

Comparative discourse analysis of AI-generated and student-authored essays: Pedagogical implications for academic writing

Anthony Townley, Alexander Kilpatrick

Nagoya University of Commerce and Business, The University of Aizu

Aichi, Japan, Fukushima, Japan.

anthony_townley@nucba.ac.jp, alex@u-aizu.ac.jp

Abstract

As generative artificial intelligence (AI) tools become embedded in higher education, concerns have emerged about their implications for academic writing development. Grounded in a constructivist view of learning, this study examines the comparative strengths and weaknesses of 12 argumentative essays generated by ChatGPT and 12 essays written by international undergraduate students in response to the same assessment task in an L2 Japanese university context. Using computational linguistic methods, we analysed lexical diversity, syntactic complexity, discourse marker use, and grammar error rates across the corpus of 24 essays. While the AI-generated essays demonstrated relatively strong grammatical accuracy and surface-level coherence, student essays exhibited significantly greater lexical diversity, sentence complexity, and use of temporal and additive discourse markers. Qualitative analysis further revealed that students projected a stronger sense of authorial identity and voice, grounded in contextual knowledge and evaluative reasoning, which was largely absent from AI-generated texts. These findings suggest that although AI tools may be useful for proofreading or idea generation, over-reliance risks undermining the development of critical reasoning, disciplinary engagement, and scholarly voice. The study supports pedagogical approaches that integrate AI cautiously while prioritizing writing as a core process of cognitive and metacognitive development in higher education.

Keywords: AI writing tools, comparative discourse analysis, academic writing pedagogy

1. Introduction

Within writing research, academic writing is widely understood as a process of cognitive, social, and metacognitive development rather than merely a textual product (Matsuda, 2015; Pessoa et al., 2017). From this perspective, the rapid uptake of generative AI tools raises important questions about how emerging technologies interact with the processes through which writers develop disciplinary knowledge, construct meaning and project authorial identity.

Over-reliance on generative AI technologies has the potential to jeopardize the development of these essential critical thinking and writing skills, leading to long-term implications for students' intellectual growth and preparedness for real-world challenges after graduation from university (van Niekerk et al., 2025).

In response to these pedagogical concerns, this study was designed to evaluate the quality of ChatGPT-generated essays in comparison to essays written by L2 undergraduate students in English in a university context in Japan. Drawing on computational and qualitative discourse analysis, the study analyses lexical diversity, syntactic complexity, discourse marker use, and grammar error rates, alongside qualitative discourse features of authorial identity and voice.

The primary objectives of this study are guided by the following research questions:

RQ1. Is there any evidence of comparative weaknesses in the GPT essays, which can help inform academic writing teachers to educate students about the limitations of using AI tools in support of engaging them more in the learning

processes to develop their academic research and writing skills.

RQ2. What are the strengths in the AI-generated essays that can guide teachers and students towards a balanced use of technology, whereby AI tool support, rather than replace, the cognitive skills necessary for academic success.

Accordingly, this study does not evaluate AI-generated writing solely in terms of correctness or deficiency but is designed to investigate how comparative differences can inform key dimensions of writing development for L2 learners, including complexity, coherence and voice.

2. Generative AI technology for Academic Writing

One of the most ubiquitous AI tools used in educational contexts is Open AI's ChatGPT, which is a language model (LLM) based on the Generative Pre-trained Transformer (GPT) architecture (Wang et al., 2023). GPT models are designed for natural language processing tasks such as language understanding and text generation (Tlili et al., 2023). Current versions (4.0 and 5.0) have been pre-trained on extensive amounts of data to learn language patterns and associations, which it uses to generate human-like text when prompted. In a practical sense when you input a prompt into ChatGPT, it analyses input and rapidly generates a response drawn from information acquired via machine learning in engagement with the internet (Pavlik, 2023).

2.1 Pros and Cons

ChatGPT has shown remarkable capacities in comprehending and producing language with the functionality to interpret the meaning of words in context, judge the grammaticality of sentences, identify the thematic structure of arguments and reuse sentence

structures to produce humanlike texts for specific writing tasks (Nguyen et al., 2024). Other functions include the ability to generate content and provide suggestions for better sentence structures and vocabulary choices (Marzuki et al., 2023) and helping students to improve the overall clarity and coherence of manuscripts (Dale & Viethen, 2021). GPT has also been used for error correction feedback (Alexopoulou et al., 2017) and to assist students with self-directed learning activities for writing purposes (Wang, Li & Bonk, 2024).

Irrespective of these benefits, we also need to acknowledge the limitations of using AI tools for academic writing purposes. While ChatGPT generates responses that are grammatically correct and semantically meaningful, they are not always accurate. This is because large language-based AI models don't have 'knowledge' in the usual sense of the word; they don't store or retrieve data; they don't search the internet for information as researchers would do. They are primarily good at predicting the next word(s) in a sequence based on what they have learned (Spital et al., 2023). The most serious flaw that Chomsky (2023) identified in the cognitive evolution of machine learning is the total absence of "critical capacity" which is an important attribute of human intelligence:

Their [AI] deepest flaw is the absence of the most critical capacity of any intelligence: to say not only what is the case, what was the case and what will be the case - that's description and prediction - but also what is not the case and what could and could not be the case. Those are the ingredients of explanation, the mark of true intelligence.

