Applied Corpus Linguistics 4 (2024) 100106

Contents lists available at ScienceDirect

Applied Corpus Linguistics

journal homepage: www.elsevier.com/locate/acorp

Articles

Identifying ChatGPT-generated texts in EFL students' writing: Through comparative analysis of linguistic fingerprints

Atsushi Mizumoto [a,*], Sachiko Yasuda [b], Yu Tamura [a]

[a] Kansai University, Japan
[b] Kobe University, Japan

ARTICLE INFO

Keywords:
ChatGPT
Generative AI (GenAI)
L2 writing
Academic misconduct
Natural language processing (NLP)

ABSTRACT

The emergence of generative AI (GenAI) poses new challenges for L2 writing teachers. This study investigates the distinguishability of essays written by Japanese EFL learners from those generated by ChatGPT. Partially replicating Herbold et al. (2023), 140 first-year university students wrote essays and completed a survey on ChatGPT use. Among them, 125 wrote independently, 13 used ChatGPT for proofreading, and two asked ChatGPT to write the entire essay. To create a comparative dataset, 123 additional essays were generated by ChatGPT, imitating the two texts. The resulting 263 essays were then analyzed using the natural language processing (NLP) technique, including automated linguistic analysis and machine learning classification using random forest. The results reveal significant differences between human-written and ChatGPT-generated essays across all linguistic features, with the latter being easily identifiable. This study emphasizes the need for clear guidelines on the ethical use of AI in L2 writing, highlighting the potential risk of inappropriate AI use and the importance of fostering a mutual understanding of AI use with learners regarding responsible AI integration in academic work.

Introduction

Generative AI (GenAI), such as ChatGPT, has ushered in a new era in second language (L2) learning and teaching, particularly in the field of L2 writing. These advanced AI systems provide valuable support for language learners, potentially enhancing their writing skills through instant feedback, error correction, and content generation (Barrot, 2023; Steiss et al., 2024). However, their capabilities also raise significant questions about academic integrity, the nature of language learning, and the future of writing instruction (Warschauer et al., 2023). While AI provides invaluable assistance, it is designed to complement, not replace, human instruction, thereby enriching the educational process without supplanting it.

The integration of ChatGPT and similar AI tools into L2 writing practices presents a "double-edged sword" (Derakhshan & Ghiasvand, 2024). On one hand, these technologies offer learners valuable assistance throughout various stages of the writing process, including generating ideas, drafting and revising essays, and refining their language use (Su et al., 2023). Proficient writers, as demonstrated in L2 writing research, are more likely to be engaged in all stages of the writing process recursively compared to novice writers, and this significantly impacts the quality of the final products they produce (Flower & Hayes, 1981; Williams, 2003). During the writing process, these automated writing evaluation systems offer corrective feedback on learner writing. They identify errors and infelicities in the writing and provide suggestions for improvement tailored to individual students' needs. Since it can be time-consuming and challenging for human instructors to offer personalized feedback and support (Ferris, 2010), the potential for enhancing the writing process and improving writing outcomes is especially appealing in contexts where direct instruction time may be limited.

On the other hand, the ease with which ChatGPT can generate coherent text, typically with few errors, poses challenges to traditional notions of authorship and academic honesty. There are growing concerns about students relying too heavily on AI-generated content, potentially bypassing the crucial cognitive processes involved in language learning and writing development. Moreover, the ability of these tools to produce human-like text raises questions about how teachers can accurately assess students' true language abilities and writing skills (Fleckenstein et al., 2024).

* Corresponding author.
E-mail address: mizumoto@kansai-u.ac.jp (A. Mizumoto).

https://doi.org/10.1016/j.acorp.2024.100106
Received 8 July 2024; Received in revised form 3 September 2024; Accepted 18 September 2024
Available online 26 September 2024

This tension between the potential benefits and risks of AI in L2 writing is further complicated by the rapid pace of technological advancement. As large language models (LLMs) such as GPT become increasingly sophisticated, there is ongoing debate about the relationship between human-authored and AI-generated text. Recent studies, such as those by Mizumoto et al. (2024), have demonstrated the high performance of ChatGPT in various language tasks. However, contrary to the notion that the line between AI and human writing is becoming indistinguishable, some research suggests a more complex view. For instance, Berber Sardinha (2024) argues that despite superficial similarities, linguistic analysis can reveal significant differences between AI-generated and human-authored texts. This evolving situation challenges practitioners and researchers to reassess their approaches to writing instruction, assessment. It also highlights the importance of identifying linguistic markers that differentiate AI from human writing.

In light of these challenges, our study investigates the distinguishability of essays written by learners from those generated by ChatGPT. Building upon the work of Herbold et al. (2023), we analyze a corpus of essays written by university EFL (English as a foreign language) learners, including those who used ChatGPT to varying degrees in their writing process. By comparing these essays with a set of ChatGPT-generated texts, we seek to identify key linguistic features that differentiate human-authored from AI-generated content in the context of L2 writing. When analyzing key linguistic features, we focused on both form-based features and discourse aspects of writing that can greatly impact the overall quality of writing. Specifically, we selected seven measures for analysis: lexical diversity, clausal syntactic complexity, embedded syntactic complexity, complex nominals (nominalizations), modals, epistemic markers, and discourse markers. In addition to these seven linguistic features, we also investigated accuracy and fluency features of students' writing.

