data acquisition, where we utilised the OpenAlex API [10] to construct a dataset covering an 11-year period (2015-2025). This dataset comprises 1,000 abstracts per year for each of three distinct domains: AI, Medicine, and Business. Second, we apply three established readability metrics to this data - the Flesch-Kincaid Grade Level (FKGL), the Automated Readability Index (ARI), and the Dale-Chall Formula, and plot the mean scores for each domain over the 11-year period to analyze the resulting trends.

2.1 Corpus Design and Data Acquisition

To construct our dataset, we used the open-source tool OpenAlex API [10] to gather data across an 11-year period (2015-2025) for three distinct academic domains. We used OpenAlex's topic filtering as opposed to title keywords to build three distinct datasets compiled in a CSV file. These were works tagged

with "Artificial Intelligence" (AI domain), "Medicine" (Medicine domain), and "Business" (Business domain).

For each of the three domains, we randomly sampled 1,000 papers that included an abstract for each year from 2015 to 2025, inclusive. This resulted in a total corpus of 33,000 abstracts.

2.2 Data Pre-processing and Analysis

The abstract data from OpenAlex was provided in the abstract_inverted_index format, so we first used a Python script to reconstruct the original, ordered plaintext abstracts.

We used the `textstat` Python library to calculate the three established scores for our analysis: the FKGL [7], the ARI [8], and the Dale-Chall Formula [9]. For the final trend analysis, the individual abstract scores were aggregated to calculate the mean readability score for each of the three domains for each year from 2015 to 2025. These means were then plotted on a time-series graph to visually inspect trends and identify potential inflection points.

The analysis in this paper is primarily based on the visual inspection of plotted trends. This paper could and should be extended during the next review process to include formal statistical modeling. Regression analysis should be performed to confirm the statistical significance of the observed trends or to test for a definitive inflection point after 2022. This statistical validation remains a critical step for future work.

2.3 Rationale for Readability Metrics

Standardized measures for readability have been developed since the 1930'ies, with the goal to improve the reading abilities of pupils, students, and the adult population. These measures face two challenges: They have to be applied easily to text, and they have to be related to the ease or difficulty of the reading experience for certain defined populations. The most influential measure goes back to the work of Rudolf Flesch, but was further developed and supplemented by other approaches.

Flesch–Kincaid Grade Level (FKGL). The FKGL is one of the most widely used readability measures. It was developed for the U.S. Navy in the 1970s to help produce clear and accessible technical manuals. It is based on the length of sentences, and the length of words. The logic behind it is straightforward: longer sentences require more memory to follow, and longer words take more time to pronounce or decipher. The combination of these two features results in an estimate of how many years of education a reader would need to comfortably understand the text. In the FKGL scale, a score of 8 means the text should be readable for an eighth grader, 12 corresponds to high school level, and 16 or more suggests university-level difficulty. In practice, very few scientific abstracts score below 14. Most fall between 16 and 20.

It is in general a good thing to examine sentence length. This is related to readability when measured independently, but it does not mean that a text with short sentences is automatically easier to read. Very short sentences can make a text sound robotic, while too many long ones can feel exhausting. Readability exists somewhere between those extremes. However, this depends on how the clauses are stitched together. Sentences can be complex because of their grammatical structure. For example, subordinated sentences can be experienced as more complex than coordinated sentences. A sentence like *This study explores the emergent coherence of recursive neural systems.* is not particularly long, and it does not contain any sub-clause. However, it has a subject that is not animate, which is very rare for a transitive verb (of course, *this study* is scientific lingo standing for the authors of the paper). It also expresses an event with a nominalization, which is felt to be more complex (compare the example sentence to *we explore how recursive neural system become more coherent*). Such factors for complexity will not be captured by the FKGL. Also, it is not difficult to compose a text with simple sentences that is very hard to read because it lacks information about how the sentences are related to each other, the so-called coherence markers.