Thus, ChatGPT is regarded as having "uneven factual accuracy" (Vincent, 2022), which means that it does not try to ensure that the content of its text is true, robust, verifiably, valid, generalizable, etc. (Hutson, 2021). In machine learning terms, it can "hallucinate" present legitimate-sounding material that it is not real (Welborn, 2023). Within these cognitive constraints, ChatGPT's modus operandi is to offer generic and descriptive responses to prompts, which means that incremental prompting is needed to gradually focus its attention and train it to give you responses tailored to the user's interests and level of understanding.

2.2 Pedagogical Concerns

In terms of pedagogical outcomes, the use of AI technology has the potential to stymie the learning process for students to construct coherent writing in response to a research task. Writing is the primary means of formal communication in almost all academic and work contexts and acquiring strong writing skills is one of the most important abilities a student can realize though teaching and learning in higher education (Zahari et al., 2025). Academic writing skills serve not only as a medium for articulating knowledge and comprehension within a student's fields of study, but also as a framework for constructing and presenting sophisticated arguments for other academic and work tasks. Being able to write well also encompasses the ability to engage in critical thinking, synthesise research findings, and evaluate scholarly discourse" (Du Preez, 2012; van Niekerk et al., 2025).

In academic contexts, students are expected to apply critical thinking to the research and writing process and not just restate the opinions of others when deciding how to use

research information. Only relevant and valid research sources should be evidenced and referenced correctly. In order to avoid bias, academic writing should also consider all available evidence and when there are contrary opinions, clear reasoning should be given as to why a specific opinion is valued over another. In terms of semantics, it is important for academic writing to be grounded in accurate and relevant discipline-specific knowledge, often achieved through specialized disciplinary terminology (Pessoa et al., 2017).

The concern for academic writing skills development is that students are now able to exploit AI technology to produce essays to an acceptable level for assessment. This is both academically inappropriate and counterproductive for educational outcomes. As teachers having to deal with the misuse of AI tools for learning activities and written assessment in higher education, the authors of this study set out to investigate empirical research findings that can inform teaching and learning practices and help encourage students to develop the cognitive abilities to succeed in academic research and writing by comparing the strengths and weaknesses of essays generated by ChatGPT and written in English by a cohort of L2 undergraduate students in response to the same assessment task.

3. Research Materials and Participants

3.1 The Environmental Law Course

This study sourced student essays from the Environmental Law course, which forms part of the curriculum for a Global Bachelor of Business Administration (GBBA) degree program that is taught in English for international students at a private university in Japan. Most students enrolled in the GBBA program are non-native speakers of English who were admitted with a minimum English language requirement of EIKEN Grade pre-1, TOEFL iBT: 75, Duolingo English Test: 110, PTE Academic: 50, IELTS: 6.0 (Overall Band), or TOEIC®: 750. The GBBA curriculum encompasses both business and liberal arts courses, primarily utilizing the Harvard case method pedagogy to engage students in critical analysis and class discussions of the facts presented in the cases (Desiraju & Gopinath 2001). Written assessment is also assigned to evaluate student knowledge of course content and academic research and writing skills.

3.2 Writing assignment prompt – Argumentative Essay

As part of the formal assessment for the Environmental Law course, students are required to write an argumentative essay in response to the following task description:

Evaluate the coordination of more than one law (local and/or national) to regulate the same environmental problem in a particular country context. Recommend additional legal and/or business-related strategies to help address the problem.

- Write 3 - 4 pages (A4) – 12-point font 1.5-line spacing.
- Introduction + Body Paragraphs + Conclusion.
- Use citations.

3.3 Research Participants

The Environmental Law course is taught to first-year students and the twelve students listed in Table 1

participated in this research study with informed consent by agreeing to write the argumentative essay assessment without any AI technology assistance. Students admitted to the GBBA program with a waiver come from countries where English is the native language or have completed three years of formal education in the English language. Waiver students highlighted with an asterisk completed an International Baccalaureate Diploma with English as the language of instruction.

Student Nationality	Essay topic	English proficiency to enroll in the GBBA program
1 : American	Transportation pollution in California	Native
2 : Filipino	Agricultural pesticides in Australia	Waiver*
3 : Hungarian	Invasive species in Hungary	Waiver*
4 : Indian	Deforestation in India	Waiver*
5 : Filipino	Bomb fishing in the Philippines	Waiver*
6 : Jamaican	Deforestation in Peru	Waiver
7 : Indian	Loss of wetlands in Kerala	TOEFL (iBT) 108 (01/2023)
8 : Vietnamese	Loss of biodiversity in Vietnam	TOEIC 865 (10/2023)
9 : Indonesian	Groundwater quality in Indonesia	IELTS 7.5 (03/2023)
10 : Burmese	Agriculture in Myanmar	IELTS 6.5 (08/2023)
11 : Pakistan	Deforestation in Pakistan	Duolingo DET 110
12 : Mongolian	Air pollution in Ulaanbaatar	Pearson PTE 75

Table 1: English proficiency levels of research participants

The participating students represent a diverse sample of nationalities, English proficiency levels and experiences with academic writing. It is significant to note that only one student is a native speaker of English, whereas the other participants completed the argumentative essay assessment in English as a second language (ESL).

4. Methodology

Statistical analysis examined linguistic features such as lexical diversity, syntactic complexity, and discourse cohesion in all AI-generated and student-authored texts, which are frequently measured in applied linguistics to assess writing proficiency and development, particularly in L2 contexts (O’Leary & Steinkrauss, 2022). Lexical diversity reflects the range of vocabulary used by a writer, while syntactic complexity, often operationalized through sentence length and clause density, indicates the capacity to construct elaborated structures. The effective use of discourse markers contributes to cohesion by signalling relationships between ideas and guiding readers through the progression of arguments.

