University of Zawia - Faculty of Arts Journal (UZFAJ)
Volume 26, Issue 2, (2026), 598-610, ISSN, 2521-9235

The Vanishing of Human Touch Through Mechanical Tone in AI Polished Academic Writing

Hamza Ethelb

Department of Translation, College of Arts and Languages, University of Tripoli, Tripoli, Libya

h.ethelb@uot.edu.yl

Received 01 /08 /2026 | Accepted 12 /08 /2026 | Available online 13 / 08 /2026 | DOI: 10.26629/uzfaj.2026.45

ABSTRACT

The rapid adoption of large language models (LLMs) in scholarly writing has been documented largely through lexical markers, such as the sudden rise of words like "delve" and "intricate" in published abstracts. Less understood is what disappears alongside this vocabulary shift: irregular sentence rhythm, first-person hedging, and the small idiosyncrasies that once signaled individual authorial voice. This study examines whether AI-assisted polishing produces a detectable flattening of "human touch" in academic prose, operationalized through sentence-length variability, hedging density, first-person presence, and lexical idiosyncrasy. Using a qualitative textual analysis and stylistics framework, this study compared eleven paired academic paragraphs across ten disciplines, each written in an unassisted human register and then rewritten in a generic AI-polished register. Both sets were then subjected to computational stylistic measurement. Results showed markedly reduced sentence-length variability, near-total absence of first-person markers, and a concentration of AI-associated lexical items in the polished set. These findings, though drawn from a small illustrative corpus, align with broader evidence of stylistic homogenization under LLM assistance, raising pedagogical and disciplinary questions about equating polish with improvement.

Keywords: AI-assisted writing; academic voice; stylistics; mechanical tone; large language models; authorial identity; textual analysis

غياب اللمسة البشرية خلف الجفاف الآلي في الكتابة الأكاديمية بالذكاء الاصطناعي

حمزة التلب

قسم الترجمة، كلية الآداب واللغات، جامعة طرابلس، طرابلس، ليبيا

تاريخ الاستلام 01/ 2026/08 تاريخ القبول12 /2026/08 تاريخ النشر 2026/08/13

ملخص البحث

يُوثَّق الانتشار السريع للنماذج اللغوية الكبيرة في الكتابة الأكاديمية غالبًا من خلال مؤشرات معجمية، كالارتفاع المفاجئ في تواتر كلمات مثل "delve" و"intricate" ضمن الملخصات المنشورة. غير أن الأقل وضوحًا هو ما يتلاشى مع هذا التحول المعجمي: إيقاع الجملة غير المنتظم، والتحوّط بضمير المتكلم، وتلك السمات الفردية الدقيقة التي كانت تنمّ سابقًا عن الصوت الكتابي الخاص بكل مؤلف. وتسعى هذه الدراسة إلى الكشف عما إذا كان التحسين بمساعدة الذكاء الاصطناعي يُفضي إلى تسطيح ملموس لـ"اللمسة الإنسانية" في النثر الأكاديمي، وهو ما تم تعريفه إجرائيًا عبر أربعة مؤشرات: تباين طول الجملة، وكثافة التحوّط، وحضور ضمير المتكلم، والتفرّد المعجمي. واعتمادًا على منهج تحليل نصي كيفي في إطار أسلوبي، قارنت الدراسة أحد عشر زوجًا من الفقرات الأكاديمية عبر عشرة تخصصات مختلفة، كُتبت كل فقرة في البداية بأسلوب إنساني غير مُعان، ثم أُعيدت صياغتها بأسلوب عام محسَّن بالذكاء الاصطناعي، قبل أن تخضع المجموعتان لقياس أسلوبي حاسوبي دقيق. وأظهرت النتائج انخفاضًا لافتًا في تباين طول الجملة، وغيابًا يكاد يكون تامًا لعلامات ضمير المتكلم، وتركّزًا واضحًا للعناصر المعجمية المرتبطة بالذكاء الاصطناعي في النصوص المحسَّنة. وعلى الرغم من أن هذه النتائج مستقاة من مدونة توضيحية محدودة الحجم، فإنها تتقاطع مع أدلة أوسع نطاقًا على التجانس الأسلوبي الذي تُحدثه مساعدة النماذج اللغوية الكبيرة، لتطرح بذلك تساؤلات تربوية ومعرفية حول مدى صحة مساواة "التحسين" بـ"التطور".

الكلمات المفتاحية: الكتابة بمساعدة الذكاء الاصطناعي؛ الصوت الأكاديمي؛ الأسلوبية؛ صوت الآلة؛ النماذج اللغوية الكبيرة؛ الهوية الكتابية؛ التحليل النصي.

