

The Marker Drift Problem: Measuring LLM-Assisted Writing with Unstable Lexical Indicators

Serhii Nazarovets¹

¹ Borys Grinchenko Kyiv Metropolitan University, 18/2 Bulvarno-Kudriavska Str., 04053 Kyiv, Ukraine.

<http://orcid.org/0000-0002-5067-4498>

Correspondence: serhii.nazarovets@gmail.com; s.nazarovets@kubg.edu.ua

Abstract

Lexical markers are increasingly used to estimate the prevalence of large language model (LLM)-assisted writing across large publication corpora. However, longitudinal applications of this approach rely on an important assumption: that the relationship between these markers and LLM use remains sufficiently stable over time. Recent research suggests that once publicly identified, individual markers may lose their diagnostic value due to adaptation by users and the evolution of language models. Furthermore, lexical markers may reflect not only LLM use, but also linguistic competence, AI-assisted editing, and broader forms of stylistic convergence. This paper introduces the Marker Drift Problem to describe a methodological challenge in which indicators change alongside the phenomenon they are intended to measure. Marker drift raises questions about temporal comparability, as observed changes in marker prevalence may reflect changes in LLM use, changes in the indicators themselves, or both. Without attention to the temporal stability and construct validity of lexical indicators, increasingly precise estimates of LLM-assisted writing may be based on unstable measurement instruments.

Keywords: large language models; scientific writing; lexical markers; marker drift; text detection

The rapid adoption of large language models (LLMs) in academic writing has created a parallel methodological challenge: how can their use be measured at scale? One increasingly influential approach relies on lexical markers – words or expressions whose frequency increased following the widespread adoption of generative AI. Such markers offer a scalable method for estimating LLM-associated writing across large publication corpora. However, their use for longitudinal measurement relies on an important assumption: the indicators themselves must remain sufficiently stable over time.

The problem is particularly important for longitudinal studies. If the diagnostic relationship between a lexical marker and LLM use changes over time, estimates derived from the same marker set may not be directly comparable across periods. An observed decline in a marker may reflect reduced LLM use, but it may equally reflect user adaptation, model evolution, or the diffusion of previously LLM-associated language into human writing. Conversely, the emergence of new markers may create the appearance of changing prevalence even when the underlying level of LLM assistance remains stable. Marker drift therefore raises a

problem of temporal comparability: the measurement instrument is not necessarily invariant across the period being measured.

McCreery and Naser (2026) employ the approach proposed by Kobak et al. (2025), according to which the prevalence of large language models in academic writing can be assessed through the analysis of characteristic lexical markers. The relationship between marker density and citation rates in engineering publications, which constitutes the central finding of the article, is derived from this approach.

However, this approach relies on an important assumption: the markers must remain sufficiently stable to be used for assessing the prevalence of LLM-associated writing. This assumption warrants further discussion.

One of the most interesting findings of the article itself is the observation that certain words, previously regarded as hallmarks of LLM usage, begin to lose popularity once their diagnostic role becomes the subject of public debate. A similar conclusion was reached by Geng and Trotta (2025), who demonstrated a decline in the frequency of words such as *'delve'*, *'intricate'*, *'pivotal'* and *'realm'* after they became widely recognised as markers of ChatGPT-like writing.

This observation is significant not only for AI-text detection. The set of markers changes under the influence of user reactions and the evolution of the language models themselves. Longitudinal comparisons of successive ChatGPT versions likewise illustrate how LLM outputs and capabilities can change substantially over relatively short periods (Nazarovets & Teixeira da Silva, 2026). As a result, the indicators used for measurement also evolve. Under such conditions, estimates of LLM prevalence may reflect not only the scale of language model usage, but also the relevance of a specific set of lexical indicators at a given time.

This development is consistent with the findings of AI detection research. A number of studies indicate that any successful detection strategy creates incentives for further adaptation (Kehkashan et al., 2025). As soon as certain linguistic features begin to be used as markers of LLM texts, users have reason to avoid such constructions, and new generations of language models gradually move away from patterns that have become easily recognisable.

In this context, marker drift appears not as a random anomaly, but as a natural consequence of the interaction between detection tools and the object being detected. As noted by Majovský et al. (2024), the development of text generation and detection systems is taking on the characteristics of a technological race, in which each new generation of detectors stimulates the emergence of new evasion strategies. At the same time, Liu, Li and Li (2025) demonstrate that the effectiveness of many indicators depends on the specific model, domain, and degree of human intervention in the text.

The challenge extends beyond the disappearance of individual markers. The indicator set itself is continuously evolving, while the boundary between human and LLM-associated writing is becoming increasingly blurred. Geng and Poibeu (2025) point out that editing,

paraphrasing and other hybrid forms of interaction between the author and the language model are making the interpretation of linguistic features increasingly difficult.

This leads to a more fundamental question: what exactly do lexical markers measure? Liang et al. (2023) demonstrated that popular AI text detectors systematically classify texts by non-English-speaking authors as AI-generated more often, suggesting a possible influence of linguistic competence on such assessments. A similar argument is put forward by Hashemi, Shi and Corriveau (2025), who highlight the distinction between fully generated texts and those that have merely been edited or linguistically enhanced using an LLM. Under these conditions, identical markers may reflect fundamentally different forms of interaction between a human and a language model.