4.1 Data Preparation and Linguistic Feature Extraction

All essays were programmatically processed in R to ensure standardized preparation and extraction of linguistic features. Texts were first cleaned and concatenated at the paragraph level, discarding non-prose elements. Tokenization and lexical processing were conducted with the *quanteda* package (Benoit et al., 2018), enabling systematic computation of type–token ratio (TTR) as a measure of lexical diversity. Sentence segmentation was approximated using punctuation delimiters, from which sentence counts, word counts, and average sentence lengths were derived. Subordinate clause introducers (e.g., *that*, *which*, *because*, *although*, *if*) were identified to proxy clause counts, and clause density was calculated as the ratio of clause count to sentence count. Together, these measures provided a multidimensional profile of syntactic complexity and fluency for each essay.

4.2 Corpus Compilation

The analytic corpus comprised 24 argumentative essays evenly split between two sources (12 human-authored, 12 GPT-generated). Documents were ingested programmatically in R to preserve provenance and ensure reproducible preprocessing. Each file was read with the *officer* package (Gohel et al., 2025) and summarized automatically. Only rows flagged as paragraph content were retained and concatenated to form a single plain-text representation per document, thereby excluding non-prose elements such as tables, headers/footers, and figure captions. Filenames were preserved and each essay was labelled with a source identifier (Human or GPT). Prior to analysis, paragraph text was trimmed of leading/trailing whitespace, internal runs of whitespace were collapsed, and basic Unicode/encoding normalization was applied to reduce artefacts from differing editors. The ingestion workflow logged file paths and filenames to enable traceability; the full R script used for extraction and cleaning is provided in the supplementary materials (see OSF link). The resulting dataset (one row per essay) served as the single source for all downstream feature extraction and statistical testing.

4.3 Discourse Marker Annotation

To quantify discourse coherence, we annotated texts for the presence of semantic markers across four functional categories, reflecting common rhetorical relations:

- *Causal*: *because, so, since, therefore, as a result, if*
- *Adversative*: *however, but, although, nevertheless, on the other hand*
- *Additive*: *and, also, moreover, furthermore, in addition, as well as*
- *Temporal*: *then, next, after, before, finally, eventually*

These categories were operationalized as regular expression patterns and applied to each text, yielding token-level frequency counts for each category. The marker sets were selected to align with prior research on discourse coherence and argumentative structure.

Discourse coherence was operationalized through automated annotation of semantic markers grouped into four functional categories: causal, adversative, additive,

and temporal. Each category was encoded as a regular-expression pattern with explicit word-boundary anchors (e.g., `\b(...)\b`) to avoid spurious partial matches (for example matching `so` inside `some`). Multiword phrases (e.g., `as a result`, `on the other hand`, `as well as`) were included verbatim in the patterns so that each multiword occurrence was counted as a single marker. Patterns were applied to lowercased text and occurrences were tallied, producing token-level frequency counts per essay. These raw counts were stored as separate variables (additive, adversative, causal, temporal) and later inspected visually on a random subset of essays to verify that the regular expressions captured intended uses while minimizing false positives (for example markers embedded in citations or quoted material). For comparative analyses we used the raw frequency counts and examined length-normalized densities (marker count divided by word count) to control for variation in essay length; marker features were then included among the dependent variables in the multivariate analyses reported below.

4.4 Lexical and Syntactic Feature Extraction

To capture linguistic complexity and fluency, we computed a suite of lexical and syntactic metrics from each essay. Lexical diversity was operationalized as the Type-Token Ratio (TTR), calculated with the `quanteda` package (Benoit et al., 2018), which tokenized essays into word units and derived the ratio of unique word types to total tokens. Sentence segmentation was approximated via terminal punctuation markers (`.`, `!`, `?`), from which a sentence count was obtained. Word counts were estimated through boundary detection of alphanumeric tokens. From these measures, average sentence length was derived as words per sentence. Syntactic embedding was approximated by counting the presence of subordinate clause introducers (`that`, `who`, `which`, `because`, `although`, `while`, `when`, `if`, `as`), yielding a clause count for each text. Clause density was then calculated as clause count per sentence, allowing direct comparison across essays of differing lengths. Collectively, these measures provide interpretable proxies for lexical range, sentence structure, and clause-level subordination, which are central indicators of argumentative writing quality.

- *Lexical Diversity: Measured as the Type-Token Ratio (TTR).*
- *Sentence Count: Estimated by counting terminal punctuation (`.`, `!`, `?`).*
- *Word Count: Based on word-boundary detection.*
- *Average Sentence Length: Calculated as word count divided by sentence count.*
- *Clause Count: Proxied by frequency of subordinate clause introducers (`that`, `who`, `which`, `because`, `although`, `while`, `when`, `if`, `as`).*
- *Clause Density: Computed as clause count per sentence.*

These metrics provide interpretable proxies for syntactic complexity and fluency, which are central to discourse quality.

4.5 Grammar Error Detection

Grammar errors were quantified via the `LanguageTool` API, a rule-based multilingual grammar checker. To accommodate the API's input constraints, each essay was

partitioned into 5,000-character segments and checked iteratively. The number of grammar errors returned per segment was summed to obtain a total error count per document. To expedite processing, grammar checks were parallelized using the `furrr` package (Vaughan & Dancho, 2022) with multisession execution across available CPU cores. An error rate was computed for each essay as the number of grammar errors divided by total word count.