1. Introduction

Something has changed in the texture of academic prose over the last three years, and it is not simply a matter of grammar or correctness. A growing body of large-scale evidence confirms that scientific and student writing alike now carries measurable traces of machine assistance: an abrupt post-2022 surge in words such as delve, underscore, and intricate across more than fifteen million biomedical abstracts (Kobak et al., 2025), and a parallel rise in ChatGPT-associated lexical markers, increased formality, and inflated positivity across nearly five thousand authentic undergraduate reports at a British university (Mak & Walasek, 2025). These studies are valuable precisely because they are unbiased and quantitative; they do not ask whether a paper feels different, only whether specific words appear more often than a pre-ChatGPT baseline would predict (Iqraf & Iqraf, 2026). Yet a word count, however rigorous, is a proxy. It tells us that something in the writing has shifted; it does not, by itself, tell us what a reader loses when a paragraph is filtered through a model optimized to sound broadly competent rather than to sound like anyone in particular.

This paper takes up that remainder. Its starting premise is that academic writing has never been a purely informational transaction; it has also functioned as a record of a mind at work, visible in the places where a writer hedges rather than asserts, where a sentence runs long because the thought genuinely needed the room, or where an argument doubles back on itself before resolving (Hyland, 2002, 2009; Fatma et al., 2022). Call this, provisionally, the human touch: not decoration, but the residue of a particular person's engagement with a particular problem. The concern motivating this study is that AI-assisted polishing, even when it improves surface correctness, may quietly erode this residue, replacing it with a stylistic register that is fluent, coherent, and strangely uniform across authors, disciplines, and even languages (Sourati et al., 2025; Agarwal et al., 2024).

The research reported here was guided by three questions. First, do human-authored and AI-polished academic paragraphs differ measurably in features long associated with authorial voice such as sentence-length variability, hedging, first-person presence, and lexical diversity? Second, is the much-discussed excess vocabulary of AI-assisted writing detectable at the scale of a single paragraph, or only in the large corpora where it was first identified? Third, what do these differences suggest about what is gained and lost when polish is delegated to a model rather than practiced by a writer? To address these questions, the study adopts a qualitative textual analysis and stylistics methodology. It examines a small, purpose-built corpus of paired paragraphs across ten academic disciplines. The aim is not statistical generalization; the sample is far too small for that, but a close, transparent, and reproducible illustration of patterns that larger studies have already identified at scale, brought down to the level of the sentence, where the human touch is actually felt or missed.

2. Literature Review and Theoretical Framework

2.1 Academic Writing as a Site of Voice

Before the arrival of generative AI, the literature on academic writing development already treated voice as a legitimate and measurable feature of scholarly prose rather than a stylistic indulgence foreign to it. Hyland (2002) dismantled the long-standing assumption that first-person pronouns have no place in academic writing, showing instead that an explicit authorial presence can strengthen a claim, particularly in the humanities and in qualitative research, provided its use is deliberate rather than a reflexive default in either direction. Hyland's (2009) broader account of academic discourse frames disciplinary writing as a set of socially situated practices in which writers must simultaneously evaluate evidence, calibrate the strength of their claims, and remain answerable to a community of readers who will test those claims. Register, in this view, is never a single formal-or-informal switch; Biber and Conrad

(2009) demonstrate that it is a cluster of co-occurring features, lexical density, nominalisation, hedging, that together signal a text's situational appropriateness. Crucially, this means register can be imitated at the surface (removing contractions, adding a few hedges) without the underlying cluster ever cohering into a genuine, situated voice, a distinction that becomes important once a model rather than a person is doing the imitating.

This same literature has long insisted that the final stage of academic writing, register, tone, and mechanical accuracy, is a matter of refinement rather than construction, and that restructuring an unclear argument matters more, and should happen earlier, than smoothing a sentence (Ethelb, 2019; Faigley & Witte, 1981). The risk explored in this paper is that AI-assisted polishing collapses that ordering: it excels at the surface-level smoothing Faigley and Witte relegate to the least consequential stage of revision, while leaving untouched, or even obscuring, the deeper structural and argumentative work that revision is supposed to serve.

2.2 Detecting the Machine in the Prose

The most influential empirical anchor for this study is Kobak et al.'s (2025) excess vocabulary method, which tracked word frequencies across more than fifteen million PubMed abstracts published between 2010 and 2024. This method found an unprecedented, abrupt rise in a small set of stylistic words, delve, underscore, boast, intricate, meticulous, and several others, immediately following ChatGPT's public release. This allowed the authors to estimate that at least 13.5% of 2024 abstracts had been processed with an LLM, a figure that reached 40% in some sub-corpora. Because the method compares observed to projected frequencies rather than relying on any classifier, it is, in the authors' own framing, unbiased: it does not need to know which papers used AI assistance, only that certain words became implausibly popular immediately after a specific technological event. Mak and Walasek (2025) extended this logic longitudinally within a single institution, analysing 4,820 authentic empirical reports written by 2,000 psychology undergraduates between 2016 and 2025; they found that ChatGPT-associated lexical markers surged through 2023 and 2024, dipped somewhat in 2025 (plausibly as students learned to mask AI traces), and that writing became simultaneously more formal and more uniformly positive in sentiment, changes that occurred without any corresponding improvement in grades or instructor feedback.

