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}01@estudiant.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 - Computer Science, 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 Computer Science 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 similar upward spike in 2023–2025. These findings suggest that the integration of AI tools correlates with more complex sentence structures and denser vocabulary, potentially hindering accessibility. The data also suggests that Computer Science researchers likely adopted these tools earlier than scientists in other fields.

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 urgent 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 more "vague" or "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 topic 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 an acceleration of this trend. 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, and 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 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 data set covering an 11-year period (2015-2025). This data set comprises 1,000 abstracts per year for each of three distinct domains: AI, Medicine, and Business. Second, we apply three established readability metrics to these 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 data set, we used the open-source tool OpenAlex API [10] to collect data over 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 data sets compiled in a CSV file. These were works tagged with "Computer Science", "Medicine", and "Business".

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 OpenAlex abstract data were provided in the abstract_inverted_index format, so we first used a Python script to reconstruct the original, ordered plain-text 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 trend analysis, individual abstract scores were aggregated to calculate the mean readability score for each domain per year.

To quantify the rate of change and identify structural breaks, we went beyond visual inspection and performed a Segmented Linear Regression. We defined 2023 as the structural break point coinciding with the widespread availability of ChatGPT, and calculated the slope (m) for the pre-LLM (2015-2023) and post-LLM (2023-2025) periods. Additionally, we calculated the Readability Acceleration (Δm) to empirically verify the inflection point.

2.3 Rationale for readability metrics

Standardized measures for readability have been developed since the 1930s, 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 that 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.* It 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 systems become more coherent*). Such factors for complexity will not be captured by this scale. 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 of 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 treatment*, 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; this scale 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 Automated Readability Index 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 data sets 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: Computer Science, Medicine, and Business. To rigor-

ously identify the impact of Large Language Models (LLMs), we performed a three-step analysis: (1) Visual inspection of yearly trends, (2) Segmented Linear Regression to quantify pre- and post-adoption gradients, and (3) Acceleration analysis to identify the precise year of maximum inflection.

3.1 Readability trends (2015–2025)

Figures 1, 2 and 3 display the mean yearly score for the ARI, FKGL, and Dale–Chall formulas.

The data reveals distinct patterns of change. The Computer Science domain shows a consistent, gradual rise in complexity (decreasing readability) from 2015 to 2022, followed by a sharp acceleration in all three metrics from 2023 onward. In contrast, Medicine and Business show high year-to-year volatility between 2015 and 2022 with no clear linear direction. However, all three domains converge on a similar pattern in the final epoch: a significant, synchronized increase in linguistic complexity from 2023 to 2025.

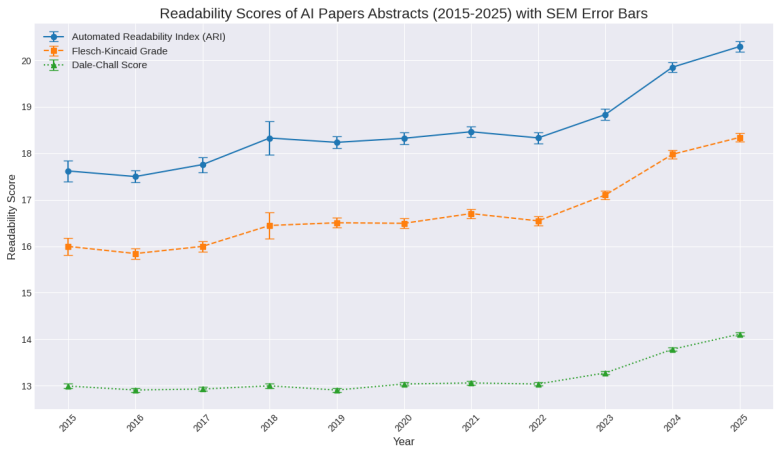

Fig. 1. Readability trends in **Computer Science** (2015–2025). The baseline trend of increasing complexity steepens dramatically after 2022.

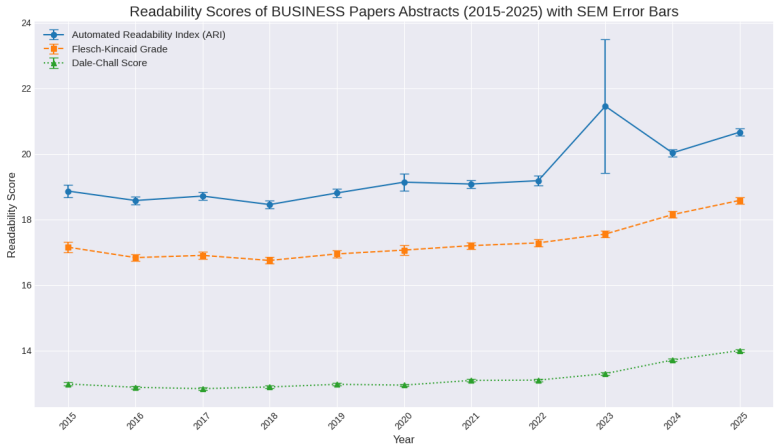

Fig. 2. Readability trends in Business (2015–2025). Similar to Medicine, the domain exhibits a "hockey stick" inflection post-2022.