The second dimension of the FKGL are long words, where length is measured by syllables per word. It is plausible that long words impede readability. This is not only because the fact that they are longer; rather, long words also tend to be rarer. This is because od a well-known tendency in human language towards efficiency; it would be detrimental to have very long words that are very frequent. Hence, longer words tend to be more difficult to process because they are rarer, and it takes more time and effort to mentally access their meaning in the reading process. However, it should be mentioned that FKGL is inconsistent, as it counts words by spaces. It misses compounds like *attention deficit disorder tratement*, as English allows for spaces within many compounds except simple, well established ones like *blackbird*. This is in contrast to German, where our example would be realized as one string (*Aufmerksamkeitsstörungsbehandlung*) and would count as a highly complex word. It should be mentioned that complex compounds are a prominent feature of scientific texts; the FKGL has a blind spot for this dimension of complexity.

Another limitation of FKGL is that it cannot distinguish between a long but elegant argument and a long but confusing one. It ignores meaning entirely. Yet it remains useful because it offers a repeatable way to track structural change. If the average FKGL rises sharply after 2023, that signals a shift in how scientific authors, possibly assisted by AI, structure their thoughts on paper.

Automated Readability Index (ARI). The ARI was designed for machines. It comes from the early days of computational linguistics, when computers could not easily count syllables or stress patterns. ARI replaced that with a simpler approach: counting characters and sentences.

In essence, ARI measures how visually dense a text is. It counts word length simply by characters per word, not by syllables, as the characters generate the

visible density of words. The longer the words, the higher the score; the longer the sentences, the higher the score. The result is another grade-level number, roughly equivalent to FKGL.

This formula is especially sensitive to terminology. Words like *electrophysiology*, *multimodal*, or *metacognitive* immediately increase the ARI value. In scientific abstracts, this often reflects how specialized the field has become. The presence of AI-related terms adds another layer. Expressions such as *neural transformer architecture*, *emergent agentic behavior*, or *quantum-inspired coherence* inflate the score further.

ARI therefore captures the growing visual weight of scientific language. It measures how compact or inflated the surface of a text has become. ARI tells us how thick the vocabulary feels to the eye.

When we compare abstracts from 2015 to 2025, a rising ARI score suggests an evolution toward longer and more complex words. This may not mean the ideas themselves are more complex, but that the phrasing has become heavier. AI writing tools might contribute to this shift. Many of these systems tend to select formal, high-register words. A human might write *"We tested this idea using simulations."* A model might produce *"This concept was evaluated through computational experimentation."* The second version feels more official, but it nearly doubles the number of characters.

The strength of ARI lies in its simplicity and reproducibility. It works well on large datasets and avoids subjective assessment. Its limitation is that it cannot tell necessary precision from unnecessary ornamentation. Still, it shows how scientific language evolves in its visible surface — not what it says, but how much space it takes to say it.

Dale–Chall Readability Formula. The Dale–Chall Formula focuses on vocabulary familiarity. Developed by Edgar Dale and Jeanne Chall in the 1940s, it estimates difficulty based on how many words in a text appear on a list of about 3,000 "easy" English words known to most fourth-grade students. Words not on this list are counted as "difficult." The more difficult words a text contains, the higher its score, and the harder it is presumed to read.

Unlike FKGL and ARI, the Dale–Chall test is grounded in cognitive psychology. It assumes that readers struggle less with syntax and more with unfamiliar vocabulary. It measures the distance between everyday language and the jargon of a field.

Scientific writing often relies on specialized terms. In recent years, especially with the rise of AI-related research, certain keywords recur: *recursive*, *evolutionary*, *fractal*, *complexity*, *quantum*, *emergent*, and *coherence*. New terminology is constantly introduced as well — what we might call "AI-ish" language. These words carry specific meaning within a discipline but increase the lexical difficulty of the text for non-specialists.

