ASCILITE 2026

***The Future is Connection*:**

Collaboration, Creativity and Compassion in a Digital Society

Sophistication without clarity: Student writing versus unedited GenAI delegation

Zixuan Guo
University of Melbourne

Kamila Misiejuk
FernUniversität in Hagen, Germany

Sonsoles López-Pernas, Mohammed Saqr
University of Eastern Finland

Guanliang Chen
Monash University

Eduardo Oliveira
University of Melbourne

Writing is central to how students develop and demonstrate communicative competence, yet that development is increasingly difficult to observe as GenAI enters the writing process. This paper establishes a pre-GenAI stylistic baseline of student academic writing and characterises how unedited GenAI responses to the same assessment prompts differ from it. Using stylometric and writing analytics techniques, we analysed over 118,000 words collected in 2021 from 18 higher education students in a software engineering subject. Study 1 used the seven students who completed all four tasks to establish a within-cohort reference: the size and direction of stylistic change in the same students' writing across two semesters and two rhetorical genres. Study 2 paired each of 59 authentic submissions with a length-matched, zero-shot response to the identical prompt from three large language models. Student writing moved toward shorter words, fewer rare terms and markedly higher readability. Unedited model outputs moved in the opposite direction on each measure, gaining lexical sophistication through a narrower working vocabulary at substantial cost to readability; the human–GenAI gap in readability exceeded the students' own shift between semesters.

Keywords: generative AI, academic writing, stylometry, writing analytics, AI in education

Introduction

The increasing integration of generative artificial intelligence (GenAI) tools into educational settings has introduced both challenges and opportunities in how students produce academic writing (J. Kim & Lee, 2023). GenAI tools offer support across various stages of the writing process, from brainstorming and rewording to editing and even generating entire texts. GenAI tools therefore raise questions not only about authorship and academic integrity, but about what happens to writing development itself when part of the writing process is handed over. Academic writing is not merely a medium for assessment; it is central to how students construct arguments, develop critical thinking, and demonstrate conceptual understanding. As such, the style, structure, and language of student texts offer valuable insight into students' development as learners. While much of the current literature focuses on the ideal of human-AI synergy and collaborative co-authorship, a pervasive reality in higher education is "naïve" GenAI misuse—where students delegate entire writing tasks to large language models (LLMs) with little to no editing (Rowland, 2023). What is less well understood is what delegation does to the writing itself. Writing development is ordinarily observed as a trajectory, a direction of travel across a course of study, but that trajectory is now hard to observe since student corpora collected from December

2022 cannot be assumed free of GenAI involvement. A corpus of student writing collected before GenAI is therefore a non-renewable reference point: it records what student academic writing looked like when it was, unambiguously, the students' own.

In this context, we draw on writing analytics to examine changes in students' written reports across two semesters, focusing on vocabulary use and stylistic variation as indicators of writing growth. This analysis was first conducted on a pre-GenAI dataset from 2021, allowing us to establish a baseline of natural development in student writing over time. We then generated GenAI-produced texts in response to original assignment prompts, without incorporating students' own writing. This methodology was deliberately chosen to establish a "zero-shot" AI baseline—capturing the stylistic signature of minimally prompted, unedited LLM generation. By simulating one end of the use spectrum, full delegation with minimal prompting and no editing, we obtain a reference profile against which both authentic student writing and, in future work, student-edited or scaffolded GenAI use can be compared. This research design not only serves our primary goal of isolating machine stylometrics but also strictly adheres to ethical constraints regarding the protection of students' original intellectual property from commercial LLMs. By comparing the longitudinal pre-GenAI dataset with these simulated GenAI outputs using stylometric features, we explore whether and how the stylistic trajectory of natural student growth contrasts with the default outputs of LLMs. While our approach draws from explainable AI and authorship attribution research, it is uniquely situated within writing analytics and pedagogy. This paper contributes to the growing literature on GenAI in education by offering an empirical lens on the stylistic signatures of unscaffolded GenAI-produced writing, offering educators a descriptive reference point, rather than a means of detecting GenAI use in individual texts, for recognising and rewarding the features toward which students' own writing develops.

Background literature

Stylometric analysis and writing style attribution

Stylometric analysis offers a way to describe writing style quantitatively and, therefore, to compare how different authors or different text-production processes shape a text in response to the same task. It involves the quantitative analysis of writing style using computational methods to detect patterns in word choice, syntax, punctuation, and other linguistic features (Pan et al., 2025). Stylometry has long been used in authorship attribution studies, under the assumption that every writer exhibits a consistent and recognisable idiolect (Laramée, 2018). Academic writing, as a specialised form of communication, is particularly well-suited for stylometric analysis due to its structural conventions and stylistic consistency. It is central to university education, enabling students to articulate arguments, reflect critically, and demonstrate disciplinary understanding (Oliveira et al., 2026). Because of this, stylometric metrics have been used both to track change in writers' academic voice over a course of study and to characterise how texts produced under different conditions differ.