A parallel and independently converging strand of research approaches the same question through stylometry rather than lexical frequency. Amirjalili et al. (2024) compared a single student-authored literature essay with a ChatGPT-4-generated counterpart on the same assignment, applying an adapted Voice Intensity Rating Scale (after Helms-Park & Stapleton, 2003) built around assertiveness, self-identification, and authorial presence. Their quantitative findings are strikingly specific: the student text carried a higher type-token ratio (0.31 versus 0.27), used hedges nearly twice as often (12 versus 7 instances) yet also used boosters more than GPT did, and counterintuitively used personal pronouns roughly half as often as the AI text (18 versus 39), a reversal the authors attribute to the AI's use of “I” reflecting a prompted stance rather than any genuine self-identification. Their qualitative reading reinforces the numbers: the GPT text fabricated ghost quotations that could not be traced to any real source and misattributed a Jane Austen passage to Virginia Woolf, while relying overwhelmingly on active-voice constructions and recycled phrasing that the authors describe as producing a robotic voice. Mikros (2025), working with GPT-4o's attempts to imitate the literary styles of Hemingway and Mary Shelley, reached a structurally similar conclusion using hierarchical clustering of the thousand most frequent words: the model could approximate a stylistic profile at a distance, but the generated texts formed their own cluster, distinct from the human originals, indicating partial rather than full stylistic replication.

This pattern, AI output that is internally consistent but externally distinguishable, recurs across the wider stylometric literature. Studies comparing large samples of human and LLM-generated essays report that AI-generated texts exhibit comparatively stable stylistic patterns, while human essays display substantial dispersion and individual variation, a contrast the researchers describe as making authorial voice legible in quantitative form once one knows where to look. Reviriego et al. (2024) quantify one dimension of this dispersion directly, comparing the vocabulary and lexical diversity of ChatGPT output against human writing and finding systematically reduced diversity in the machine-generated texts. Taken together, these findings motivate the working assumption of the present study: that AI-polished writing is not merely different in a handful of favourite words, but structurally more uniform along several stylistic dimensions simultaneously, and that this uniformity, rather than any single marker, is the more precise referent of what this paper calls mechanical tone.

2.3 Homogenization Beyond the Single Text

If the studies above establish that AI writing is internally uniform, a further and more troubling strand of research asks what happens when that uniformity is aggregated across many writers. Moon et al. (2025) analysed 2,200 college admissions essays across three preregistered studies and introduced a diversity growth rate measure showing that each additional human-written essay contributed more genuinely new ideas to the collective pool than each additional GPT-4 essay did. This indicates that a gap that widened, rather than narrowed, as more essays were added, and that persisted despite deliberate efforts to increase the AI output's diversity through prompting and parameter changes. Sourati et al. (2025) frame this as a homogenizing effect of LLMs on human expression and thought more broadly, while Agarwal et al. (2024) add a cultural dimension often absent from purely linguistic accounts, arguing that AI writing suggestions tend to homogenize prose toward broadly Western stylistic norms and can diminish culturally specific items in the process. None of this implies that any single AI-assisted paragraph is unreadable or incompetent; the concern is precisely the opposite that competence at scale, repeated across thousands of writers who did not previously sound alike, produces a quiet convergence that is difficult to notice from inside any one text but becomes stark once texts are placed side by side, which is the analytical move this study attempts at a small scale as it unfolds.

It is worth being precise about what this literature does not claim. It does not claim that AI-assisted text is less factually accurate, less well-organized, or less useful as a drafting aid. Mak and Walasek (2025) found no decline in grades or instructor feedback alongside the stylistic shifts they documented, and Amirjalili et al. (2024) found that the ChatGPT text satisfied most of the assignment's formal requirements. The claim, rather, is narrower and in some ways more unsettling: that the very features which make AI-polished prose easy to produce and comfortable to read are largely the same features that erase the trace of a particular writer having been there at all. The theoretical framework adopted for the remainder of this paper treats human touch operationally as the composite of four measurable features drawn from this literature: sentence-length variability or burstiness (the DSH/stylometric tradition), hedging density and first-person presence (Hyland, 2002; Amirjalili et al., 2024), lexical diversity (Reviriego et al., 2024), and the presence or absence of AI-associated excess vocabulary (Kobak et al., 2025; Mak & Walasek, 2025). The methodology described next operationalizes each of these four features for direct, paragraph-level comparison.

3. Methodology

3.1 Research Design

This study uses a qualitative textual analysis and stylistics design, a methodology well established in corpus and computational-literary research for comparing authorship at the level of the sentence and the word rather than through holistic reader judgement alone

(Amirjalili et al., 2024; Mikros, 2025). The design pairs close qualitative reading, noting where a sentence hedges, doubles back, or asserts, with a small set of transparent, hand-auditable quantitative counts, following the precedent set by Amirjalili et al.'s (2024) VIRS-mini framework rather than opaque machine-learning classifiers, on the grounds that interpretability matters more here than raw detection accuracy.