This research aims to enhance our understanding of AI detection in academic writing and explore its potential impacts on L2 learners' writing processes and outcomes. By examining essays that involve different levels of AI assistance—from no use, to partial assistance (e.g., editing, proofreading), to complete reliance—we aim to shed light on the ways in which ChatGPT might influence learners' language production and writing processes. Our study also offers insights into how ChatGPT can be effectively integrated into the writing process while maintaining academic integrity and fostering genuine language development. Ultimately, the findings of this study will contribute to the ongoing debate about how to maximize the potential of AI in language education while preserving the integrity and educational value of L2 writing tasks.

Literature review

Since ChatGPT's public release on November 30, 2022, the number of studies utilizing GenAI in a wide variety of fields has been increasing exponentially. The high performance of ChatGPT, particularly GPT-4.0, in various language tasks is increasingly being demonstrated. For example, Mizumoto and Eguchi (2023) used the GPT-3 text-davinci-003 model to automatically score 12,100 essays from the TOEFL11 corpus, written by English learners. The scores were then compared to benchmark levels, revealing that when GPT scores were combined with linguistic features, the model achieved a Quadratic Weighted Kappa of 0.605 (95 % CI [.589, 0.620]) with the original TOEFL11 levels, indicating a substantial level of agreement in automated essay scoring. This has also been confirmed in newer studies using GPT-4 (Tate et al., 2024; Wang and Gayed, 2024; Yamashita, 2024).

With an increasing number of publications reporting the effects of using ChatGPT for L2 learning and teaching, Yang and Li (2024) recently conducted a systematic review of 44 studies, published in 2022 and 2023. They reported that the most common ChatGPT tasks reported in the reviewed studies were content generation, feedback, teaching support, assessment or grading, and recommendation (such as recommending learners with personalized learning materials). The studies highlighted several benefits, including improving perceptions and attitudes towards L2 learning, enhancing four language skills, facilitating autonomous learning and increasing L2 practice, providing an interactive, enjoyable, and engaging learning environment, and reducing teachers' workload while enabling more flexible teacher roles.

However, the systematic review by Yang and Li (2024), at the same time, reported the challenges and limitations associated with the use of ChatGPT in L2 learning. Among these, the most notable concerns include issues of plagiarism, where learners may overly rely on AI-generated content without fully engaging in the learning process. Additionally, some studies reported that ChatGPT's feedback might lack the depth and accuracy required for nuanced language learning, potentially leading to the reinforcement of incorrect language use if not carefully monitored. There are also concerns about the tool's ability to generate culturally insensitive or biased content, which could hinder the learning experience for students from diverse backgrounds. Moreover, the over-reliance on ChatGPT could lead to a reduction in critical thinking and problem-solving skills, as learners might depend too heavily on AI-generated outputs rather than developing these skills independently.

Among the four language skills, writing was the most dominantly researched skill in Yang and Li's (2024) review, largely due to ChatGPT's versatile conversational-based response feature. While ChatGPT has been used as a tool for enhancing writing skills, providing formative feedback, and assisting in tasks such as outline preparation, content revision, proofreading, and post-writing reflection (Barrot, 2023; Su et al., 2023; Zou and Huang, 2023), these applications are not without their drawbacks. The tool's feedback may sometimes lack the granularity needed to address complex language issues, and its ability to facilitate higher-order cognitive processes in writing remains questionable. Additionally, while ChatGPT offers benefits such as enhanced language exposure and personalized learning (Steiss et al., 2024), there is a risk that learners might become passive recipients of AI-generated content rather than active participants in their language development. This underscores the need for careful integration of ChatGPT into the curriculum, ensuring that it complements rather than supplants traditional language learning methods.

Until now, there has been significant confusion and debate surrounding the use of ChatGPT in L2 instruction, particularly in writing. Nevertheless, the consensus is that its use should not be prohibited. As Warschauer et al. (2023) claim, "even if we could ban it, we shouldn't" (p. 5). In fact, numerous studies recommend that teachers and educational institutions should establish guidelines for the ethical use of GenAI (e.g., Zou and Huang, 2023). Thus, a balanced approach that integrates ChatGPT with traditional instruction is advocated, focusing on creativity, authorship, and critical thinking.

To ensure that GenAI tools like ChatGPT are used effectively, strategies such as the "human-in-the-loop approach" (Ranade & Eyman, 2024) and the "human-centered approach" (UNESCO, 2023) have been proposed. The "human-in-the-loop approach" involves active human oversight, where teachers guide and refine the AI's outputs, ensuring that the technology supports rather than supplants human creativity and judgment. The "human-centered approach" emphasizes the importance of aligning AI tools with human values and educational goals, making sure that AI serves to enhance rather than detract from the learning experience. These strategies foster AI literacy among teachers, helping them navigate the opportunities and challenges posed by AI in education (Kasneci et al., 2023). Moreover, they promote a collaborative relationship between AI and educators, where each complements the other's strengths. Additionally, this shift highlights the need to redefine and reinvent assessment methods to better align with the evolving educational paradigm (Foung et al., 2024).

The appropriate use of ChatGPT to support writing may enable L2 learners to enhance their skills, fostering self-regulated, autonomous learners—an ideal outcome for L2 language learning and teaching.

However, the use of ChatGPT is not without drawbacks and can be a "double-edged sword" (Derakhshan & Ghiasvand, 2024). It is crucial to recognize that effective use of ChatGPT as a complementary tool in L2 learning requires a certain level of proficiency (e.g., writing ability) (Woo et al., 2024), learner motivation, and AI literacy. In contexts where English is taught as a foreign language, if learners are not motivated enough to improve their skills, unrestricted use of ChatGPT may lead to patchwriting or plagiarism (Pecorari, 2023), which undermine original thought and academic integrity, or even encourage passive learning (Crosthwaite & Baisa, 2023).