Several studies have identified core stylistic features that are effective in capturing authorial voice. Das and Ismael (2015) found that features like word and sentence length, punctuation frequency, and lexical richness are strong indicators of writing individuality. Pan and colleagues (2025) further demonstrated the value of part-of-speech tags, function-word n-grams, and structural patterns in distinguishing between authors in long and short texts. More comprehensively, Neal et al. (2018) categorised stylistic features into six domains: lexical, syntactic, semantic, structural, domain-specific, and additional indicators, providing a framework for robust stylistic analysis. To evaluate the significance of these features, statistical methods like the t-test are commonly used to assess whether stylistic changes are statistically significant. Wang et al. (2014) used the t-test for feature selection in text classification, and Kim (2015) outlined its assumptions, including normality, variance homogeneity, and independence. These assumptions must be met for valid inference. When used appropriately, such tests have the potential to highlight whether changes in writing metrics are due to human development or AI interference.

A small body of work has compared LLM-generated and human academic text directly. Herbold et al. (2023), comparing ChatGPT-3.5 and ChatGPT-4 essays with student essays on the same argumentative prompts, found the model texts showed higher sentence complexity, more nominalisation and a more formal register, while students used more modal and epistemic markers. At corpus scale, Kobak et al. (2025) document a measurable inflation of LLM-associated vocabulary in biomedical abstracts from 2023 onward. What this literature lacks is a comparison anchored in authentic, pre-GenAI student assessment writing, paired to the same assignment prompts and length-matched, and set against the direction in which the same students' writing moved across a course. That is the comparison this paper provides. It is also important to distinguish this descriptive, corpus-level aim from detection: tools that attempt to classify individual texts as AI-generated perform unreliably (Elkhatat et al., 2023; Weber-Wulff et al., 2023), and nothing in the present design depends on such classification.

GenAI in academic writing: collaboration versus delegation

The notion of human-AI co-authorship in writing tasks forms part of a broader discussion on human-AI collaboration, which highlights the interactive partnership between humans and generative AI systems. This collaboration blends computational efficiency with human creativity, contextual awareness, and judgment to accomplish shared tasks (Lee et al., 2024). In educational settings, GenAI tools are increasingly used by students across multiple stages of the writing process, including planning, drafting, and revision. While this integration offers opportunities for ideation and scaffolding, it also blurs the boundaries of authorship, particularly when learners delegate entire tasks to AI rather than engaging in iterative co-writing.

While concerns persist around academic integrity, disclosure, and the accuracy of GenAI outputs, many scholars have argued that responsible use of these tools could enhance equity, particularly for students from non-English-speaking backgrounds. Instead of viewing AI solely through the lens of detection or misconduct, there is a growing shift in the literature toward understanding the nuances of student-AI interaction and its implications for writing quality and learning. For example, Stanford's CoAuthor project (HAI, 2023) has compiled large-scale interaction datasets to examine patterns in human-AI co-authorship. Similarly, Shibani et al. (2023) introduced the CoAuthorViz tool, which visualises writing behaviours during collaboration with GPT-3. However, while these initiatives capture the ideal of "human-in-the-loop" co-authorship, they do not address the prevalent reality of unscaffolded delegation—where students bypass the writing process entirely by submitting raw, unedited AI outputs. To effectively design pedagogical strategies and transparent assessment systems (Elkhatat et al., 2023), educators must first understand the baseline stylistic differences between natural student writing and unedited, zero-shot AI generation.

Despite these emerging efforts, much of the current research remains conceptual or descriptive, with limited empirical evidence on how GenAI affects writing development over time. Furthermore, the few studies that compare human- and AI-generated academic text directly are not anchored in authentic, pre-GenAI student assessment writing, paired to the same assignment prompts and length-matched, nor set against the direction in which the same students' writing moved across a course. Addressing this gap, the present study combines stylometric analysis with writing analytics to establish, first, a within-cohort reference for how far and in which direction the same students' writing shifts across an assessment sequence completed without GenAI, and second, to compare authentic student texts with unedited GenAI responses to the identical prompts, asking whether the human–GenAI difference is the same kind of shift, or a different one. Specifically, it investigates the following research questions (RQs):

RQ1: How does the stylistic profile of student academic writing shift across an assessment sequence completed without GenAI tools, and by how much?
RQ2: How do the stylistic characteristics of student-authored texts compare with those of unedited GenAI responses generated independently to the same prompts, and how does the direction and size of that difference relate to the shift observed in RQ1?