The Dale–Chall score thus acts as a proxy for jargon density. If the score rises over time, it means that scientific communication is drifting further from general accessibility. The strength of this formula is its close relation to how

people actually read. It estimates how often a reader is likely to encounter a word they do not know. However, its dictionary is dated and culturally specific, reflecting mid-century American English. Still, when used comparatively across years or within the same domain, it remains a sensitive indicator of lexical shift.

3 Results

We visualized the temporal evolution of average readability scores across three representative domains - Artificial Intelligence, Medicine, and Business. Each figure displays the mean yearly score for the ARI, FKGL, and Dale–Chall formula from 2015 to 2025, including standard error of the mean (SEM) error bars.

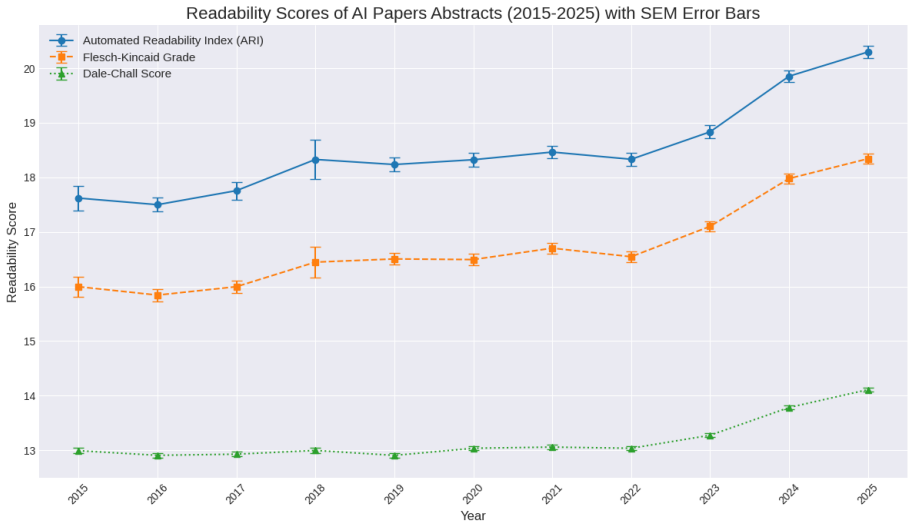

Fig. 1. Readability trends in **Artificial Intelligence** (2015–2025). The marked rise after 2022 aligns with increased AI tool use in research writing.

The data reveals distinct patterns of change across the three domains. The AI domain (Fig. 1) shows the most consistent trend of a slow but steady rise in complexity (decreasing readability) from 2015 to 2022, followed by a clear and sharp acceleration in all three metrics from 2023 onward. In stark contrast, the Medicine (Fig. 2) and Business (Fig. 3) domains show no clear linear trend from 2015 to 2023, exhibiting high year-to-year volatility instead. However, all three domains appear to converge on a similar pattern in the final two years, showing a significant increase in complexity from 2023 to 2025.