In light of the growing concern over learners submitting ChatGPT-generated writing as their own work, issues of plagiarism and cheating have become significant in academic research. Detecting AI-generated text is complex, posing a central challenge to maintaining academic integrity. Researchers typically use one or a combination of three methods to address this issue: (a) employing AI detectors, (b) conducting human evaluations, and (c) classifying text based on linguistic features.

Recent studies utilizing AI detectors have yielded varied results. Liang et al. (2023) evaluated seven widely used GPT detectors on TOEFL essays from non-native English speakers and US eighth-grade essays. While the detectors accurately classified US student essays, they incorrectly labeled more than half of the TOEFL essays as "AI-generated," with an average false-positive rate of 61.3 %. This finding highlights potential biases in AI detection systems against non-native English writers. In a more comprehensive study, Kar et al. (2024) tested 10 AI-detector tools on both human-written and ChatGPT-generated content. The sensitivity of these detectors varied dramatically, ranging from 0 % to 100 %. In another study, Ibrahim (2023) investigated the potential of two types of RoBERTa-based AI text detection platforms in identifying AI-assisted plagiarism in ESL composition classes. While both platforms demonstrated some ability to identify AI-generated texts, their detection accuracy was inconsistent across the dataset. The inconsistency in accuracy across different types of texts underscores the ongoing difficulties in distinguishing between AI-generated and human-written content.

As false positives or false negatives are indeed inevitable in the detection of AI-generated text, it leads to the conclusion that "the reliance on AI detectors is not the answer" (Godwin-Jones, 2024). The limitations of AI detection tools were further underscored when OpenAI, the developer of ChatGPT, discontinued its AI Text Classifier in July 2023, reflecting that the technology is not viable in its current form.

Human evaluation has emerged as another approach to identifying AI-generated content. Casal and Kessler (2023) examined reviewers' ability to distinguish between AI- and human-generated writing in research abstracts. Their findings revealed that reviewers were largely unsuccessful in identifying AI versus human writing, with an overall positive identification rate of only 38.9 %. Similarly, Fleckenstein et al. (2024) found that teachers could not reliably identify texts generated by ChatGPT among student-written texts, though more experienced teachers made more accurate judgments. These results highlight the limitations of relying on human judgment for detecting AI-generated text, even among language experts and experienced educators, underscoring the complexity of addressing issues related to the use of AI-generated text in academic and educational settings.

The third approach involves using linguistic features in the texts produced by AI and humans. Berber Sardinha (2024) proposed, through multidimensional analysis, the importance of consireding register when distinguishing AI-generated and human-authored texts. The findings revealed significant disparities between these two types of texts. Berriche and Larabi-Marie-Sainte (2024) proposed a technique employing intrinsic stylometric features of documents (i.e., lexical, syntactic grammatical, syntactic structural features) to detect ChatGPT-based plagiarism. Using classical and ensemble classifiers, they achieved 100 % accuracy in distinguishing between human and ChatGPT writing styles on a dataset of TOEFL essays. Similarly, Desaire et al. (2023) developed a method for discriminating text generated by ChatGPT from academic scientists' writing, achieving over 99 % accuracy using carefully selected linguistic features and machine learning.

Another significant contribution to this approach comes from Herbold et al. (2023), who conducted a large-scale comparison of human-written versus ChatGPT-generated argumentative essays. Their study utilized a corpus of 90 essay topics, with each topic having one human-written essay and corresponding AI-generated essays from ChatGPT. They employed 111 high school teachers to rate the essays based on seven criteria, including topic completeness, logical structure, language mastery, and complexity. The results showed that ChatGPT-generated essays were consistently rated higher in quality than human-written essays across all criteria. At the same time, the study revealed distinct linguistic differences between human and AI-generated content. AI essays showed higher sentence complexity and more nominalizations, while human essays used more modal and epistemic constructions.

While the Herbold et al. (2023) study provides robust evidence of the usefulness of employing linguistic features to detect differences between human and AI-written essays, it was conducted in a German educational context with relatively high English proficiency learners. There is a need to explore whether similar results would be obtained in different cultural and linguistic settings, particularly in Asian EFL contexts. Furthermore, the linguistic features they employed did not include accuracy and fluency, which are conventionally included in the CAF (complexity, accuracy, and fluency) framework in L2 writing assessment. Additionally, their study compared fully human-written essays with fully AI-generated ones, but did not explore scenarios where students might use AI tools for partial assistance (e.g., editing, proofreading). Given these gaps, a replication study focusing on the Asian EFL context and incorporating NLP-based detection methods would be valuable for several reasons:

1. It would test the generalizability of Herbold et al.'s findings in a different cultural and linguistic context.
2. By incorporating NLP-based detection methods, it could provide practical tools for educators to identify AI-generated or AI-assisted content.
3. Exploring partial AI assistance scenarios would offer insights into how students actually use these tools in practice and how this impacts the detectability and quality of their writing.
4. Such a study could provide valuable insights for EFL pedagogy, helping educators understand how to effectively integrate AI tools into language learning while maintaining academic integrity.

For these reasons, we conducted a partial replication of Herbold et al. (2023) in the current study, aiming to address these research gaps and contribute to the growing body of knowledge on AI-generated text detection in educational contexts.

This study

This study seeks to explore the following research questions, grounded in the literature review provided:

RQ1: In line with previous findings, are there discernible differences between essays composed by EFL learners and those generated by ChatGPT?

RQ2: Can NLP techniques effectively detect essays written by learners who used ChatGPT to compose the entire essay?

RQ3: What distinctive features can be identified in the essays of learners who utilize ChatGPT for editing and proofreading their self-written work, as opposed to relying on ChatGPT to compose the entire essay?