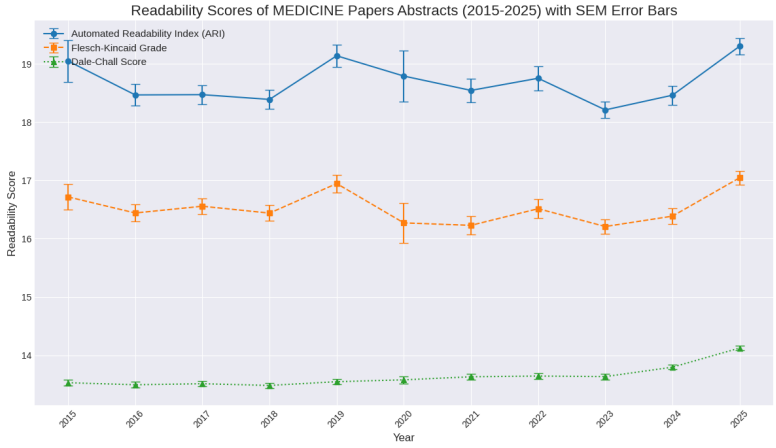

Fig. 3. Readability trends in Medicine (2015–2025). Volatility characterizes the pre-2022 era, followed by a unified upward spike.

3.2 Gradient analysis: structural breaks

To quantify the rate of change, we performed a segmented linear regression setting 2023 as the structural break point (coinciding with the widespread availability of ChatGPT). We calculated the slope (m) for the Pre-LLM period (2015–2023) and the Post-LLM period (2023–2025).

Figures 4, 5 and 6 illustrate these regression lines. In all three domains, the solid red line (Post-LLM) exhibits a significantly steeper positive gradient than the dashed blue line (Pre-LLM).

In Computer Science, the slope for the Automated Readability Index (ARI) jumps from a moderate steady growth ($m \approx 0.04$) to a rapid ascent ($m \approx 0.9$). In Business, the contrast is even more stark; the pre-2023 gradient was effectively negligible (indicating stable readability), while the post-2023 gradient indicates a rapid surge in text complexity.

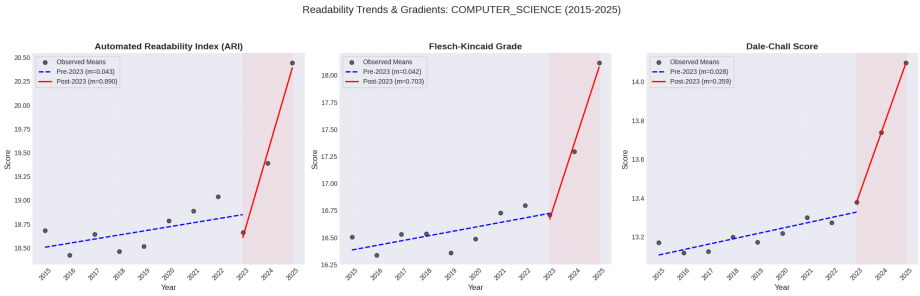

Fig. 4. Gradient analysis for Computer Science. The blue dashed line represents the pre-LLM trend; the red solid line represents the post-LLM trend. The steepening slope indicates a structural shift in writing style.

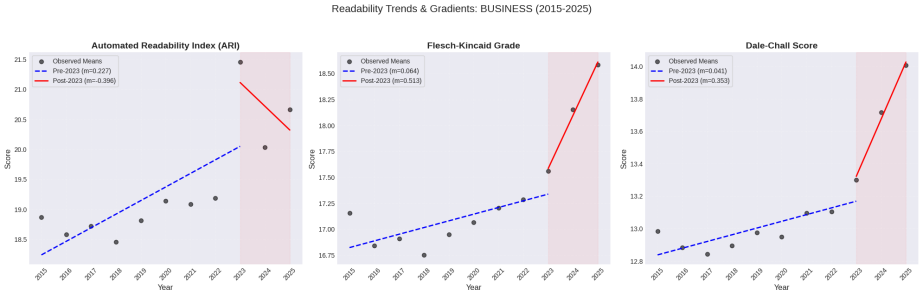

Fig. 5. Gradient analysis for Business. Note the shift from a flat trend ($m \approx 0$) to a strong positive incline after the cutoff.