4.6 Statistical Analysis

Although both groups were given the same prompt and identical instructions, the GPT model produced texts well below the required length. While human-authored essays averaged 1781 words ($SD = 299$), AI-generated essays averaged only 1012 words ($SD = 106$). Because of this significant length disparity, all subsequent analyses were conducted on normalized rates rather than raw counts. Accordingly, discourse marker frequencies (causal, adversative, additive, temporal) and grammar errors were expressed as rates per 100 words. Clause density (clauses per sentence) was retained in its original form, and clause rate (clauses per 100 words) was computed as an additional measure of syntactic complexity. Grammar error rate was likewise expressed as errors per 100 words. After normalization, independent-samples linear regression models were used to test differences between human-authored and GPT-generated essays for each linguistic and structural feature. Essay source (Human vs. GPT) served as the independent variable, with each normalized feature entered separately as the dependent variable. Results are reported in Table X, including group means, standard deviations, t values, and p values.

4.7 Data Availability

The R scripts used for preprocessing, feature extraction, and statistical analysis are available through the Open Science Framework repository (OSF): https://osf.io/3aysw/?view_only=46108add23264775bf7c2bef7ec4ae38.

4.8 Authorial Voice and Identity Analysis

Although authorial identity and voice were not operationalized as computational variables, qualitative discourse analysis was employed to examine how writers positioned themselves rhetorically in relation to their topics, sources, and audiences. This extended discourse analysis aligns with writing research that treats identity and voice as emergent features of writing rather than strictly quantifiable features (Matsuda & Tardy, 2007; Matsuda, 2015; Li & Deng, 2019).

In this study, a writer's identity is understood as the textual construction of the writer as a knowledgeable participant in disciplinary discourse with vocabulary choices, argument style, and citation practices (Pittam et al., 2019). For example, writers can integrate multiple sources and use critical argumentation to establish an authorial position in their written work (Othman & Lo, 2023). From this perspective, differences between student-authored and AI-generated essays are not merely stylistic but reflect fundamentally different relationships to knowledge construction and epistemic engagement.

In comparison, voice in academic writing refers to the distinctive presence of the author in the text. Even though

the use of personal pronouns and overt subjectivity are customarily avoided in academic writing, a writer can establish a presence in the text through the discursive tone and style they make claims, evaluate sources, and connect ideas (i.e. urgent, neutral, persuasive, descriptive). The use of specific signal phrases to emphasize certain arguments over others also helps to show critical academic engagement with the topic and improve clarity (Hyland, 2002, 2008; Ivanič, 1998).

5. Results

5.1 Quantitative Discourse Analysis

Preliminary analysis confirmed a substantial difference in essay length between sources. Human-authored essays were on average 769 words longer than GPT-generated essays ($t = 8.05$, $p < .001$). Because of this large discrepancy, subsequent analyses were conducted on normalized rates rather than raw counts. Table 2 summarizes the comparisons of normalized linguistic and structural features between human (Hn), and GPT essays.

Several features showed significant differences. Human-authored texts contained more additive markers ($t = -3.87$, $p = .001$) and displayed substantially higher clause density ($t = 6.91$, $p < .001$), indicating more frequent use of syntactic embedding. They also exhibited longer average sentence lengths ($t = 4.53$, $p < .001$) and greater use of temporal markers ($t = 4.30$, $p < .001$). Interestingly, GPT essays contained slightly more causal markers ($t = 4.42$, $p < .001$), although the absolute rates were very low for both groups.

Variable	Mean (Hn)	SD (Hn)	Mean (GPT)	SD (GPT)	t value	P value
Additive Markers	5.997	4.48	0.69	1.168	- 3.873	0.001
Adversative Markers	0.324	0.29	0.165	0.154	- 0.514	0.612
Average Sentence Length	12.24	17.607	1.334	3.883	4.529	< 0.001
Causal Markers	0.008	0.249	0.028	0.187	4.423	< 0.001
Clause Density	0.133	0.409	0.029	0.135	6.913	< 0.001
Grammar Errors	0.108	0.327	0.373	0.349	1.49	0.15
Lexical Diversity	0.455	0.369	0.02	0.038	- 6.931	< 0.001
Temporal Markers	0.018	0.105	0.041	0.058	4.296	< 0.001

Table 2: Comparative linguistic and structural features of human-authored and GPT-generated essays

In terms of lexical measures, the human-authored essays had significantly higher lexical diversity ($t = -6.93$, $p < .001$), while GPT essays produced more uniform surface structures. Grammar error rates did not differ significantly between groups ($t = 1.49$, $p = .15$). Adversative markers were also comparable across sources ($t = -0.51$, $p = .61$).

5.2 Qualitative Discourse Analysis

The AI-generated and human-authored essays concerned with environmental problems examined in Mongolia and Pakistan were selected for further qualitative comparative analysis. The main features for discourse analysis included the author's evaluative stance on the environmental problems and relevant laws for reform as features of identity, and the tone and style of discursive choices to emphasize research-based claims (urgent, neutral, persuasive, descriptive) as a function of voice.