Methodology

Participants

Participants were recruited from the Master of Software Engineering (SWEN) program at the University of Melbourne in Australia in 2021. A total of 18 students voluntarily took part in the study after providing informed consent, in accordance with the university's human ethics protocol (Ethics ID: 24272). While all 18 students contributed to the general corpus, only seven students completed all four assessment tasks across both semesters. To ensure a strict within-subject comparison in Study 1, only these seven students (df=6) were maintained. The resulting corpus for Study 2 consisted of 59 authentic human-authored submissions (considering all text available from all participants), which were compared against 59 length-matched GenAI outputs. All inferential tests were conducted on per-document metric values (Study 1: per-student paired values; Study 2: per-document pairs). For descriptive purposes, Table 1 reports a combination of per-document means and corpus-level values (raw totals, and pooled ratios computed as aggregate divided by aggregate), as indicated in the table; all inferential statistics were computed on per-document values. Recruitment was conducted via email, and participation involved the use of written submissions already produced as part of their regular coursework. Demographic data indicated a predominantly male cohort (16 males, two females), with ages ranging from 23 to 25 years or older. Due to the small sample size and significant gender imbalance, evaluating gender-based stylistic differences was not statistically viable. Therefore, the analysis maintained a strict within-subject design, focusing on overall individual stylistic evolution.

Data collection

The primary dataset for this study was collected from students enrolled in the year-long unit Advanced Software Project. The unit is structured around large-team software development projects for real clients, using agile methodologies. Assessment comprises 10 tasks; however, for this study, we focused exclusively on the individual assessments, which were designed to evaluate communication and reflective skills. These included: (A1) Personal Objectives (10%) [average words: 150], (A2) Ethics Report (2%) [average words: 700], (A3) Professional Communication Report (13%) [average words: 2,300], and (A4) Individual Contribution Statement (15%) [average words: 750]. A1 was completed synchronously during a supervised session in Week 5, while the remaining submissions were completed asynchronously via the Learning Management System. These writing samples span both Semester 1 (A1 and A2) and Semester 2 (A3 and A4). Due to varying completion rates across the cohort, the final human-authored dataset comprised 59 individual documents. For Study 2, each of these 59 documents was paired with a corresponding AI-generated response using the identical prompt and a ±10% word-count constraint. For the paired analysis reported in Table 1, each student document was matched with the ChatGPT-4o-latest output.

Stylometric metrics

To investigate changes in students' academic writing styles, we extracted stylistic features across three levels of linguistic analysis: word, sentence, and structure. This multi-level categorisation aligns with prior work on stylometry (e.g., Lagutina et al., 2019; Neal et al., 2018) and allows for a detailed, granular view of how writing complexity and voice evolve over time or shift with AI involvement. Word-level metrics capture lexical diversity and sophistication. Key features include the type-token ratio (TTR), unique word count, average word length, rare word frequency, and syllabic complexity. We note that raw TTR and raw frequency counts are sensitive to text length, which varies across the assessment tasks in this study; interpretations of these metrics are therefore made with caution, and length-robust alternatives such as Measure of Textual Lexical Diversity (MTLD) are identified as a refinement for future work. Sentence-level metrics reflect syntactic variation and stylistic rhythm. We calculated average sentence length, sentence length variability, use of punctuation, and passive versus active voice distribution. Additionally, indicators such as average use of articles ("a," "an") and hyphenation patterns were included, based on prior findings in stylometric attribution studies (e.g., Khan et al., 2021). These low-frequency features are reported for completeness; no interpretive weight is placed on

them. Structure-level metrics assess the overall organisation and readability of texts. These include standard counts (word, sentence, paragraph), paragraph density, readability scores, and punctuation diversity. All stylometric features were extracted computationally through a scripted pipeline applied uniformly to every document. Each feature — including TTR, syllable and rare-word measures, and Flesch Reading Ease — was calculated deterministically after preprocessing (case normalisation, punctuation handling, stopword removal, and lemmatisation, applied selectively per metric). Because extraction was fully automated and rule-based, it involved no manual coding, and inter-rater reliability was therefore not applicable at this stage.

Design

This research was organised into two studies (Figure 1). Study 1 investigated natural writing development by analysing student-authored texts across two semesters, completed within the same year and as part of the same unit, prior to the existence of GenAI tools. Study 2 examined how GenAI-generated texts compare stylistically to student-authored writing. Rather than co-authoring based on student submissions, this study simulated student-like responses by prompting GenAI models (GPT-3.5-turbo-1106, GPT-4-turbo-2024-04-09, and ChatGPT-4o-latest) with the original assignment prompts to generate full texts. Since we did not have access to a private LLM instance at the time of this research, this approach enabled a controlled comparison of writing styles between human-written and AI-generated outputs while avoiding the use of actual student data in AI input, in line with ethical constraints. As illustrated in Figure 1, the research design maintains a strict separation: student submissions (SWEN dataset) serve only as the human comparator in both studies and were never supplied to any model, while Study 2 generates independent AI outputs based exclusively on assignment prompts and personas (detailed below), ensuring no student intellectual property was processed by commercial LLM providers.