By exploring these research questions, this study seeks to contribute to the growing body of literature surrounding the impact of language models like GPT on EFL learners' writing practices and the potential for NLP-based detection methods to differentiate between human-authored

and AI-assisted texts.

Methods

Participants

The participants in this study were 140 first-year university students enrolled at a private university located in the western part of Japan. The cohort included 48 males and 92 females, all from a single faculty of foreign language studies, specifically focusing on English. These students were all taking a mandatory course designed to enhance their grammar and vocabulary skills in preparation for participating in a one-year study abroad program in the following year. The data for this study was collected through convenience sampling, as writing was a component of the course assignments. The English proficiency of the participants was assessed using the TOEFL ITP test, with scores generally ranging from the Common European Framework of Reference (CEFR) B1 to B2 levels. All participants had received a minimum of eight years of mandatory English education in Japan's primary and secondary schools.

Procedure

Participants were assigned a task as part of their coursework to write an essay of 200 to 300 words within a 70 min time limit. They were instructed to write the essay outside of class time and submit it via the Learning Management System (LMS). The essay topic was: "Do you agree or disagree with the following statement? Use reasons and specific details to support your opinion. It is important for college students to have a part-time job." The task directions specifically stated not to use any reference materials, such as dictionaries, while writing. However, there were no ethical guidelines provided regarding the use of GenAI, such as ChatGPT. This omission was intentional, leaving open the possibility that some participants might use GenAI tools to assist in their essay writing.

The following week after submitting the essay task, participants were informed about the purpose of the study and surveyed regarding their use of ChatGPT as summarized in Table 1. They were reassured that their survey responses would not impact their course grades and were urged to provide honest answers to support the research. Written consent was obtained from all participants. According to the survey results, 15 out of 140 participants acknowledged using ChatGPT for various purposes (Table 1). Notably, among these 15 participants, two allowed ChatGPT to write their entire essays. However, the majority (n = 125, 89.3 %) did not use ChatGPT and authored their essays independently.

The two individuals who "let ChatGPT write everything" are considered to be outsourcing the work to ChatGPT, which can be viewed as inappropriate AI use or cheating. As the purpose of this study was to compare the essays written independently by learners and those written by ChatGPT, we utilized the two original essays as seed content to instruct ChatGPT to generate similar essays. Subsequently, 123 essays were created using Python with OpenAI's API (Application Programming Interface), GPT-3.5 Turbo model, which has capabilities equivalent to the free browser version of ChatGPT (GPT-3.5) used by the two participants. The prompts included the two students' essays written entirely by ChatGPT as examples, and instructions were added stating, "The vocabulary and grammar difficulty, essay length, and structure should closely resemble the two examples provided below." Along with the two original essays, this led to a total of 125 ChatGPT-generated essays (i.e., 2 original essays + 123 generated ones), closely matching the writing level of the two initially considered to have committed academic misconduct (see the online supplementary material for the Python code at https://osf.io/bj8kq/). These 125 essays were used for comparison against another set of 125 essays written by participants without using ChatGPT, as well as essays from learners who partially used ChatGPT for editing and proofreading. This approach created a specialized mini-corpus for the comparison purposes in the analysis, allowing us to address all research questions, including the distinctive features of essays partially assisted by ChatGPT (RQ3). The breakdown of these essays is illustrated in Fig. 1.

Through these procedures, 125 essays written by learners without using ChatGPT (Human), 125 essays generated using ChatGPT (ChatGPT), and 13 essays by learners who partially used ChatGPT (Mix) were created. Fig. 2 shows the word count for each type. The median word count for Human essays is 216 words (Min = 83, Max = 348), the median for ChatGPT essays is 280 words (Min = 215, Max = 348), and the median for Mix essays is 208 words (Min = 169, Max = 334). From this, it can be seen that the essays generated by ChatGPT have the highest word count and come closest to the task requirement (200 to 300 words). In contrast, the Human essays demonstrate the greatest variability, often not meeting the minimum word count, while the Mix essays display word counts comparable to the Human essays but with slightly higher consistency.

Table 1
Responses to the questionnaire (n = 15).

How ChatGPT was used	Number of Responses
Generated ideas	5
Asked it to write a sample essay for reference	3
Used it for proofreading (correction)	6
Checked the logical structure of the essay	4
Let ChatGPT write everything	2

Note. Some respondents selected more than one option.

Data analysis

As described above, a mini-corpus was created containing essays of three distinct types: Human, ChatGPT, and Mix. To address the three research questions posed in this study, an analysis was conducted using the complexity, accuracy, and fluency (CAF) framework. This study is positioned as a partial replication of the study by Herbold et al. (2023). Utilizing the Python code provided in their publication (https://doi.org/10.5281/zenodo.8343644), we calculated various indicators of syntactic complexity. Additionally, measurements such as lexical diversity, semantic properties, and the use of discourse markers were also analyzed, adhering to the methodologies outlined by Herbold et al.