5.2.1 Mongolia

The student-authored essay on air pollution in Mongolia exemplifies how the linguistic features identified in Section 5.1 support the construction of a situated and evaluative authorial identity. For instance, higher clause density and longer sentence structures were used to situate her analysis and arguments of the environmental problem within broader socioeconomic realities in Mongolia, such as poverty, migration, and public health. It is evident from her language choices that the student is personally concerned about the impact of air pollution in Ulaanbaatar and she was able to incorporate multiple layers of evaluative reasoning to connect analysis of socioeconomic conditions, legal frameworks, and public health outcomes into a cohesive argument. More specifically, her discussion of poverty, migration, and heating practices was not presented as isolated facts, but as interconnected factors contributing to air pollution. This level of syntactic embedding facilitates what Hyland (2002) describes as stance through evaluation, allowing the writer to interpret and prioritize information rather than simply report it. Similarly, the significantly higher use of additive and temporal discourse markers contributes to the development of a progressive argumentative structure, guiding the reader through causal relationships and policy implications. She also establishes a strong academic identity by referencing and critically engaging with both constitutional rights in Mongolia and international principles (Stockholm Declaration, UNICEF reports) to evaluate the coordination of different laws.

The writer establishes her voice by taking an evaluative stance taken toward government policies, acknowledging both successes (e.g., refined coal briquette subsidies) and failures (e.g., carbon monoxide poisoning incidents, ineffectiveness of past expenditures). There is also variation in tone with the use of informative, critical, and advocacy-oriented information to make her voice heard. For instance, the author uses persuasive appeals to draw attention to vulnerable populations, health impacts, and public trust issues in Mongolia when evaluating the ineffectiveness of existing laws and stating that "enhancing accessibility to clean heating alternatives through improved distribution systems and supportive policies for disadvantaged groups is crucial". In terms of structure, the student essay contains more detailed longer sentences than the AI counterpart, with more clause embedding and connective devices to create a persuasive narrative throughout the essay.

In contrast, the AI-generated essay projects a more administrative identity by providing a structured overview of national and local laws without a compelling stance about the effectiveness of these laws. The significantly lower clause density and reduced use of additive markers constrain the development of extended reasoning, resulting in a text that presents legal frameworks in a segmented and descriptive manner. Although the organization appears logical, the lack of syntactic and discourse-level elaboration limits the writer's ability to establish evaluative connections between ideas and there is no engagement with social or human dimensions of poverty, inequality, and health risks in Mongolia that would situate the author as critically engaged with the issue. Furthermore, the voice is impersonal and uniform, relying on bullet-point lists without paragraph development and formal descriptions of laws with little critical or evaluative language. The discursive result is that the AI essays read like a policy briefing of information that lacks the distinctive identity and critical voice that the student established in her argumentative essay.

5.2.2 Pakistan

A similar contrast in identity and voice is evident in the AI-generated and human-authored essays dealing with Pakistan. The higher lexical diversity and clause density identified in the quantitative analysis provide the linguistic resources through which the student constructs a more contextually grounded identity to discuss deforestation as a historical, environmental, and social problem in Pakistan. The student also frames the issue of deforestation in Pakistan as both a national challenge (rapid loss of biodiversity, soil erosion) and a local one (illegal logging and floods) in the northern part of the country. In doing so she references specific statistics and historical context (Forest Act 1927, NDMA reports, deforestation rates), to position herself as informed and critically engaged. In terms of legal reform, the author takes an active stance in recommending practical, context-sensitive solutions such as public awareness campaigns on the basis that "the literacy rate in Pakistan is only 58% and it is important for the government to educate people..."

The higher frequency of additive discourse markers in student essays also plays a critical role in shaping a critical and persuasive voice throughout the essay, addressing weaknesses in law enforcement and urging legal reforms. For instance, the student uses urgent language to state that "the deforestation rate in Pakistan in the last decade has rapidly increased that has not only posed a serious threat to the environment of the country but also the ecosystem and the health of the citizens" and evaluative language "not enough," "ineffective enforcement," "catastrophic aftereffects" to emphasize the importance of addressing deforestation effectively.

In contrast, the AI-generated essay about air pollution in Pakistan reflects the same structural constraints to project a more administrative identity by presenting national and provincial laws in a generalized representation of the topic. The reliance on a bullet-point list of terminology to present gaps in the coordinated laws (e.g., "Jurisdictional Overlaps," "Weak Enforcement," "Business Disengagement") results in lower lexical diversity and reduced use of discourse markers. Furthermore, there is very little critical evaluation of the current regulation of air

pollution and the proposed reforms read like a briefing note of terminology without any personal stance or lived context, unlike the student essay.

Very similar to the AI essay generated for Mongolia, the voice is mechanical, structured, and neutral without any strong judgment or advocacy. As a result, this type of academic writing also fails to convey the distinctive identity and compelling voice that are evident in the student's writing, particularly in the discursive ways she recommends legal reforms to address deforestation in her home context of Pakistan. While grammatically accurate, the text does not demonstrate the epistemic engagement required to construct a persuasive or contextually grounded argument. This integrated analysis reinforces the central claim of the study that AI-generated texts do not replicate the discursive features for original authorial voice and identity.

6. Discussion

6.1 Comparative Strengths

The analysis demonstrates that human-authored essays displayed stronger linguistic and structural qualities compared to those generated by ChatGPT. More specifically, the students produced texts with more complex syntactic structures and greater clause density, which together indicate academic writing quality (O'Leary & Steinkrauss, 2022). The longer average sentence length in student essays further reflects their ability to construct detailed arguments in unified paragraphs that extend beyond simple declarative structures. In comparison, most of the AI essays listed terms related to the evaluation of current laws and recommendations for legal reform in bullet-lists without any paragraph development that is essential for academic writing.