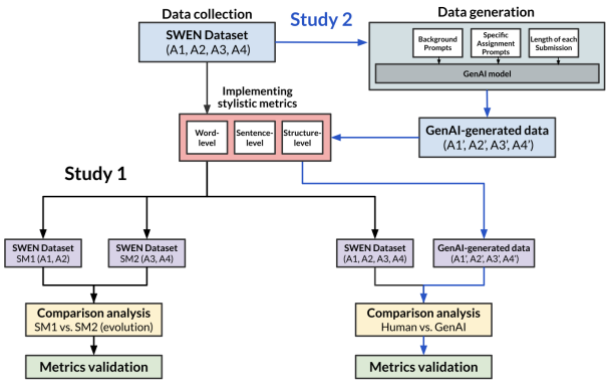


Figure 1. Research design

To investigate RQ1, we segmented the SWEN dataset into two subsets (Figure 1): Semester 1 (SM1), comprising Semester 1 submissions (A1 – Personal Objectives and A2 – Ethics Report), and Semester 2 (SM2), comprising Semester 2 submissions (A3 – Professional Communication Report and A4 – Individual Contribution Statement). SM1 and SM2 represent different rhetorical genres (reflective vs. professional reporting) in Study 1. Consequently, observed shifts in stylometric metrics reflect a combination of natural student maturation and the differing linguistic demands of the tasks. To maintain clarity, we refer to these shifts as “stylistic evolution” within the curriculum rather than isolated longitudinal development. We applied the full set of stylometric metrics (previously discussed) to both SM1 and SM2 datasets. Our goal was to quantify the size and direction of stylistic shift in the same students’ writing across the sequence, in the absence of GenAI, so that the human–GenAI differences in Study 2 could be read against a within-cohort reference rather than in isolation. To evaluate stylistic changes, we conducted paired-samples t-tests for both studies. Study 1 employed a paired within-subject design (n =7; df=6), while Study 2 used a paired-corpus design (n=59; df=58), matching each authentic student submission with a GenAI output generated from the identical prompt and length constraint. Normality of the difference scores was assessed using Shapiro-Wilk tests; metrics for which distributional assumptions were not satisfied were not tested inferentially and are marked "–" in Table 1. A paired design was prioritised to control for prompt and task demand, isolating stylistic differences attributable

to authorship (human vs. GenAI). We note one limitation: because several documents originate from the same students, the difference scores are not fully independent (Kim, 2015), which may understate standard errors; a mixed-effects model with a student-level random effect would address this and is a priority for future work with larger cohorts. These results establish a baseline for stylistic variation across the assessment sequence in the absence of GenAI, and a reference point for Study 2, where authentic student writing is contrasted with independently generated GenAI outputs.

To examine how GenAI ghostwriting differs from students' authentic academic writing (RQ2), we generated new responses to the original assignment prompts using GenAI tools, rather than expanding or modifying existing student submissions in Study 2. This approach aimed to simulate how students might use GenAI to delegate writing tasks entirely, thereby enabling a controlled comparison between human-authored and unedited AI-generated submissions (Figure 1). Three LLMs were accessed via the OpenAI API and integrated into a scripted workflow. These models were prompted to generate responses for each of the four assignments in the SWEN dataset (A1', A2', A3', A4'), replicating the tasks based solely on the prompt instructions and without access to students' original responses. To address the critical role of prompt construction, we adopted a "zero-shot persona" template. Models were not "primed" with student examples but were provided with a system message establishing a professional persona (e.g., "Act as a Master of Software Engineering student"). While this deviates from a "pure" default LLM output, it simulates a more ecologically valid scenario of "naïve" delegation where a student provides the prompt and context but no original writing. Default API parameters (e.g., temperature = 0.7) were maintained to capture the models' inherent stylistic tendencies rather than heavily fine-tuning the outputs. We deliberately did not optimise or iterate the prompts. More specific prompting produces more specific output, but the aim here was to profile the unedited default; prompt refinement and editing are themselves forms of student engagement that this baseline is intended to help study in future work. Two researchers with subject-matter expertise independently reviewed the generated texts to ensure quality and consistency. Their review confirmed that the responses met the assignment requirements and were appropriate in tone, structure, and content. To maintain comparability with original student submissions, we constrained the AI-generated outputs to remain within ±10% of the original word counts. Although the models were not directly editing student texts, they were instructed to simulate a student-like writing style, producing responses that reflected plausible student output when using GenAI support. To improve domain alignment, we also provided contextual information about the subject, course, and assessment to each model during generation. Corpus-level count metrics in Study 1 were likewise reported without inferential testing, as they are determined by the differing word-count requirements of the assignments; in Study 2, count metrics were tested because document lengths were matched by design.