Table 2 outlines the measures used for the automated linguistic analysis of essays. Seven of the nine measures were sourced from Herbold et al. (2023) and were computed using their Python code, which is accessible in the aforementioned public repository. The measures are as follows (see Herbold et al. for more details):

(1) Lexical diversity (LD): LD assesses vocabulary richness by calculating the ratio of unique words to total words in a text. Herbold et al. used the Measure of Textual Lexical Diversity (MTLD), which is less affected by text length compared to the type-token ratio (McCarthy & Jarvis, 2010). Higher LD scores reflect a broader vocabulary range.
(2) Syntactic complexity (clauses): This measure counts specific clausal structures per sentence, including conjuncts, clausal modifiers of nouns, adverbial clause modifiers, clausal complements, clausal subjects, and parataxes. A greater number of clauses suggests increased syntactic complexity.
(3) Syntactic complexity (depth): Calculated by determining the maximum depth of a sentence's dependency parse tree using Python's spaCy dependency parser. Deeper trees indicate more complex syntactic structures and embeddings, such as embedded relative clauses.
(4) Nominalizations: This counts nouns derived from verbs or adjectives, often ending in -ion, -ment, and -ance, indicative of formal, abstract writing. This study used nominalizations per sentence as a measure.

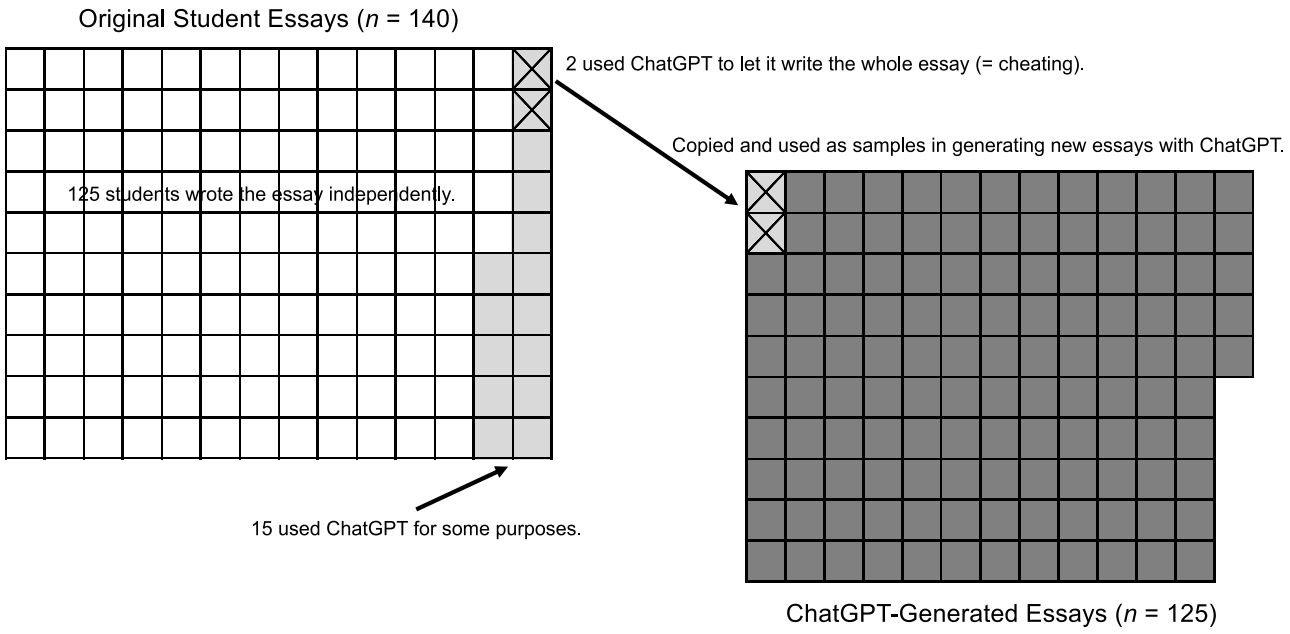

Fig. 1. The Breakdown of the Essay Mini-Corpus Used in the Study
Note. Each cell represents an essay produced either by a learner or ChatGPT. The white cells indicate essays written independently by students without the use of ChatGPT. The cells in gray indicate essays where ChatGPT was used to some extent, while the two cells marked in X (☒) represent cases of inappropriate AI use where two students asked ChatGPT to write the entire essay.