In addition, the more frequent use of temporal markers highlights the ability of students to establish cohesive links between ideas and to guide the reader through the progression of arguments in a logical manner. From a writing-development perspective, these features indicate a greater capacity for elaborated reasoning and more sustained arguments concerning the regulation of environmental problems in the home contexts of students. These findings point to stronger epistemic engagement, more reliable reasoning, and the emergence of authorial voice, as students construct more elaborate arguments about law reforms for environmental issues that are of concern to them.

By contrast, the ChatGPT-generated essays performed comparatively well on selected surface-level measures reported in Table 1, particularly grammatical accuracy, with lower normalized error rates than human-authored texts. GPT generated a wide range of topic-related terminology that may support brainstorming, although overall lexical diversity remained significantly lower than in student essays. However, these strengths remain largely procedural and do not extend to the higher-order features of writing development captured in the analysis of student-written essays. In particular, GPT texts exhibited lower clause density, shorter sentence structures, and limited use of additive and temporal discourse markers, indicating reduced syntactic elaboration and weaker discourse progression. As a result, they lacked the contextually

grounded argumentation and authorial positioning evident in the student essays.

6.2 Comparative Weaknesses

Although human-authored essays demonstrated stronger syntactic cohesion, one recurring issue was the presence of more grammatical errors compared to AI-generated texts. This outcome is to be expected for students writing in their second language without assistance from AI tools. The variability in grammatical accuracy and structural clarity across the student essays reflects differences in language proficiency levels and writing experience of different student participants. This pattern is consistent with L2 writing research that indicates that increased syntactic and rhetorical complexity may co-occur with reduced accuracy during developmental stages of language acquisition (Norris & Ortega, 2009) and L2 writing development remains sensitive to instructional conditions and feedback practices (Polio, 2017).

AI tools may therefore serve a productive pedagogical role to support lower-order writing concerns, such as grammar, sentence structure, and lexical choice, thereby enabling learners to focus more fully on higher-order processes of argument development, evaluation, and disciplinary positioning. However, such use must be carefully guided to ensure that AI functions as a scaffold for writing development rather than a substitute for the cognitive and epistemic work involved in composing academic texts.

Also significant to note is that AI-generated essays showed particular weaknesses in evaluation of research and the development of sustained arguments. As noted above, the ChatGPT essays averaged only 1012 words ($SD = 106$) compared to 1781 words ($SD = 299$) for human-authored essays in response to the same prompt instructions. This length disparity corresponded with lower clause density and shorter sentence structures, indicating reduced syntactic elaboration and limited capacity for extended evaluative reasoning. In addition, the over-reliance on causal markers in repetitive and formulaic patterns, such as bullet-point lists, indicate a restricted range of rhetorical resources that constrain the development of nuanced argumentation.

In comparison, student essays used a wider range and higher frequency of discourse markers, particularly additive and temporal connectors. The use of more explicit signalling of relationships between ideas improves cohesion throughout the essays, which is essential for effective argumentative writing. In contrast, the more limited use of discourse markers in AI-generated texts indicates weaker discourse organization and less explicit rhetorical structuring. Taken together, these results suggest that AI-generated writing approximates surface-level fluency and accuracy but does not reflect the deeper linguistic development associated with the L2 student-authored essays. This distinction is critical, as it highlights that grammatical correctness alone is not sufficient to indicate L2 writing development. Instead, writing proficiency involves the ability to construct complex, coherent, and rhetorically effective discourse (Wolfe-Quintero, Inagaki, and Kim, 1998).

6.3 Pedagogical Implications

Collectively, these statistical and qualitative findings support teaching and learning practices that emphasize writing as a process of knowledge construction rather than mere text production (Pessoa et al., 2017; Zahari et al., 2025). The ability of students in this study to apply reasoning across diverse environmental law contexts and sustain complex and persuasive argumentation demonstrates how writing tasks function as sites for intellectual engagement. By contrast, the performance of ChatGPT reflects a descriptive, rather than critical capacity, to engage with the research topic in argumentative essays, consistent with arguments that machine learning models cannot produce explanations or counterargument reasoning (Chomsky, 2023). The study therefore affirms the value of pedagogy that enables students to cultivate the cognitive and metacognitive skills necessary for academic writing, even as AI technologies become more prevalent in academic settings. These academic skills are not only essential for success in higher education, but also in professional work contexts after graduation (Zahari et al., 2025; Du Preez, 2012).

Writing instruction should therefore distinguish between acceptable AI-supported activities (e.g., proofreading, grammar checking, brainstorming, outlining) and activities that should remain the responsibility of the student (e.g., evaluating evidence, constructing arguments, synthesising sources, and developing authorial voice). Similar to the academic writing pedagogy used in the Environmental Law course, teachers can also assist students by teaching the genre conventions for argumentative essays and providing them with annotated examples of strong essays and highlighting elements like thesis statements, topic sentences, transitions, citations, and paragraph cohesion. Feedback on grammar, clarity, coherence, and argumentation should also be given to students with the opportunity for them to learn from mistakes and revise drafts of the essay (Zahari et al., 2025).

6.4 Academic Writing Outcomes

When students write their own essays, the pedagogical outcomes extend beyond the immediate assessment to help them develop a range of cognitive skills as a process of learning. First, independent writing fosters research skills, as students must locate, assess, and correctly cite credible sources. This process enhances academic integrity and guards against fabricated or unreliable references, which are a persistent problem in AI-generated texts (King, 2023; Welborn, 2023). The writing process then supports the development of critical thinking by requiring students to evaluate research information, consider counterarguments, and express authorial stance and perspective rather than relying on neutral text (Pessoa et al., 2017). This intellectual engagement strengthens their ability to evaluate issues from multiple perspectives.