Findings

Study 1: Stylistic shift across the assessment sequence (within-cohort reference)

The results from Study 1 indicate that students' writing style shifted markedly across the two semesters (Table 1), with the clearest evidence coming from the proportional metrics, which are not affected by the differing lengths of the assignments. Three word-level indicators of lexical accessibility all declined significantly from Semester 1 to Semester 2: average word length (SM1: 5.43 vs. SM2: 4.84; $p = .012$), syllables per word (1.82 vs. 1.61; $p = .006$), and the proportion of rare words (0.05 vs. 0.03; $p < .001$). By Semester 2, submissions favoured shorter, more familiar vocabulary, consistent with both growing student proficiency and the plainer register demanded by professional reporting genres. This lexical shift was accompanied by improved mechanical accuracy, with typo proportion falling significantly (0.03 vs. 0.02; $p < .001$), pointing to more careful proofreading and revision. Critically for RQ1, these changes translated into a significant gain in overall readability: Flesch Reading Ease rose from 45.05 to 59.94 ($p = .015$), a shift from the "difficult" band into "plain English" on conventional interpretations of the scale.

This evolution was selective rather than wholesale. Sentence-level features remained stable across semesters — average sentence length (20.56 vs. 22.47; $p = .63$), sentence length variability ($p = .62$), and lexical word

proportion (p = .52) showed no significant change — indicating that students did not simply shorten their sentences but chose more accessible vocabulary within a stable syntactic frame. Corpus-level counts (marked * in Table 1), including total word, sentence, paragraph, and unique word counts, all increased and are reported descriptively, as these are determined by the differing word-count requirements of the assignments. The decline in pooled Type-Token Ratio (0.31 vs. 0.14) is likewise not interpreted developmentally, as TTR decreases mechanically with text length. Taken together, the proportional metrics describe a coherent trajectory toward clarity across the assessment sequence: by the end of the academic year, and in the absence of GenAI tools, students produced writing that was lexically simpler, more accurate, and substantially more readable. We do not partition this shift between student development and the differing rhetorical demands of the tasks; with seven students and two genres that is not possible, and it is not required for the shift to serve its purpose here. What Study 1 establishes is the direction and magnitude of change that occurs within the same students' own writing across a year (for readability, +14.9 Flesch points; for average word length, −0.59 characters), which is the reference against which Study 2 is read.

Study 2: Simulated AI delegation (zero-shot baseline)

The initial phase of Study 2 assessed the ability of different generative AI models to replicate the length of original student submissions. GPT-3.5-turbo-1106 and GPT-4-turbo-2024-04-09 frequently fell short, producing outputs significantly below the expected word count. A manipulation check confirmed that length matching succeeded: per-document word counts did not differ between paired human and AI texts ($t(58) = -0.17$, $p = .87$), and corpus-level syllable and unique-word volumes were likewise comparable ($p = .25$ and $p = .31$, respectively). Observed stylistic differences therefore cannot be attributed to text length. Against this length-matched baseline, the AI-generated texts diverged from authentic student writing in a consistent direction. First, surface complexity was inflated across every indicator tested: AI texts exhibited longer words (Original: 4.88 vs. AI: 5.57; $p < .001$), more syllables per word (1.62 vs. 1.86; $p < .001$), a higher proportion of rare words (0.03 vs. 0.04; $p < .001$), and a higher density of content words relative to function words (lexical word proportion 0.61 vs. 0.64; $p < .001$).

Second, this inflated complexity was not matched by greater lexical variety: per-document Type-Token Ratio was significantly lower in the AI texts ($p < .001$), and unique word counts did not differ ($p = .31$), suggesting that the models achieved their sophisticated surface through repetition of a narrower vocabulary of complex terms rather than through genuine lexical breadth — sophistication without diversity. Third, and most notably, this lexical inflation came at a severe cost to readability: Flesch Reading Ease fell from 59.23 in the student texts to 36.28 in the AI texts ($p < .001$) — a drop from the plain English band typical of clear professional communication into the difficult band associated with dense academic prose. The AI texts were thus simultaneously more complex at the word level and substantially harder to read, the precise inverse of the developmental trajectory observed in Study 1. The magnitude of the divergence is as informative as its direction. The students' own writing moved +14.9 Flesch points between semesters; the human–GenAI gap on the identical prompts was −22.9 points. The GenAI texts (Flesch 36.3) therefore sit below even the Study 1 cohort's Semester 1 starting point (45.1): further from where the students finished than where they began, and in the direction their own writing moved away from.

Discussion

Across an assessment sequence completed without GenAI, students' own writing moved toward plainer vocabulary, fewer errors and markedly higher readability. Unedited GenAI responses to the same prompts moved the other way: longer and rarer words, a narrower working vocabulary, and readability in the difficult band. The two profiles are almost inverses on the metrics where the students changed most. However, Study 2 revealed that unedited use of GenAI diverged from this natural developmental path on the very metrics where students improved: where student writing moved toward shorter words, fewer rare terms, and greater readability, rising from the difficult band into plain English on the Flesch scale, default LLM outputs

ASCILITE 2026

The Future is Connection:

Collaboration, Creativity and Compassion in a Digital Society

Table 1. Stylometric metrics, t-test and p-value results from Study 1 and Study 2. Values shown are descriptive statistics, t-values, and p-values. Statistically significant ($p < 0.05$) results are shown in bold. Missing values (–) indicate that no inferential test is reported, either because distributional assumptions were not satisfied or, for corpus-level counts in Study 1, because these are determined by differing assignment length requirements.