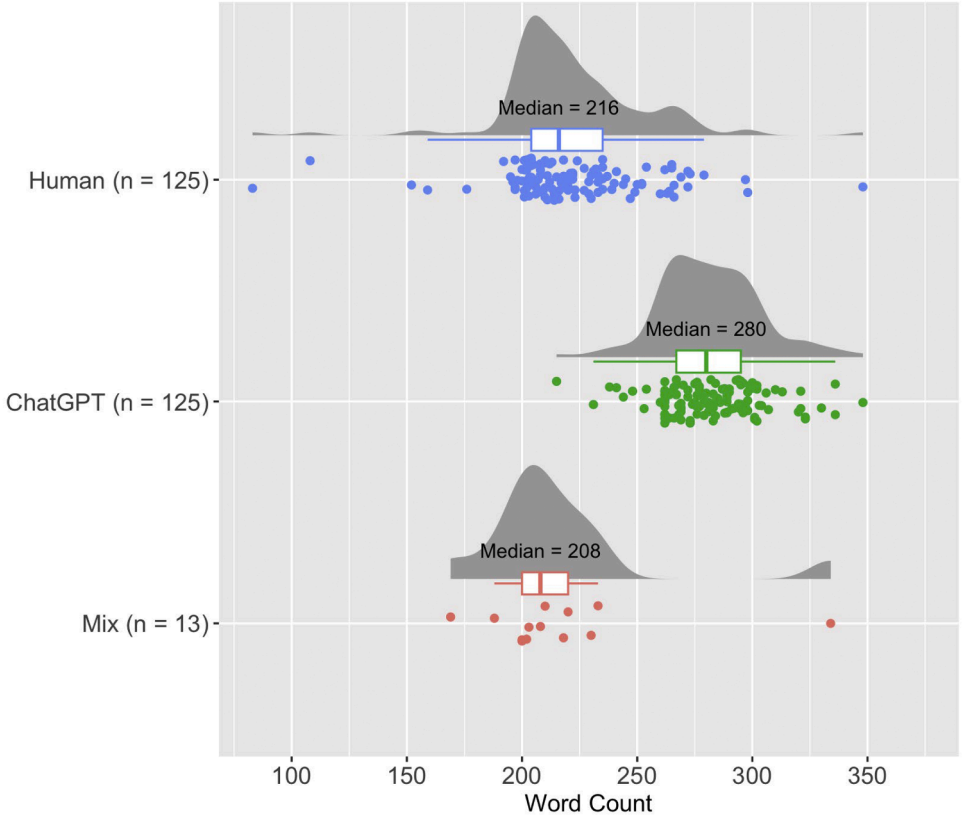

Fig. 2. Distributions of word count for essays by humans, ChatGPT, and partial use of ChatGPT (Mix).

(5) Modals: This measure tallies modal verbs (e.g., can, could, may, might, must, shall, should, will, would) and expressions of modality (e.g., definitely, likely, possibly), reflecting the writer's attitude toward their assertions.

(6) Epistemic markers: These are expressions of the writer's judgments about the certainty or reliability of a statement, such as 'I think' or 'in my opinion.' The frequency of epistemic markers per sentence was measured.

Table 2
Measures used for automated linguistic analysis of essays.

Construct	Measure	Abbreviation
Lexical diversity	1. Lexical diversity	LD
Complexity	2. Syntactic complexity (clauses)	sent_complex_tags
	3. Syntactic complexity (depth)	sent_complex_depth
	4. Nominalizations	nom_per_sent
Semantic property	5. Modals	modals_all
	6. Epistemic markers	ep_per_sent
Discourse property	7. Discourse markers	dm_per_sent
Accuracy*	8. Errors per 100 words	error_per100
Fluency*	9. Total words in an essay	word_count

* Constructs not included in Herbold et al. (2023)'s study.

(7) Discourse markers: Discourse markers are words or phrases (e.g., however, moreover, thus) that signal connections between clauses, enhancing textual cohesion. More discourse markers suggest greater explicit logical coherence. This was measured by discourse markers per sentence.

Fluency was quantified by the total word count, as this measure has been straightforwardly operationalized in L2 writing as the total number of words (Plakans et al., 2019). Accuracy was operationalized as errors per 100 words, in accordance with Mizumoto et al. (2024). Mizumoto et al. demonstrated that this metric can be effectively computed using ChatGPT (GPT-4). Specifically, Research Question 2 (RQ2) investigates whether NLP techniques can accurately identify essays written by learners with the help of ChatGPT. For this purpose, accuracy was calculated exclusively through automated calculations, without manual coding. Scoring the quality of the essays was not included because, while it is possible, it was beyond the scope of this study.

The constructs marked with an asterisk (*) in Table 2, fluency and accuracy, were not included in Herbold et al. (2023) but were added to expand the scope of the analysis to align the CAF framework and to address the research questions posed in this study. The abbreviations are used consistently throughout the study for conciseness and ease of reference.

To address Research Question 1, which focuses on the discernible differences between essays composed by EFL learners and those generated by ChatGPT, we applied statistical tests to examine the differences between the two types (Human vs. ChatGPT) across all nine measures listed in Table 2. Additionally, we employed a random forest model to determine if essays written by learners and ChatGPT can be correctly identified.

For Research Question 2, which explores whether NLP techniques can effectively detect essays written by EFL learners who employ ChatGPT in their writing process, we investigated whether the random forest model could correctly identify the two learners who relied entirely on ChatGPT to produce their essays.

Lastly, to address Research Question 3, which aims to identify the distinctive features in essays written by learners who utilize ChatGPT for editing and proofreading their self-written work, as opposed to those who rely on ChatGPT to compose the entire essay ($n = 2$), we analyzed the linguistic features of the essays written by learners who used ChatGPT ($n = 13$). Furthermore, we examined qualitative differences in the essays written by those learners.

All text analysis in this study were conducted using Python (version 3.8.10). For statistical analysis and visualization, we employed R (version 4.3.1). To ensure reproducibility and transparency in our data analysis (In'nami et al., 2022), all data and Python/R code used in this study are shared on OSF (https://osf.io/bj8kq/).

Results

Fig. 3 displays the inter-correlations among linguistic features, alongside their histograms, as investigated in this study. Some variables exhibited non-normal distributions, prompting the use of Spearman rho correlation (ρ). These correlations ranged from moderate to high, depending on the variable combinations. According to Plonsky and Oswald (2014), effect sizes for correlation coefficients in L2 studies are