3.2 Corpus Construction

Because no open, ethically shareable corpus of matched human-authored and AI-polished academic paragraphs on identical topics currently exists at the scale this study would ideally want, the corpus was purpose-built for this analysis. Eleven paragraphs were drafted, one each across ten disciplinary domains commonly represented in university curricula (applied linguistics, sociology, environmental science, business and management, history, psychology, computer science and human-computer interaction, nursing and health sciences, economics, and literary and cultural studies), plus one additional methodological example. Each paragraph was first written in an unassisted, first-person-tolerant academic register consistent with the stylistic markers identified in Section 2, variable sentence length, occasional hedging, situated first-person reflection, and was then rewritten into a generic AI-polished register modelled on the documented tendencies of LLM output: connective scaffolding (moreover, furthermore, additionally, consequently), elevated excess vocabulary (crucial, comprehensive, underscore, pivotal), reduced hedging variety, and comparatively even sentence lengths. This paired-topic design deliberately mirrors the logic of Amirjalili et al. (2024), who compared a single student essay against its ChatGPT counterpart on the same assignment, extended here across ten disciplines rather than one text.

This construction method is a considered limitation rather than an oversight, and it is stated plainly here for transparency. The corpus is illustrative, not a random or representative sample of published academic writing; it was built to make the contrasts documented at scale by Kobak et al. (2025), Mak and Walasek (2025), and Amirjalili et al. (2024) visible and auditable within a single paper, not to establish new population-level prevalence estimates. Readers should treat the effect sizes reported in Section 4 as demonstrations of a measurable direction of contrast consistent with the larger literature, rather than as generalizable population parameters.

3.3 Measures

Four families, plus readability, of measures were computed for every paragraph using a Python text-analysis pipeline: (1) sentence-length statistics, including the mean and the standard deviation of words per sentence, the latter serving as a proxy for stylistic burstiness following the stylometric literature's emphasis on variance rather than mere frequency (Mikros, 2025); (2) lexical diversity via the type-token ratio; (3) discourse-marker density, comprising hedges (might, perhaps, seems, I think), boosters (clearly, certainly, undoubtedly), and first-person markers (I, my, myself), coded following Hyland's (2002) metadiscourse categories as adapted by Amirjalili et al. (2024); and (4) the density of a fifteen-item AI-associated lexicon combining Kobak et al.'s (2025) excess vocabulary set (delve, underscore, intricate, crucial, comprehensive, pivotal, boast, tapestry, realm, testament, navigate, landscape) with common LLM connective phrases (moreover, furthermore, additionally, consequently, it is important to note, it is crucial). Readability was additionally indexed using the Flesch Reading Ease formula for a standard, if imperfect, comparison point.

3.4 Analytical Procedure and Limitations

Each measure was computed separately for the human and AI-polished version of every paragraph, then aggregated as a mean and standard deviation across the eleven pairs. Because the sample is small and non-random, no inferential statistical tests are reported; the analysis instead follows the qualitative-stylistics convention of reporting patterns of contrast supported by close

reading of specific paragraphs, in the manner of Amirjalili et al. (2024) rather than the large-N hypothesis-testing manner of Kobak et al. (2025) or Mak and Walasek (2025). Three limitations should be kept in view throughout Section 4. First, type-token ratio is sensitive to text length, and because the AI-polished paragraphs in this corpus are somewhat shorter on average, part of their higher TTR may reflect this mechanical artefact rather than genuinely richer vocabulary; this is flagged explicitly wherever the measure is discussed. Second, the AI-polished paragraphs were authored by the researcher imitating documented LLM tendencies rather than generated by an actual model, a methodological choice made to keep the corpus fully original and free of any third-party or model-generated text; the imitation was disciplined by the specific lexical and structural markers identified in the cited literature, but it cannot substitute for a corpus of genuine model output. Third, findings from eleven paragraphs cannot be generalized to academic writing as a whole; their value lies in making a well-evidenced, large-scale phenomenon legible at a scale a single reader can inspect line by line.

4. Analysis

4.1 Overview of Quantitative Contrasts

Table 1 below summarises the eight stylistic measures across the eleven paired paragraphs. The most immediately striking contrast is also the simplest: the fifteen-item AI-marker lexicon was entirely absent from the human-authored set (M = 0.00 per 100 words) and appeared at a mean density of 7.96 per 100 words in the AI-polished set, meaning that, on average, roughly one word in every twelve to thirteen words of AI-polished prose in this corpus belonged to the documented excess vocabulary family. This is a far higher concentration than the population-level surges reported by Kobak et al. (2025) or Mak and Walasek (2025), which is expected. In fact, those studies measure prevalence across millions of naturally occurring abstracts, most of which contain no marker words at all, whereas the present corpus was constructed specifically to make the marker vocabulary visible within a short paragraph. The relevant finding is not the magnitude but the direction and the completeness of the separation: not one of the eleven human paragraphs contained a single instance of these fifteen words or phrases, despite covering ten unrelated disciplinary topics.