		Study 1				**Study 2**		
Metric	**SWEN (SM1)**	**SWEN (SM2)**	**t-value**	**p-value**	**SWEN (SM1 and SM2)**	**AI-Authored (SM1 and SM2)**	**t-value**	**p-value**
Word-level				**Word-level**				
Type-Token Ratio*	0.31	0.14	-14.47	**<0.001**	0.0915	0.0907	3.37	**<0.001**
Unique Word Count*	1693	3214	-	-	5459	5320	1.02	0.31
Average Word Length	5.43	4.84	-3.56	0.012	4.88	5.57	9.61	**<0.001**
Rare Word Proportion*	0.05	0.03	-3.36	**<0.001**	0.03	0.04	5.73	**<0.001**
Rare Word Count*	256	647	-	-	1666	2497	-	-
Syllables Proportion*	1.82	1.61	-4.08	0.006	1.62	1.86	9.54	**<0.001**
Syllables Count*	9865	36359	-	-	96839	109157	1.17	0.25
Typo Proportion	0.03	0.02	-3.22	**<0.001**	0.01	0.01	-	-
Lexical Word Proportion	0.63	0.62	-0.68	0.52	0.61	0.64	5.62	**<0.001**
Sentence-level				**Sentence-level**				
Average Sentence Length	20.56	22.47	0.50	0.63	22.76	20.18	-	-
Sentence Length SD	18.99	14.48	-0.52	0.62	13.79	10.44	-	-
Punctuation per Sentence	3.75	2.50	-	-	2.94	3.60	-	-
Average Hyphens per Sentence	0.44	0.20	-2.53	0.04	0.20	0.26	-	-
Average 'a' per Sentence	0.36	0.46	0.36	0.73	0.51	0.37	-	-
Average 'an' per Sentence	0.05	0.06	0.87	0.42	0.07	0.05	-2.13	0.037
Passive Voice Proportion	0.21	0.14	-2.35	0.06	0.16	0.09	-	-
Active Voice Proportion	0.79	0.86	2.35	0.06	0.84	0.91	-	-
Sentence Complexity	1.09	0.88	-	-	0.87	1.01	-	-
Structure-level				**Structure-level**				
Readability (Flesch Reading Ease)	45.05	59.94	3.38	0.015	59.23	36.28	-6.38	**<0.001**
Word Count*	5412	22610	-	-	59670	58675	-0.17	0.87
Sentence Count*	310	1139	-	-	2938	3408	-	-
Paragraph Count*	206	612	-	-	1435	1438	-	-
Length of Paragraph	2.06	2.32	1.67	0.15	2.53	2.83	-0.41	0.68
Punctuation Diversity Score	22.0	16.0	-1.43	0.20	22.4	21.6	2.83	0.006

Note: Both studies used paired-samples t-tests (Study 1: df=6; Study 2: df=58). * denotes corpus-level values (raw totals, or pooled ratios computed as aggregate ÷ aggregate); all inferential tests were conducted on per-document values

moved in the opposite direction on each, falling from plain English into the difficult band despite a marked increase in lexical sophistication and rare word usage. These findings suggest that while GenAI tools can effortlessly augment surface-level vocabulary, they may inadvertently create a "stylistic illusion of competence." When students use LLMs as unscaffolded ghostwriters, the resulting text mimics academic rigour through lexical density but obscures the communicative clarity expected in higher education. The AI texts achieved their sophisticated surface through denser repetition of a narrower vocabulary, exhibiting significantly lower per-document lexical diversity than the students they were simulating; sophistication without variety.

The observed decline in readability likely stems from the LLMs' lack of grounding in the lived learning context: although contextual information about the subject and assessment was supplied in the prompts, the LLMs had no access to the classroom discussions, team experiences, and iterative feedback that students draw upon, operating instead solely on the prompt text and their pre-trained weights. This results in outputs that are lexically sophisticated but lack the communicative clarity and domain specificity expected in authentic student writing. Without task-specific fine-tuning or iterative prompting, the LLMs appear to revert to generalised academic conventions that prioritise surface complexity over relevance and clarity. As a result, unedited outputs may appear polished while failing to meet pedagogical or disciplinary standards.