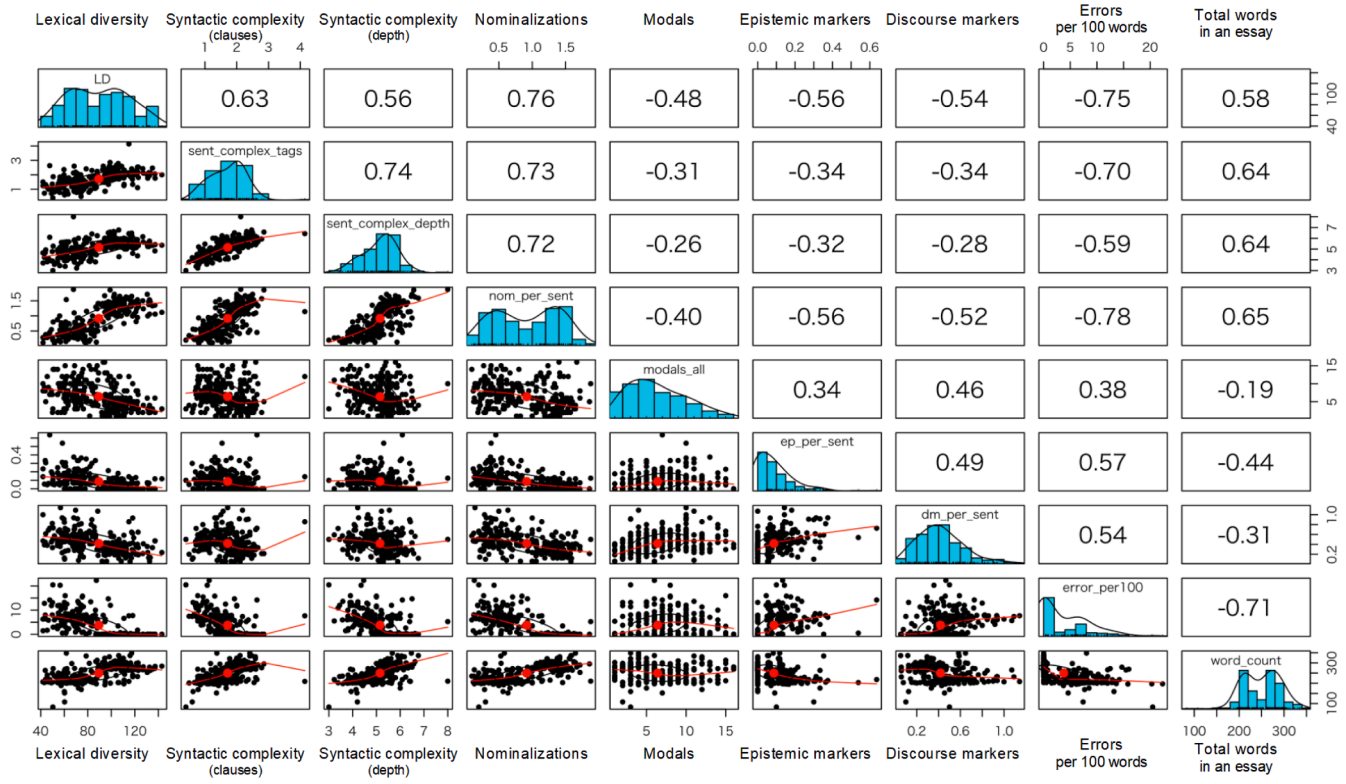

Fig. 3. Correlation matrix with scatter plot of matrices (SPLOM) and histograms.

0.25 (small), 0.40 (medium), and 0.60 (large). Most correlations in our study fell within the medium to large effect sizes. Notably, the correlations involving modals, epistemic markers, and discourse markers (numbered 5, 6, and 7 respectively, as listed in Table 2) with other variables were negative, except for the inter-correlations among themselves. Specifically, these markers showed negative correlations with lexical diversity, syntactic complexity (measured by clauses), syntactic complexity (measured by depth), and nominalizations.

Generally, high negative correlations are found between writing scores (i.e., writing proficiency) and accuracy measures (Polio & Shea, 2014), since fewer errors indicate higher proficiency in writing. In the correlations presented in Fig. 2, the accuracy measure (i.e., errors per 100 words) showed high correlations with measures for lexical diversity and syntactic complexity, indicating theoretically sound relationships. This suggests that the measures for semantic and discourse properties might not be functioning effectively.

Moreover, the same pattern of correlations was observed with fluency (measured by total words in an essay), further supporting the possibility that measures for semantic and discourse properties were not performing as expected. Fluency, particularly total words or word count, is recognized as a predictor of both proficiency and writing level among the three CAF dimensions (Plakans et al., 2019). The role of word count as a robust predictor of L2 writing score variance and overall proficiency is well-established (Crossley et al., 2014; Ferris, 1994; Gebril & Plakans, 2013; Goh et al., 2020), affirming the relationship between word count and L2 writing proficiency. That is, learners who write longer essays generally receive better writing scores. Therefore, the negative correlations between the measures of modals, epistemic markers, and discourse with total words in this study appear illogical.

Interestingly, these inter-correlation patterns (i.e., semantic and discourse properties showing negative correlations with other measures) were consistent with the findings from Herbold et al. (2023), when the correlations among variables were recalculated, as correlations were not reported in the original paper, with the data available on a repository (10.5281/zenodo.8343644).

Research Question 1 aimed to compare essays written by EFL learners with those generated by ChatGPT across nine linguistic features. We first conducted statistical tests to examine the differences between the human and ChatGPT scores for each indicator. The histograms in Fig. 3 indicate that some variables were not normally distributed, necessitating the use of the Mann-Whitney U test. We applied this test to the nine variables, setting the significance level at 0.0056 (0.05/9) using the Bonferroni correction; values below this threshold indicated statistical differences.

Table 3 summarizes the results of these tests, revealing statistical differences in all nine linguistic features, with mostly large effect sizes that indicate substantial differences between the essays of EFL learners and those created by ChatGPT. Notably, the ChatGPT-generated essays exhibited greater lexical diversity, higher syntactic complexity (in two measures), more nominalization, much fewer errors (i.e., higher accuracy), and a greater word count, in comparison with the human-written essays.

Table 3
Comparisons of medians.