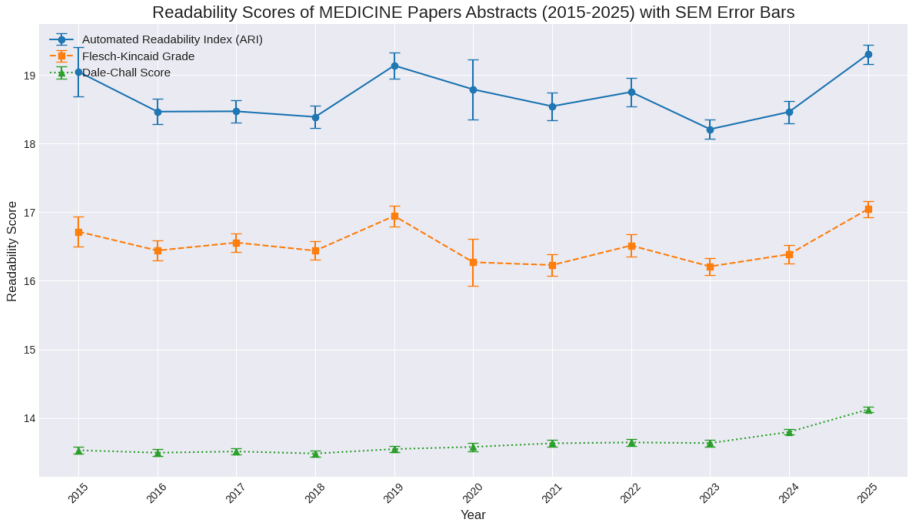

Fig. 2. Readability trends in **Medicine** (2015–2025). The scores show significant year-to-year volatility, with notable peaks in 2019 and 2025.

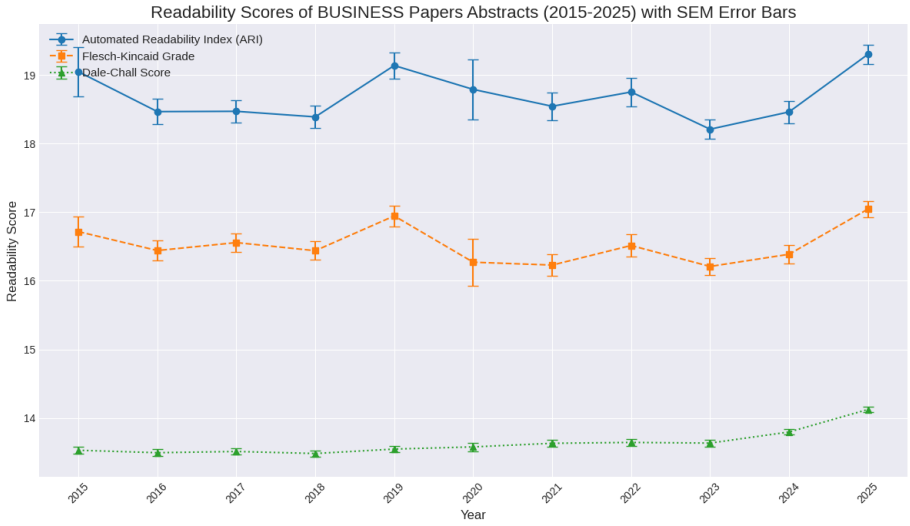

Fig. 3. Readability trends in **Business** (2015–2025). Scores show high volatility, mirroring the trends in Medicine, with no consistent trend until a final increase in 2024-2025.

4 Discussion

The interpretation of our results will focus on whether the data supports our working hypothesis.

4.1 Interpretation of Findings

The results of our analysis provide support for our central hypotheses. Our first hypothesis, which predicted a significant inflection point towards decreased readability after 2022, confirming preprint trends [6], is strongly supported for the corpus of AI studies, and is noticeable as a plausible trend for Literature on Medicine and Business, as visualized in Figures 1, 2, and 3, all three domains: AI, Medicine, and Business—demonstrate a sharp, convergent increase in linguistic complexity in the 2023–2025 period.

Furthermore, the data offers support for our second hypothesis regarding the nature of this change. The pattern is not uniform across disciplines. The AI domain (Fig. 1) stands in stark contrast to the other two. From 2015 to 2022, it exhibits a slow but consistent rise in complexity, establishing a clear baseline. After 2022, this trend sharply accelerates, visually confirming the inflection point we predicted for this early adopter field. Conversely, the Medicine (Fig. 2) and Business (Fig. 3) domains show year-to-year volatility from 2015 to 2023.

Overall, the data indicate that the widespread adoption of AI writing tools correlates with a decline in readability, as evidenced by a measurable increase in linguistic complexity across the scientific literature. We can also see that authors who write about AI appear to be early adopters of the techniques their subject offers.