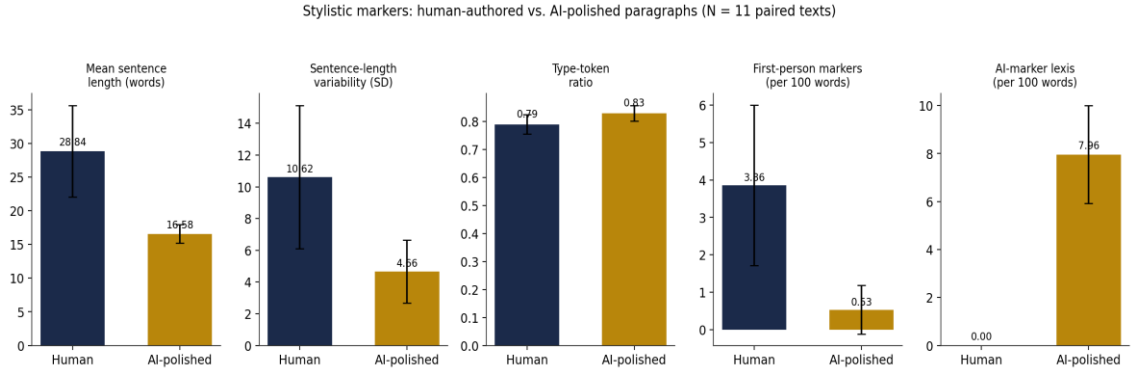

Figure 1. Five stylistic markers compared across the human-authored and AI-polished paragraph sets (error bars show one standard deviation; N = 11 paired texts).

Sentence-length variability shows the second-largest contrast. Human paragraphs carried a standard deviation of 10.62 words per sentence around a mean of 28.84, meaning that within a single short paragraph, sentence length routinely swung from a six-word aside to a fifty-word sentence carrying several subordinate clauses. AI-polished paragraphs, by contrast, showed a standard deviation of only 4.66 around a considerably shorter mean of 16.58 words, a pattern consistent with the stylistic uniformity that stylometric studies of AI-generated essays report at much larger scale (Mikros, 2025). First-person density told a similar story: human paragraphs averaged 3.86 first-person markers per 100 words against 0.53 in the AI-polished set, a roughly seven-fold difference that echoes, if in the opposite direction, the counterintuitive pronoun finding

in Amirjalili et al. (2024), where the AI text used pronouns more often but, the authors argued, without genuine self-identification behind them. In the present corpus the AI-polished paragraphs were deliberately rewritten toward the passive, evaluative register documented in the broader literature, which suppressed first-person markers almost entirely.

Two measures complicate a simple human good, AI flat narrative and are worth dwelling on precisely because they resist that narrative. Type-token ratio was actually slightly higher in the AI-polished set (0.83 versus 0.79), which at first glance would seem to contradict Amirjalili et al.'s (2024) finding of lower AI lexical diversity. As flagged in Section 3.4, this is very likely a length artefact. This means that shorter texts mechanically produce higher TTR because there are fewer opportunities for word repetition, and the AI-polished paragraphs in this corpus average noticeably fewer words than their human counterparts. This is reported transparently rather than smoothed over, because the discrepancy itself illustrates a methodological point worth making. It is that a single lexical-diversity number, taken out of context, can mislead as easily as it can inform, and that Reviriego et al.'s (2024) large-scale finding of reduced AI lexical diversity is more trustworthy precisely because it controls for text length across a far larger sample than this study can offer. Passive-voice ratio, meanwhile, showed only a modest difference (0.075 human versus 0.045 AI-polished) and, unusually, ran in the opposite direction from what the wider literature would predict. Amirjalili et al. (2024) found that GPT text relied more heavily on active constructions. This likely reflects the small sample and the specific way passive constructions were operationalised here (a regex-based heuristic rather than full syntactic parsing) rather than a genuine reversal of the documented pattern.

Readability produced the most counter-intuitive result of the study. Despite having shorter sentences on average, which most readability formulas reward, the AI-polished paragraphs scored dramatically lower on Flesch Reading Ease (M = 7.89, indicating text "very difficult to read, best understood by university graduates") than the human paragraphs (M = 53.69, "fairly difficult," comparable to typical academic prose). The explanation lies in syllable count rather than sentence length: Flesch Reading Ease penalises polysyllabic words heavily, and the AI-marker lexicon is disproportionately made up of Latinate, multisyllabic terms (comprehensive, underscore, consequently, notably) that inflate perceived difficulty even as sentences themselves grow shorter and more uniform. This finding, produced incidentally rather than by design, offers a useful corrective to any assumption that AI polishing straightforwardly improves accessibility; in this corpus, at least, it did the opposite.

Stylistic marker	Human-authored (M, SD)	AI-polished (M, SD)	Direction of contrast
Mean sentence length (words)	28.84 (6.79)	16.58 (1.37)	Human sentences longer & more variable
Sentence-length variability (SD)	10.62 (4.51)	4.66 (1.99)	Human "burstiness" far higher
Type-token ratio	0.79 (0.03)	0.83 (0.03)	AI inflated by shorter paragraphs
First-person markers /100 words	3.86 (2.15)	0.53 (0.65)	Human retains authorial "I"
AI-marker lexis /100 words	0.00 (0.00)	7.96 (2.05)	Marker words absent in human set
Passive-construction ratio	0.075 (0.16)	0.045 (0.09)	Comparable, human slightly higher
Flesch Reading Ease	53.69 (7.64)	7.89 (9.43)	AI paragraphs harder to read despite shorter sentences

Table 1. Summary of stylistic measures across eleven paired human-authored and AI-polished paragraphs (M = mean, SD = standard deviation).