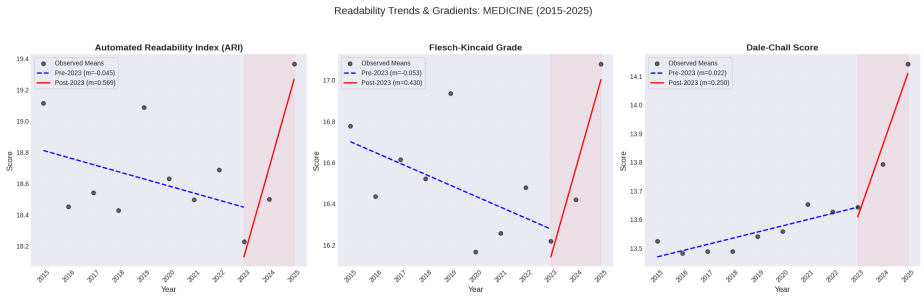

Fig. 6. Gradient analysis for Medicine. Despite earlier volatility, the post-2023 trend indicates a decisive move toward higher complexity scores.

3.3 Acceleration analysis: identifying the inflection point

To empirically verify that 2022–2023 was indeed the inflection point, we calculated the "Readability Acceleration" for a rolling window of cutoff years. Acceleration is defined as the difference between the post-cutoff slope and the pre-cutoff slope :

$$\Delta m = m_{post} - m_{pre}$$

Figures 7, 8 and 9 plot the acceleration magnitude. A peak in this graph represents the year where the trajectory of scientific writing changed most drastically.

Computer Science shows a distinct peak inflection at 2023. The acceleration is positive across all metrics, confirming that the rate of complexity growth increased significantly.

On the other hand, Business exhibits a mixed signal. While the Flesch-Kincaid and Dale-Chall metrics show distinct positive acceleration (peaking at $\sim$0.4 and $\sim$0.3 respectively), confirming a steepening trend, the ARI metric presents a sharp negative outlier in 2023. This divergence suggests that while vocabulary and sentence length increased, character-count density behaved differently in this domain.

Finally, Medicine mirrors the behavior of Computer science, with the acceleration metric peaking sharply around the 2022–2023 window.

The alignment of these peaks across disparate fields provides robust evidence that an external factor—likely the introduction of generative AI tools—precipitated a simultaneous shift in academic writing styles globally.

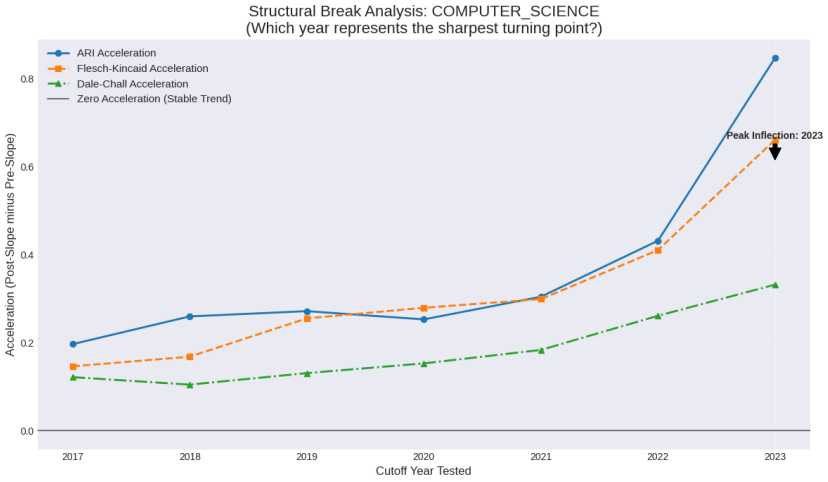

Fig. 7. Acceleration analysis for Computer Science. The peak at 2023 indicates the point of maximum change in the trend.