Over-reliance on unedited LLM generation poses a challenge for assessment, as educators may instinctively reward sophisticated vocabulary without recognising the accompanying regression in readability and lexical variety. By bypassing the cognitive effort required for authentic composition—drafting, revising, and refining arguments for readability—students risk outsourcing the very intellectual struggle essential to deep learning. Furthermore, zero-shot approaches risk homogenising students' voice and tone, replacing individual expression with generic prose that lacks personal nuance and contextual depth (Anderson et al., 2024). Most concerning is the likely outsourcing of critical thinking and argumentation that such delegation entails, which erodes precisely the productive difficulty that writing-to-learn is meant to exercise. The implication of this regression goes beyond writing but touches on creativity, innovation and ideation at large and may even have implications for cognitive development in the long run. Acquiring academic writing skills requires certain intellectual struggle as part of the process, which LLMs short circuit.

To address this, the stylometric signatures identified in this study offer actionable, data-driven applications for educators. First, assessment rubrics must be fundamentally recalibrated. Historically, academic rubrics have often rewarded "sophisticated vocabulary" and dense academic tone—features that zero-shot LLMs effortlessly replicate. Educators must shift grading criteria to heavily weight structural clarity, audience-sensitive readability, and contextual grounding, directly questioning the unreadable lexical inflation typical of AI ghostwriting. Second, writing analytics are better used formatively than forensically. We do not propose them as AI detectors, given the unreliability of single-text detection (Elkhatat et al., 2023; Weber-Wulff et al., 2023). Instead, students can use their own readability, word length, and lexical-variety trajectories as feedback on developing clarity, without any inference about tool use. In this way, analytics support student reflection rather than institutional surveillance. Finally, these stylometric baselines should be integrated into student-facing AI literacy programs. By transparently showing students that authentic academic mastery correlates with simplicity and clarity, educators can dismantle the misconception that LLMs produce "superior" academic writing, empowering students to use AI for ideation rather than raw generation.

For academics, three things follow. First, review rubric language. Criteria such as "sophisticated vocabulary" or "academic register" reward exactly the surface that unedited delegation produces at no cost, and penalise the plainer, more readable register that the students in this cohort moved toward; criteria should name clarity for a specified audience, contextual grounding, and argument instead. Second, design tasks around what a zero-shot model cannot supply: the classroom discussions, client interactions and team history that the students' texts drew on and the models' texts could not. Assessment that requires this grounding is both more authentic and less delegable. Third, the belief that GenAI produces "better" academic prose is a misconception with a measurable basis: the output is more lexically sophisticated but harder to read. This distinction should be made explicit to students.

This study has several limitations. First, the dataset was small and situated within a specific discipline (Software Engineering), with a heavily male-dominated demographic. This limits the representativeness and generalisability of the findings. Future studies should replicate this design across diverse disciplines and demographic groups. Second, the human dataset in Study 2 involved 59 texts produced by 18 students. While we used paired-samples t-tests to control for prompt-level variance, the clustering of multiple texts per student means the observations are not fully independent, a limitation that should be addressed in future work using mixed-effects models with larger cohorts. While we controlled for length in Study 2, future research could employ more advanced length-agnostic metrics. Additionally, given the number of stylometric metrics tested, no correction for multiple comparisons was applied; results near the $p = .05$ threshold should therefore be interpreted with caution, and isolated significant results on low-frequency features (e.g., article frequency) were not interpreted substantively. Third, GenAI outputs reflect model-specific tendencies, which evolve rapidly with new model releases. Finally, the analysis focused on surface-level metrics and did not capture deeper aspects of writing such as coherence, argument quality, or rhetorical style. Integrating deeper semantic natural language processing and machine learning models could address this gap.

Conclusions and future directions

This study explored the stylistic evolution of students' academic writing over time and examined key differences between student-authored texts and unedited GenAI-authored texts. Study 1 showed that, within one cohort and one year, students' own writing shifted toward simpler, more readable language. In contrast, Study 2 showed that unedited, persona-prompted GenAI texts display distinct stylistic features—such as longer words, more rare vocabulary, and lower readability—that actively diverge from this observed natural progression. These differences suggest that while GenAI tools artificially inflate lexical complexity, they do so through a narrower working vocabulary and at a severe cost to readability, producing writing that lacks the communicative clarity typically valued in authentic academic contexts. This has direct implications for assessment practice. By empirically demonstrating that authentic student writing across an assessment sequence trended toward simplicity and readability while AI ghostwriting defaulted to lexical inflation at the cost of readability, this study provides educators with concrete heuristics to redesign assessment rubrics.

Looking ahead, future research should consider larger and more diverse datasets to improve generalisability across disciplines. Integrating more advanced NLP techniques may also help capture deeper features of writing, such as coherence and rhetorical quality, which are beyond the reach of surface-level stylistic metrics. Finally, longitudinal studies that track students' writing with explicitly scaffolded AI assistance (rather than zero-shot delegation) could offer valuable insights into whether AI tools genuinely support writing development or inadvertently create dependency. Such research would inform how best to integrate GenAI into learning environments in ways that uphold both academic integrity and authentic student growth.