4.2 Limitations and Outlook

This study's conclusions must be tempered by several key limitations. It can only show a correlation between the time period and readability, and not directly comment on AI adoption. We cannot prove that individual authors used AI, nor can we rule out other confounding variables (e.g., changes in editorial guidelines).

Furthermore, our analysis focuses exclusively on abstracts, as they are more widely available and can be obtained with open source tools. Our findings may not generalize to the full text of articles, where writing style and complexity can differ significantly.

The readability formulas are objective but can be said to be blind for important factors that influence readability. They do not measure coherence, factual accuracy, or logical flow. They only count word and sentence length, thus a text could be readable but factually nonsensical. It would be interesting to see whether AI tools themselves can be used to judge the readability of texts. Possible prompts would be, "How readable is the following text for an average scientist?" However, as these AI tools are developing rapidly, it would be difficult to get a coherent measure over the years. What can be done, however, is to apply a particular current AI implementation to texts from a sequence of years in the past.

Our analysis in this paper is primarily based on the visual inspection of plotted trends. A key limitation is the absence of formal statistical modeling. We did not perform regression analysis to confirm the statistical significance of

the observed trends or to test for a definitive inflection point after 2022. This statistical validation remains a critical step for future work.

The techniques presented in this article can be applied to other text types as well. For example, it is to be expected that AI tools will also be increasingly used for textbooks in social studies, humanities, and the sciences. Such texts should value readability for the intended target group even more than abstracts of scientific papers. It would be detrimental if the use of AI tools led to a decline of readability for such texts.

5 Conclusion

This study presents an initial large-scale, data-driven assessment of how AI-assisted writing tools may be influencing the readability of peer-reviewed scientific literature. By analyzing 33,000 abstracts published between 2015 and 2025 across the three domains of Artificial Intelligence, Medicine, and Business, we identified a general decline in readability beginning around 2023, with the most pronounced and earliest effect in AI-related publications. These findings suggest that the growing integration of LLM-based tools coincides with denser, more complex linguistic structures rather than improved textual accessibility.

While our results demonstrate a strong correlation, they cannot confirm causality, and our analysis is based on the visual inspection of trends. This formal statistical validation remains the most critical step for future work. Subsequent research should also extend the analysis to additional disciplines and employ advanced NLP methods to evaluate aspects beyond surface-level readability, such as coherence, style, and semantic clarity.

Ultimately, understanding how AI reshapes the language of science is essential for ensuring that technological progress does not come at the expense of clear communication.

References

1. Thorp, H.H.: ChatGPT is fun, but not an author. Science **379**(6630), 313 (2023)
2. Gao, C.A., et al.: Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. npj Digital Medicine **6**(1), 75 (2023)
3. Hwang, T., et al.: Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. PLOS ONE **19**(2), e0297701 (2024)
4. Plavén-Sigray, P., Matheson, G.J., Schiffler, B.C., Thompson, W.H.: The readability of scientific texts is decreasing over time. eLife **6**, e27725 (2017)
5. Howard, F.M., et al.: Characterizing the Increase in Artificial Intelligence Content Detection in Oncology Scientific Abstracts From 2021 to 2023. JCO Clinical Cancer Informatics **8**, e2400077 (2024)
6. Alsudais, A.: Exploring the change in scientific readability following the release of ChatGPT. Journal of Informetrics **18**(3), 101538 (2024)

7. Kincaid, J.P., Fishburne, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. Research Branch Report 8-75, Chief of Naval Technical Training, Naval Air Station Memphis (1975)
8. Senter, R.J., Smith, E.A.: Automated Readability Index. AMRL-TR-66-220, Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base (1967)
9. Dale, E., Chall, J.S.: A formula for predicting readability. Educational Research Bulletin **27**(1), 11–20 (1948)
10. Priem, J., Piwowar, H., Orr, R.: OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. In: 26th International Conference on Science, Technology and Innovation Indicators (STI 2022) (2022)