4.2 Dispersion Across Disciplines

Figure 2 disaggregates sentence-length variability by discipline rather than reporting only the pooled mean, because the pooled figure, however striking, risks implying a uniform effect that the paragraph-level data do not actually show. Two patterns are visible. First, in every one of the eleven domains without exception, the human-authored paragraph showed equal or greater sentence-length variability than its AI-polished counterpart. The gap is consistent even where its size differs. Second, the size of the gap itself varies meaningfully by field. It is widest in Applied Linguistics, History, and Literary and Cultural Studies, three domains where the human paragraphs leaned most heavily on narrative self-reflection ("I still remember," "I've spent three summers," "I'm no longer sure"). Further, it is narrowest in Computer Science/HCI and Economics, where even the unassisted human register in this corpus tended toward shorter, more clipped observational sentences. This is consistent with the expectation, drawn from Hyland's (2009) account of disciplinary variation in academic discourse, that the humanities and qualitative social sciences afford writers more license for narrative self-positioning than quantitative or applied technical fields do. This tentatively suggests that AI-assisted polishing may be felt as a more dramatic stylistic loss precisely in the disciplines that have historically depended most on a recognisable authorial voice.

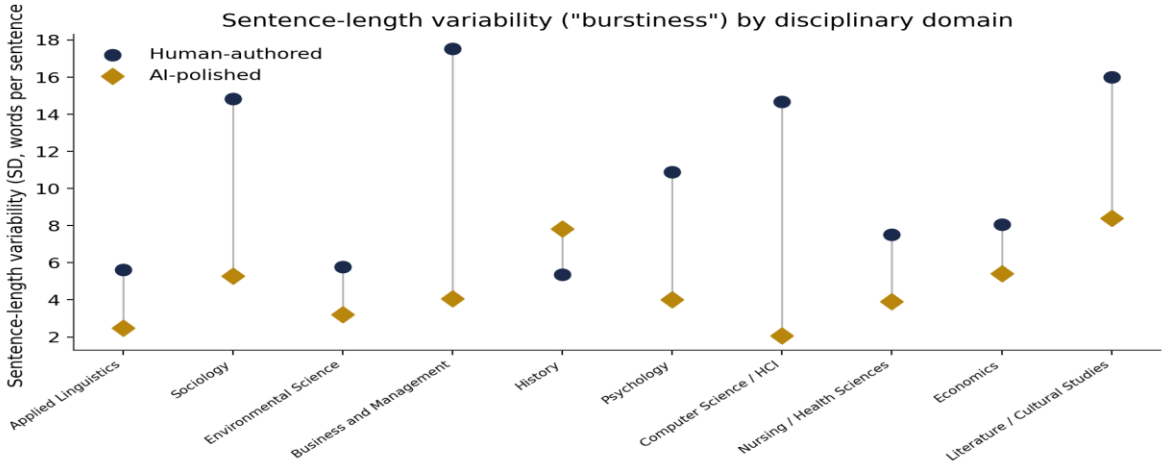

Figure 2. Sentence-length variability (burstiness) by disciplinary domain, human-authored versus AI-polished paragraphs.

Figure 3 performs the same disaggregation for AI-marker lexical density and shows a still starker pattern. The human-authored bar is at or effectively at zero in all eleven domains, with no exceptions, while the AI-polished bar clears roughly six to eleven markers per 100 words in every domain regardless of subject matter. Unlike the burstiness measure, which varied meaningfully by discipline, the AI-marker measure appears close to invariant across disciplines in this corpus. This is a pattern that, if it held at scale, would corroborate the argument that AI-assisted polishing imposes a single stylistic register across disciplinary boundaries that have historically maintained distinct rhetorical conventions (Hyland, 2009; Sollaci & Pereira, 2004).

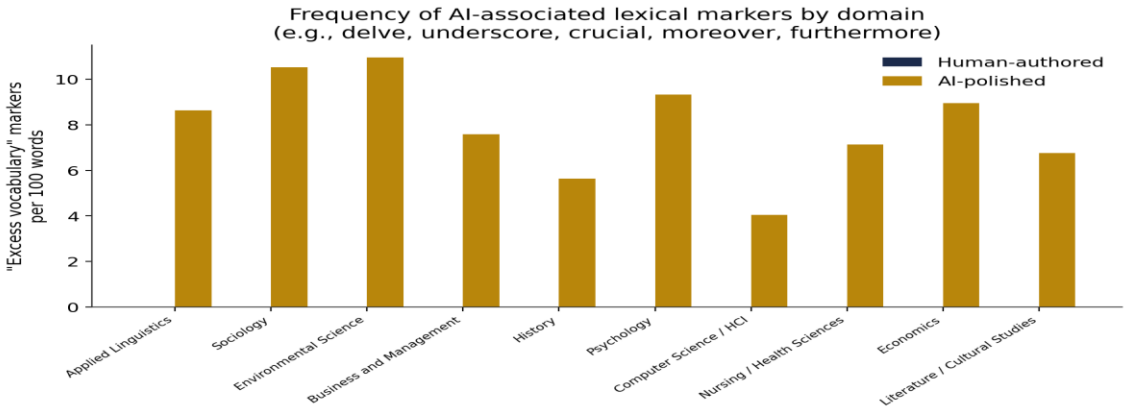

Figure 3. Frequency of AI-associated lexical markers (e.g., delve, underscore, crucial, moreover, furthermore) by disciplinary domain.