Through the writing process, student can improve grammatical accuracy, vocabulary range, and syntactic complexity (Alexopoulou et al., 2017). Students can also improve cohesion and coherence in their writing by practicing the use of discourse markers and structuring logical arguments in unified paragraphs with effective topic sentences. There is also the sense of achievement from producing original writing that enables students to project a distinctive scholarly identity and establish their academic

voice, demonstrating ownership of their arguments and active participation in disciplinary debates (Du Preez, 2012).

Essay writing also promotes metacognitive awareness, as students reflect on their strengths and limitations during the writing process. There is the opportunity for them to learn and improve writing skills when errors occur, as the proofreading and redrafting process supports long-term language growth. The cumulative effect of these pedagogical outcomes is the acquisition of transferable skills - evidence-based reasoning, problem solving and written communication skills - that are essential for both academic and professional contexts. This study therefore highlights need for higher education academic writing program to support independent and original essay writing as a core pedagogical process for cultivating these essential cognitive skills and long-term intellectual growth.

6.5 Integration of AI Tools

It is also important to acknowledge the relative strengths of GPT-generated essays in grammatical accuracy and brainstorming content (titles, outlines, counterarguments), which indicates that AI tools can serve a complementary role in writing instruction. When used responsibly, AI can assist students in proofreading, reducing grammatical errors, and generating structural outlines, thereby allowing learners to focus their attention on higher-order skills such as analysis, argumentation, and synthesis (Alexopoulou et al., 2017; Wang, Li & Bonk, 2024). However, these benefits must be carefully balanced against the risk of over-reliance on AI, which could limit the development of critical thinking and research skills if students substitute machine-generated content for their own intellectual effort (van Niekerk et al., 2025).

For educators, the challenge lies in designing learning environments that both integrate AI tools and safeguard against their misuse. This could involve teaching students to critically evaluate AI outputs, to identify inaccuracies or fabricated sources, and to treat AI-generated text as a starting point to generate ideas rather than an endpoint in the writing process. Such approaches would align with calls in the literature to promote transparency and accountability in the use of AI technologies for academic purposes (Else, 2023; Cotton et al., 2023; Malik et al., 2023; Marzuki et al., 2023). Ultimately, the pedagogical task for both teachers and learners is to ensure that AI supports rather than replaces the acquisition of academic writing competencies.

6.6 Research Limitations and Future Research

This study has several limitations that should be acknowledged. First, the analysis was based on a small sample of 12 essays from one undergraduate course at a single higher education institution, which limits the generalizability of the findings. The scope of the task is also limited to a single argumentative essay prompt in the field of environmental law, and analytical outcomes may differ across other disciplines, writing genres, or assignment types. Another limitation concerns the comparison with AI outputs. Only ChatGPT (version 4) was used for this study, and the GPT essays were generated from a single prompt, so the essay length and content quality of the AI-generated essays could be improved with more specific instructions and iterative prompt refinement (Korzynski et al. 2023). Taken together, these limitations indicate that future

research using larger, more diverse samples and multiple AI tools across disciplines would provide a fuller picture also important

7. Conclusion

This study contributes to the literature by integrating computational discourse analysis with qualitative analysis of authorial voice and identity to examine how AI-generated and student-authored texts differ as representations of writing development. The findings indicate that while AI-generated texts offer advantages in grammatical accuracy and content generation, they demonstrated weaker performance in critical reasoning, discourse cohesion, and persuasive argumentation. In contrast, student essays displayed richer lexical diversity, greater syntactic complexity, and more personal engagement with environmental law topics with a stronger projection of academic identity and voice.

In response to concerns that over-reliance on generative AI technology will undermine the development of essential skills for academic research and writing, this study affirms the importance of teachers engaging students in writing tasks as a process of intellectual development. At the same time, the findings also recognize that AI tools may provide support with surface-level writing features such as grammar and formatting. Balancing technological affordances with academic integrity is therefore crucial for successful pedagogical outcomes.

As teachers in higher education contexts, the authors hope that students will be less inclined to use language models and commit more to research and writing learning processes, if they experience or comprehend the limitations of ChatGPT and other AI writing tools. These research results can therefore help inform writing educators to commit more teaching resources to original authorship pedagogy and influence the intrinsic motivation of students to engage more in authentic writing activities to develop the cognitive skills necessary for academic and professional work success.

8. Bibliographical References

- Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67(S1), 180–208.
- Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30), 774. doi:10.21105/joss.00774, <https://quanteda.io>.
- Chomsky, N. (2023, March). "Opinion | Noam Chomsky: The False Promise of ChatGPT." *The New York Times*, <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239. <https://doi.org/10.1080/14703297.2023.2190148>