Thus, markers may signal not only the use of LLMs, but also the author's linguistic competence, editorial intervention, stylistic imitation, or the broader influence of LLMs on academic writing. This poses a methodological problem for approaches that use lexical markers as proxies for assessing the prevalence of LLM-associated writing (Kobak et al., 2025). More broadly, LLMs are increasingly used for specific scholarly tasks beyond text generation, including research evaluation and post-publication review triage (Nazarovets & Teixeira da Silva, 2026), further complicating attempts to infer the nature or extent of LLM involvement from textual features alone.

This methodological challenge is referred to here as the *Marker Drift Problem*, a situation in which the measurement tool evolves alongside the phenomenon it is designed to measure. Under such conditions, estimates of LLM prevalence may reflect not only changes in the use of language models, but also changes in the properties of the indicators themselves. In this regard, a promising avenue appears to be the search for more robust indicators, in particular approaches that rely not on individual words or phrases, but on deeper statistical characteristics of the text (Wang et al., 2026).

A complementary approach is to reduce reliance on indirect inference by encouraging transparent, voluntary disclosure of GenAI use. Structured frameworks such as the Generative AI Delegation Taxonomy (GAIDeT) allow authors to report specific tasks delegated to GenAI, rather than requiring researchers to infer such use from unstable linguistic traces (Suchikova et al., 2026). Although self-reported disclosure cannot capture undisclosed use and therefore does not replace observational methods, it can provide a valuable complementary source of evidence for studying how GenAI is used in scholarly work.

McCreery and Naser (2026) convincingly demonstrate the potential of lexical markers for studying LLM-associated writing. At the same time, their work draws attention to a broader issue: the indicators used to assess the prevalence of LLM-assisted writing may themselves change under the influence of the evolution of language models and writing practices.

The *Marker Drift Problem* describes a situation where a measurement tool evolves alongside the phenomenon it is designed to measure. Under such conditions, estimates of LLM

prevalence may reflect not only changes in the use of language models, but also changes in the properties of the indicators themselves.

For the further development of this field, understanding how long these markers remain informative and exactly what they measure is no less important than the search for new markers. Without due attention to the temporal stability and construct validity of lexical indicators, increasingly precise estimates of LLM-assisted writing may be based on unstable measurement instruments.

Conflicts of interest

The author declares no conflicts of interest.

Funding

None declared.

References

Geng, M., & Poibeau, T. (2025). On the Detectability of LLM-Generated Text: What Exactly Is LLM-Generated Text? <http://arxiv.org/abs/2510.20810>

Geng, M., & Trotta, R. (2025). Human-LLM Coevolution: Evidence from Academic Writing. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 12689–12696). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-acl.657>

Hashemi, A., Shi, W., & Corriveau, J.-P. (2025). AI-generated or AI touch-up? Identifying AI contribution in text data. *International Journal of Data Science and Analytics*, 20(4), 3759–3770. <https://doi.org/10.1007/s41060-024-00693-9>

Kehkashan, T., Riaz, R. A., Al-Shamayleh, A. S., Akhunzada, A., Ali, N., Hamza, M., & Akbar, F. (2025). AI-generated text detection: A comprehensive review of methods, datasets, and applications. *Computer Science Review*, 58, 100793. <https://doi.org/10.1016/j.cosrev.2025.100793>

Kobak, D., González-Márquez, R., Horvát, E.-Á., & Lause, J. (2025). Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27). <https://doi.org/10.1126/sciadv.adt3813>

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 100779. <https://doi.org/10.1016/j.patter.2023.100779>

Liu, X., Li, Y., & Li, K. (2025). Enhancing the Robustness of AI-Generated Text Detectors: A Survey. *Mathematics*, 13(13), 2145. <https://doi.org/10.3390/math13132145>

McCreery, Z., & Naser, M. Z. (2026). LLM-Assisted writing in engineering is associated with higher citations as evidenced from 1.17 million papers. *Journal of Informetrics*, 20(3), 101842. <https://doi.org/10.1016/j.joi.2026.101842>

Májovský, M., Černý, M., Netuka, D., & Mikolov, T. (2024). Perfect detection of computer-generated text faces fundamental challenges. *Cell Reports Physical Science*, 5(1), 101769. <https://doi.org/10.1016/j.xcrp.2023.101769>

Nazarovets, S., & Teixeira da Silva, J. A. (2026). Can Generative AI Assist Post-publication Peer Review? A Pilot Study on LLM-assisted Triage. *Journal of Academic Ethics*, 24(3), 86. <https://doi.org/10.1007/s10805-026-09761-0>

Nazarovets, S., & Teixeira da Silva, J. A. (2026). Ethical opportunities and risks of using ChatGPT for open science: evidence from a three-year mixed-methods study. *Journal of Ethics in Entrepreneurship and Technology*. <https://doi.org/10.1108/JEET-09-2025-0059>

Suchikova, Y., Tsybuliak, N., Teixeira da Silva, J. A., & Nazarovets, S. (2026). GAIDeT (Generative AI Delegation Taxonomy): A taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. *Accountability in Research*, 33(3). <https://doi.org/10.1080/08989621.2025.2544331>

Wang, Z., Yao, A., & Ren, M. (2026). AI-generated text detection via multilevel Zipf's law. *Journal of Informetrics*, 20(2), 101799. <https://doi.org/10.1016/j.joi.2026.101799>