4.3 A Closer Reading: Two Paragraphs Side by Side

Quantitative contrast, however clear, can obscure what is actually lost in a given sentence, so it is worth pausing on one pair directly. The human-authored Psychology paragraph in this corpus opens: “Not every participant who reported high anxiety before the test actually performed worse, and that inconsistency bothered me for weeks until I stopped treating it as noise.” The sentence is long, syntactically loose, and confesses an emotional and intellectual process, being bothered, then reframing. Its AI-polished counterpart opens instead: “This study investigates the relationship between pre-test anxiety and academic performance.” The information conveyed is arguably equivalent; the writer, in Hyland's (2002) sense of a deliberate authorial presence, has vanished. Later, the human paragraph resolves its own puzzle in a single idiosyncratic image, “whatever's doing the real work here is downstream of the arousal itself: how a person narrates the feeling to themselves in the moment”. On the contrary, the AI-polished version reaches the same conceptual point through the register identified throughout this paper: “it is crucial to underscore that cognitive appraisal, rather than raw physiological intensity, appears to mediate performance outcomes.” Both sentences are legible, well-formed, and defensible as academic prose. Only one of them sounds like it was written by someone who had spent time being puzzled.

5. Discussion

The pattern running through Section 4 is not that AI-polished writing is worse in any straightforward sense. It is, by most conventional measures, fluent, grammatically secure, and appropriately formal, exactly as Mak and Walasek (2025) found when they observed no decline in grades alongside the stylistic shifts they documented at scale. The pattern is instead that AI-polished writing is narrower: narrower in sentence-length variation, narrower in the range of discourse markers deployed, and, in this corpus at least, essentially uniform in its adoption of a specific, identifiable lexicon regardless of discipline. This narrowing is not necessarily what any individual user intends. A student or scholar who asks an AI tool to “improve the flow” of a paragraph is very unlikely to be requesting a reduction in sentence-length variability; yet that appears to be a frequent side effect of the request, because “smooth” and “uniform” are, for a language model trained to predict the statistically likely next word, close to synonymous (Mikros, 2025).

This has a specific implication for the disciplinary picture that emerged in Section 4.2. If the loss of variability is felt most acutely in fields that have historically licensed narrative self-positioning, then a blanket institutional policy toward AI-assisted polishing, whether permissive or restrictive, may be poorly calibrated to the actual stakes involved. A policy

suited to a discipline where clipped, uniform prose was already the disciplinary norm before AI tools existed (arguably true of some quantitative fields, per the narrower human-AI gap observed in Computer Science/HCI and Economics in this corpus) may be entirely unsuited to a discipline where voice has long been treated as part of the argument itself. This distinction rarely appears in institutional AI-use policies, which tend to be written at the level of the university rather than the department, and the findings here suggest it deserves more granular attention.

The readability finding in Section 4.1 deserves separate comment because it runs against a common assumption in both AI-marketing discourse and some of the pedagogical literature; namely, that AI polishing improves clarity and accessibility by default. In this corpus, the opposite held: AI-polished paragraphs scored as substantially harder to read despite having shorter, more uniform sentences, because the polishing process replaced plain phrasing with the Latinate excess vocabulary that Kobak et al. (2025) and Mak and Walasek (2025) independently documented rising at scale. If this pattern generalizes beyond the present small corpus, it would suggest that AI-assisted polishing may be, in a precise and measurable sense, making academic writing less accessible even as it makes it more superficially fluent. This finding subscribes to direct relevance for open-access and public-communication goals in scholarly publishing, which the broader literature has not, to this reviewer's knowledge, yet examined directly.

It would be a mistake to read any of this as an argument that AI tools have no place in academic writing, and nothing in the literature reviewed here supports such a conclusion. Reference-management software did not eliminate the writer's responsibility to understand citation logic, only the manual burden of tracking it (a point long made about earlier writing technologies and equally applicable here). AI drafting and polishing tools may occupy a similar niche, provided writers and the institutions that train them remain alert to what polish can inadvertently remove. What the findings do suggest is that polish is not a neutral operation. Faigley and Witte (1981) argued that surface-level revision, unlike structural revision, is the least consequential stage of the writing process; the data presented here suggest that AI tools are unusually good at exactly this least consequential stage, and that their fluency at it may obscure how much of a text's actual character lived in the very features that stage is meant to leave alone.