References

Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. Creativity and Cognition. In *Proceedings of the 16th C&C Conference*. https://doi.org/10.1145/3635636.3656204

Das, P., Tasmim, R., & Ismail, S. (2015). An experimental study of stylometry in Bangla literature. In *Proceedings of the 2nd EICT Conference* (pp. 575-580). https://doi.org/10.1109/eict.2015.7392018

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, *19*(1), 17. https://doi.org/10.1007/s40979-023-00140-5

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports, 13*, 18617. https://doi.org/10.1038/s41598-023-45644-9

Khan, A. I., Jain, S., Sharma, P., Deep, V., & Mehrotra, D. (2021). Stylometric analysis of writing patterns using artificial neural networks. In *Proceedings of the 2021 3ICT Conference* (pp. 29-35). https://doi.org/10.1109/3ICT53449.2021.9582095

Kim, J., & Lee, S.-S. (2023). Are two heads better than one?: The effect of student-AI collaboration on students' learning task performance. *TechTrends : For Leaders in Education & Training*, *67*(2), 365–375. https://doi.org/10.1007/s11528-022-00788-9

Kim, T. (2015). T test as a parametric statistic. *Korean Journal of Anesthesiology*, *68*, 540. https://doi.org/10.4097/kjae.2015.68.6.540

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), eadt3813. https://doi.org/10.1126/sciadv.adt3813

Lagutina, K. V., Lagutina, N. S., Boychuk, E. I., Vorontsova, I. A., Shliakhtina, E. V., Belyaeva, O., Paramonov, I. V., & Demidov, P. G. (2019). A survey on stylometric text features. In *Proceedings fo the 25th FRUCT Conference* (pp. 184–195). https://doi.org/10.23919/FRUCT48121.2019.8981504

Laramée, F. D. (2018). Introduction to stylometry with Python. *The Programming Historian*. https://doi.org/10.46430/phen0078

Lee, M., Kim, Z. M., Khetan, V., & Kang, D. (2024). Human-AI collaborative taxonomy construction: A case study in profession-specific writing assistants. In *In2Writing '24: Proceedings of the third workshop on intelligent and interactive writing assistants*, pp. 51-57. https://doi.org/10.1145/3690712.3690726

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2018). Surveying stylometry techniques and applications. *ACM Computing Surveys*, *50*(6), 1–36. https://doi.org/10.1145/3132039

Oliveira, E., Mohoni, M., López-Pernas, S., & Saqr, M. (2026). AI collaboration or cheating? Using explainable authorship verification to measure AI assistance in academic writing. *Educational Technology & Society, 29*(2), 100-117. https://doi.org/10.30191/ETS.202604_29(2).RP06

Pan, H., Araujo Oliveira, E., & Ferreira Mello, R. (2025, March). Exploring human-AI collaboration in educational contexts: Insights from writing analytics and authorship attribution. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 903-909). https://doi.org/10.1145/3706468.3706536

Rowland, D. R. (2023). Two frameworks to guide discussions around levels of acceptable use of generative AI in student academic research and writing. *Journal of Academic Language and Learning*, *17*(1), T31–T69. https://journal.aall.org.au/index.php/jall/article/view/915

Shibani, A., Rajalakshmi, R., Mattins, F., Selvaraj, S., & Knight, S. (2023). Visual representation of co-authorship with GPT-3: Studying human-machine interaction for effective writing. In *Proceedings of the 16th EDM Conference*. https://doi.org/10.5281/zenodo.8115695

Stephenson, R., & Armstrong, C. (2026). *Student generative AI survey 2026 (HEPI Policy Note).* Higher Education Policy Institute. https://www.hepi.ac.uk/reports/student-generative-ai-survey-2026/

Wang, D., Zhang, H., Liu, R., Lv, W., & Wang, D. (2014). t-Test feature selection approach based on term frequency for text categorization. *Pattern Recognition Letters*, *45*, 1–10. https://doi.org/10.1016/j.patrec.2014.02.013

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*, 26. https://doi.org/10.1007/s40979-023-00146-z

Guo, Z., Misiejuk, K., López-Pernas, S., Saqr, M., Chen, G. & Oliveira, E. (2026). Sophistication without clarity: Student writing versus unedited GenAI delegation. In Cochrane, T., Narayan, V., Poole, T., Lai, P., Lake, P., Moore, O., & Li, L. (Eds.), *The Future is Connection: Collaboration, Creativity and Compassion in a Digital Society*. Proceedings ASCILITE 2026. Brisbane (pp. 191–201). https://doi.org/10.65106/apubs.2026.3389

Note: All published papers are refereed, having undergone a double-blind peer-review process.
The author(s) assign a Creative Commons by attribution license enabling others to distribute, remix, tweak, and build upon their work, even commercially, as long as credit is given to the author(s) for the original creation.

© Guo, Z., Misiejuk, K., López-Pernas, S., Saqr, M., Chen, G. & Oliveira, E. 2026