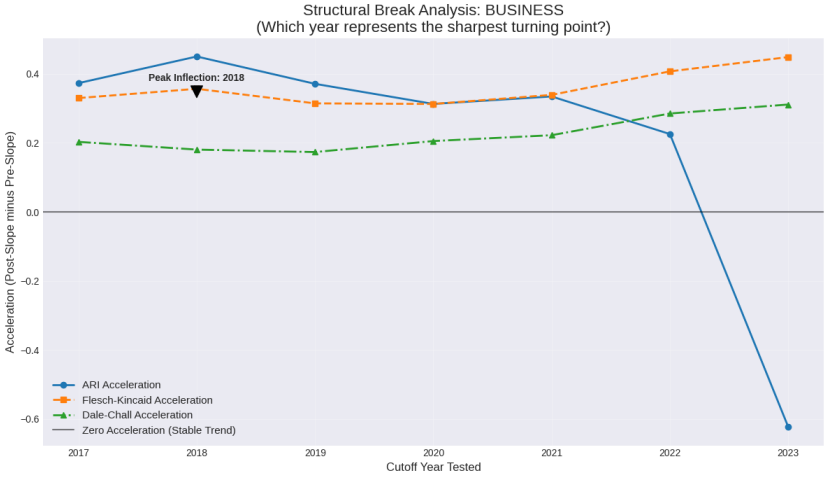

Fig. 8. Acceleration analysis for Business. Flesch-Kincaid and Dale-Chall show positive acceleration, consistent with other domains, while ARI diverges in the final year.

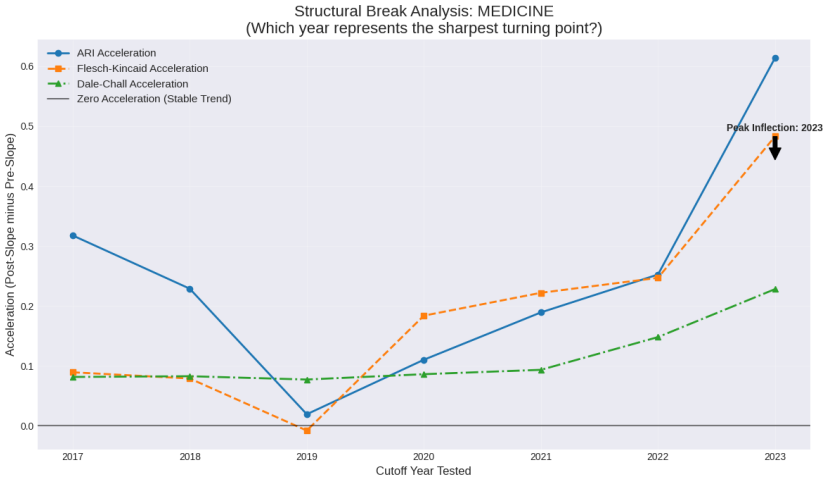

Fig. 9. Acceleration analysis for Medicine. The inflection point aligns with the other domains, suggesting a cross-disciplinary phenomenon.

4 Discussion

4.1 Interpretation of findings

The results support both of our initial hypotheses. The first hypothesis anticipated a marked shift in readability after 2022, in line with earlier observations on preprints [6]. This pattern is visible in the trend plots for all three domains (Figures 1, 3, 2), each of which shows an upswing in complexity and hence decrease in readability beginning in the 2023–2025 period. The increase is most stable and pronounced in the Computer Science corpus.

The hypothesis that this change would appear earliest or most strongly in the AI domain is also confirmed by the data. The Computer Science trends from 2015 to 2022 show a steady, gradual rise in complexity before the sharp post-2022 acceleration. Medicine and Business, in contrast, fluctuate considerably year to year during this earlier interval. These field-specific differences are quantified in our gradient analysis (Figures 4, 6, 5), where the post-2023 slopes (m_{post}) steepen most sharply for Computer Science.

The acceleration analysis further confirms 2023 as the definitive inflection point. While Business showed a mixed acceleration signal due to the ARI outlier, the Computer Science and Medicine domains exhibited clear, synchronized peaks in acceleration around 2022–2023. Taken together, these observations point to a shift toward more complex writing in the period following 2022, driven by the early adopter effect in AI-related fields.

4.2 Limitations and future outlook

This study's method must acknowledge several key limitations. First, the study establishes a correlation between the time period and readability but cannot verify the specific use of AI tools by individual authors. We cannot rule out other variables, such as changes in editorial guidelines or field-specific changes over time.

Furthermore, our analysis focuses exclusively on abstracts. While abstracts represent a standardized unit of analysis, findings may not generalize to the full text of articles, where writing style and complexity can differ significantly.

Moreover, the readability formulas (FKGL, ARI, Dale-Chall) are objective but blind to semantic quality. They only count word and sentence length. Future work could use LLMs themselves to judge readability by using prompts such as "How readable is this text for a general scientist?". Although, this may introduce new challenges regarding the consistency of the evaluator model over time.

The techniques presented in this article can be applied to other text types as well. Future studies could apply them to textbooks in various areas of study, where readability should be valued greatly, and a decline of readability due to AI usage would have a significant negative impact.

5 Conclusion

This study presents an initial 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 supported by regression analysis, they cannot confirm causality. Subsequent research could extend the analysis to additional disciplines and employ NLP methods to evaluate aspects beyond blind readability metrics.

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)