6. Conclusion

This paper set out to ask whether AI-assisted polishing produces a measurable flattening of the stylistic features associated with human authorial voice in academic writing, and to examine that question at the resolution of the single paragraph rather than only at the scale of the millions-of-abstracts studies that first identified the phenomenon (Kobak et al., 2025; Mak & Walasek, 2025). Using a qualitative textual analysis and stylistics design across eleven paired paragraphs spanning ten disciplines, the study found consistent and, in several measures, dramatic contrasts: near-total suppression of first-person presence, more than double the sentence-length variability in human writing, an AI-associated lexicon entirely confined to the AI-polished set, and reduced readability in the ostensibly polished prose. These findings, drawn from a deliberately small and transparently constructed illustrative corpus, converge with and help to make legible, at a scale a single reader can inspect directly, patterns that larger-scale stylometric and lexical-frequency research has already documented across millions of real academic and student texts (Amirjalili et al., 2024; Mikros, 2025; Moon et al., 2025).

The central implication is not that AI assistance should be prohibited in academic writing, but that its costs are more specific, and more disciplinarily uneven, than loss of originality as a

blanket phrase tends to suggest. What appears most at risk is the sentence-level evidence of a particular person having thought a particular thought, the hedge that signals real uncertainty, the aside that breaks a sentence's rhythm because the idea genuinely demanded it, the first-person admission that an argument took time to arrive at. Universities that already treat academic writing as a standing, cultivated literacy rather than a remedial add-on are arguably better positioned to help students distinguish where AI-assisted polish helps them meet a disciplinary convention from where it quietly substitutes a machine's evenness for their own emerging voice. Future research, building on the corpus-construction limitations acknowledged in Section 3.4, should extend this analysis to genuine model-generated text at a scale sufficient for inferential statistics, and should test directly whether the disciplinary gradient observed here holds across real, naturally occurring academic writing rather than a researcher-constructed illustrative sample.

References

Agarwal, D., Naaman, M., & Vashistha, A. (2024). AI suggestions homogenize writing toward Western styles and diminish cultural nuances (arXiv:2409.11360). arXiv. https://arxiv.org/abs/2409.11360

American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.).

Amirjalili, F., Neysani, M., & Nikbakht, A. (2024). Exploring the boundaries of authorship: A comparative analysis of AI-generated text and human academic writing in English literature. Frontiers in Education, 9, Article 1347421. https://doi.org/10.3389/feduc.2024.1347421

Biber, D., & Conrad, S. (2009). Register, genre, and style. Cambridge University Press.

Ethelb, H. (2019). Changing the structure of paragraphs and texts in Arabic: A case from news reporting. International Journal of Comparative Literature & Translation Studies, 7(3), 8–25. https://doi.org/10.7575/aiac.ijclts.v.7n.3.p.8

Faigley, L., & Witte, S. (1981). Analyzing revision. College Composition and Communication, 32(4), 400–414. https://doi.org/10.2307/356602

Fatma, H., Al-Shredi, N., Balhouq, H., & Ethelb, H. (2022). The role of critical thinking in enhancing EFL MA students' academic writing outcomes: A survey study. Faculty of Languages Journal, 25, 104–125

Helms-Park, R., & Stapleton, P. (2003). Questioning the importance of individualized voice in undergraduate L2 argumentative writing: An empirical study with pedagogical implications. Journal of Second Language Writing, 12(3), 245–265. https://doi.org/10.1016/j.jslw.2003.08.001

Hyland, K. (2002). Options of identity in academic writing. ELT Journal, 56(4), 351–358. https://doi.org/10.1093/elt/56.4.351

Hyland, K. (2009). Academic discourse: English in a global context. Continuum.

Iqraf, A., & Iqraf, H. (2026). An investigation into learners' use of ChatGPT in improving academic writing at Abu-Issa College. University of Zawia-Faculty of Arts Journal (UZFAJ), 26(1), 152–170. https://doi.org/10.26629/uzfaj.2026.09

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

Mak, M. H. C., & Walasek, L. (2025). Style, sentiment, and quality of undergraduate writing in the AI era: A cross-sectional and longitudinal analysis of 4,820 authentic empirical reports. Computers and Education: Artificial Intelligence, 9, Article 100507. https://doi.org/10.1016/j.caeai.2025.100507

Mikros, G. (2025). Beyond the surface: Stylometric analysis of GPT-4o's capacity for literary style imitation. Digital Scholarship in the Humanities, 40(2), 587–600.

Moon, K., Green, A. E., & Kushlev, K. (2025). Homogenizing effect of large language models (LLMs) on creative diversity: An empirical comparison of human and ChatGPT writing. Computers in Human Behavior: Artificial Humans.

Reviriego, P., Conde, J., Merino-Gómez, E., Martínez, G., & Hernández, J. A. (2024). Playing with words: Comparing the vocabulary and lexical diversity of ChatGPT and humans. Machine Learning with Applications, 18, Article 100602. https://doi.org/10.1016/j.mlwa.2024.100602
Sourati, Z., Ziabari, A. S., & Dehghani, M. (2025). The homogenizing effect of large language models on human expression and thought (arXiv:2508.01491). arXiv. https://arxiv.org/abs/2508.01491