- Dale, R., & Viethen, J. (2021). The automated writing assistance landscape in 2021. *Natural Language Engineering*, 27(4), 511-518.
- Desiraju, Ramarao, D & Gopinath, C. (2001). Encouraging participation in case discussions: A comparison of the MICA and the Harvard case methods. *Journal of Management Education*. 25(4): 394-408.
- Du Preez, I. (2012). Developing academic writing skills as part of graduate attributes in undergraduate curricula. *South African Journal of Higher Education*, 26(2). <https://doi.org/10.20853/26-2-165>.
- Else H. (2023). Abstracts written by ChatGPT fool scientists. *Nature*, 613(7944): 423-423. DOI: <https://doi.org/10.1038/d41586-023-00056-7>
- Gohel, D., Moog, S., & Heckmann, M. (2025). officer: Manipulation of Microsoft Word and PowerPoint documents (R package version 0.7.1.001). Retrieved from <https://github.com/davidgohel/officer>
- Hutson, M. (2021). Robo-writers: The rise and risks of language- generating AI. *Nature*, 591(7848): 22-25. DOI: <https://doi.org/10.1038/d41586-021-00530-0>
- Hyland, K. (2002). Authority and invisibility: Authorial identity in academic writing. *Journal of Pragmatics*, 34(8), 1091-1112.
- Hyland, K. (2008). *Disciplinary voices: Interactions in research writing*. University of Michigan Press.
- Ivanič, R. (1998). Writing and identity: The discursal construction of identity in academic writing. John Benjamins.
- King, M. R. (2023). ChatGPT. A Conversation on Artificial Intelligence, Chatbots, and Plagiarism in Higher Education. *Cellular Molecular Bioengineering*, 16, 1-2. DOI: <https://doi.org/10.1007/s12195-022-00754-8>
- Korzynski, P., G. Mazurek, P. Krzypkowska, and A. Kurasinski. (2023). Artificial Intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review*, 11(3): 25-37. <https://doi.org/10.15678/EBER.2023.110302>.
- Li, Y., & Deng, L. (2019). Writer identity construction revisited: Stance, voice, self, and identity in academic written discourse. *Chinese Journal of Applied Linguistics*, 42(3), 327-344. <https://doi.org/10.1515/CJAL-2019-0020>
- Malik, A. R., Pratiwi, Y., Andajani, K., Numertayasa, I. W., Suharti, S., Darwis, Marzuki. (2023). Exploring artificial intelligence in academic essay: Higher education student's perspective. *International Journal of Educational Research Open*, 5, 100296,
- Marzuki, Widiati, U., Rusdin, D., Darwin, & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective, *Cogent Education*, 10(2), 2236469.
- Matsuda, P. K. (2015). Identity in written discourse. *Annual Review of Applied Linguistics*, 35, 140-159. <https://doi.org/10.1017/S0267190514000178>.
- Matsuda, P. K., & Tardy, C. M. (2007). Voice in academic writing: The rhetorical construction of author identity in blind manuscript review. *English for Specific Purposes*, 26, 235-249.
- Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, 1-18. <http://doi.org/10.1080/03075079.2024.2323593>.
- O'Leary, J. A., & Steinkrauss, R. (2022). Syntactic and lexical complexity in L2 English academic writing: Development and competition. *Ampersand*, 9, 100096.
- Othman, J., & Lo, Y. Y. (2023). Constructing academic identity through critical argumentation: A narrative inquiry of Chinese EFL doctoral students' experiences. *SAGE Open*, 13(4). <https://doi.org/10.1177/21582440231218811>
- Pavlik, J. V. (2023). Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *Journalism & Mass Communication Educator*, 78(1), 84-93.
- Pessoa, S., Mitchell, T. D., & Miller, R. T. (2017). Emergent arguments: A functional approach to analyzing student challenges with the argument genre. *Journal of Second Language Writing*, 38, 42e55. <https://doi.org/10.1016/j.jslw.2017.10.013>
- Pittam, G., Elander, J., Lusher, J., Fox, P., & Payne, N. (2009). Student beliefs and attitudes about authorial identity in academic writing. *Studies in Higher Education*, 34(2), 153-170. <https://doi.org/10.1080/03075070802528270>
- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education, *Smart Learning Environment* 10(1), 15.
- van Niekerk, J., Delpont, P. M. J., & Sutherland, I. (2025). Addressing the use of generative AI in academic writing. *Computers and Education: Artificial Intelligence*, 8, 100342,
- Vaughan, D., & Dancho, M. (2022). furr: Apply mapping functions in parallel using futures (R package). Retrieved from <https://furr.futureverse.org/>
- Vincent J. (2022, December 5). AI-generated answers temporarily banned on coding Q&A site. *Stack Overflow*. <https://www.theverge.com/2022/12/5/23493932/chatgpt-ai-generated-answers-temporarily-banned-stack-overflow-llms-dangers>
- Wang, F. Y., Miao, Q., Li, X., Wang, X., & Lin, Y. (2023). What does ChatGPT say: the DAO from algorithmic intelligence to linguistic intelligence, *IEEE/CAA J. Autom. Sin.* 10(3), 575-579.
- Wang, C., Li, Z., & Bonk, C. (2024). Understanding self-directed learning in AI-assisted writing: A mixed methods study of postsecondary learners. *Computers and Education: Artificial Intelligence*, 6, 100247. <https://doi.org/10.1016/j.caeai.2024.100247>. <https://www.sciencedirect.com/science/article/pii/S2666920X2400050X>.
- Welborn A. (2023). ChatGPT and fake citations. Duke University Libraries. <https://blogs.library.duke.edu/blog/2023/03/09/chatgpt-and-fake-citations/>
- Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). Second language development in writing: Measures of fluency, accuracy, and complexity. University of Hawai'i Press.
- Zahari, M., Marini, A., Safitri, D., Dewiyani, L., & Muawanah, U. (2025). Optimizing student writing performance in higher education: A quantitative study of teacher feedback and classroom environment, *Social Sciences & Humanities Open*, 101286.