Measure	Human	ChatGPT	Effect size *r* [95 % CI]
1. Lexical diversity	69.38	108.44*	0.82 [0.78, 0.86]
2. Syntactic complexity (clauses)	1.35	2.12*	0.72 [0.65, 0.77]
3. Syntactic complexity (depth)	4.75	5.53*	0.62 [0.53, 0.69]
4. Nominalizations	0.50	1.35*	0.83 [0.79, 0.87]
5. Modals	8*	4	0.66 [0.58, 0.72]
6. Epistemic markers	0.13*	0	0.46 [0.36, 0.56]
7. Discourse markers	0.50*	0.29	0.61 [0.53, 0.68]
8. Errors per 100 words	6.88*	0	0.91 [0.89, 0.93]
9. Total words in an essay	216	280*	0.77 [0.72, 0.82]

* $p < .05$ in the Mann-Whitney U test.

It is important to highlight that, as stated above in the results of correlation coefficients, the three measures–modals, epistemic markers, and discourse markers–deviated from the theoretically expected results, compared with lexical diversity, syntactic complexity, accuracy, and fluency. These features were statistically more prevalent in the essays written by learners than in those generated by ChatGPT. This aligns with the findings of Herbold et al. (2023), necessitating a careful interpretation of these findings.

By closely comparing essays written by learners and ChatGPT (see actual essays in the online supplementary material), it became clear that learners tended to use (a) modals to hedge their opinions, (b) epistemic markers to express personal opinions and judgments based on their real-life experiences, and (c) discourse markers to organize their essays and ensure coherence. In contrast, ChatGPT's essays tended to be more factual and less personal, reflecting its training on a restricted set of texts, most of which is primarily from the crawled texts from websites, which does not necessarily include a wide range of registers, particularly not student essays. As Berber Sardinha (2024) notes, AI-generated texts are often register-restricted, and the distinctions it produces are based more on inference than on actual register-specific knowledge. While EFL compositions can be objective, the difference observed here may stem from ChatGPT's attempt to generate text using its built-in knowledge, which is not specifically tuned to the EFL essay register. This suggests that the AI was not adequately exposed to the specific features of EFL student essays during training. The AI's language generation is thus designed to be straightforward and objective, minimizing the use of language that reflected uncertainty or personal opinion unless specifically programmed to mimic such styles. This interpretation is supported by another study, which found that Chinese intermediate English learners outperformed ChatGPT in deep cohesion (Zhou et al., 2023).

To answer Research Question 2, we applied a random forest model to verify how accurately essays written by learners ($n = 125$) and those generated by ChatGPT ($n = 125$) could be classified based on these nine linguistic measures. The classification was 100 % accurate when all nine linguistic features were included (Table 4), and it remained 100 % when excluding total words, which theoretically impacts writing proficiency. Since the random forest algorithm produces varying results with each application, this analysis was double-checked by calculating the average of 2000 random forest model outcomes. With this approach, removing errors per 100 words to check classification accuracy resulted in 98.8 % of essays being correctly identified. Additionally, excluding both total words and errors per 100 words, and classifying based on the remaining seven measures, resulted in 97.8 % accuracy. These results confirm that NLP methods can reliably detect essays written entirely by ChatGPT.

To address Research Question 3, which focused on identifying distinctive features in essays of learners who used ChatGPT for editing and proofreading, we applied the previously mentioned random forest model to 13 learners. As shown in Table 5, two of these learners were classified as "ChatGPT," while the remaining 11 were categorized as "Humans." This classification led to follow-up interviews with the learners "166 (No. 5)" and "258 (No. 13)" identified as "ChatGPT" in Table 5, pseudonymously named Miwa and Nao, respectively. Miwa reported that she initially wrote her essay and then had ChatGPT completely revise it before submission. Nao, on the other hand, first wrote her essay independently but employed ChatGPT to identify and correct all errors and unnatural wording, ultimately submitting the revised version. This resulted in Nao's essay having no errors. While Nao's use of ChatGPT for revisions might appear reasonable, the

Table 4
Result of random forest classification (Confusion matrix).

Category	ChatGPT	Human	Classification Error
ChatGPT	125	0	0
Human	0	125	0

Note. No. of trees: 500; No. of variables tried at each split: 3.

Table 5
Random forest model classification of the learners who allegedly used ChatGPT partially (Mix).

No.	ID	Prediction	No.	ID	Prediction
1	124	Human	8	187	Human
2	129	Human	9	208	Human
3	134	Human	10	212	Human
4	151	Human	11	244	Human
5	166	ChatGPT	12	247	Human
6	176	Human	13	258	ChatGPT
7	183	Human			

extensive editing by ChatGPT removed all errors, producing a document that retained some essence of a learner-written essay. Consequently, these comprehensive revisions, while preserving their original ideas, justified the classification of Miwa and Nao's essays as written by ChatGPT.

Fig. 4 illustrates the distribution of data across nine linguistic features for the Human, ChatGPT, and Mix categories, with the essays of Miwa and Nao marked with bold colored circles in the Mix group. As Fig. 4 clearly shows, there is a tendency for lexical diversity, complexity (measured by two syntactic complexity indicators and nominalization), accuracy (errors per 100 words), and fluency (word count) to be higher for ChatGPT than for essays written by learners, with accuracy indicating fewer errors.

Essays written by learners tend to have higher values in semantic properties (modals and epistemic markers) and discourse properties (discourse markers) compared to ChatGPT. The Mix group, which comprises 13 individuals and thus shows more variation, has a distribution that appears to be a combination of both, likely due to the partial use of ChatGPT, and this difference is reflected in the figure.

Fig. 4 demonstrates that Miwa and Nao's essays are notable for having zero errors per 100 words. While Nao's essay demonstrates higher values for modals, epistemic markers, and discourse markers, suggesting a closer resemblance to the Human category, it was the absence of errors that primarily flagged the essay as ChatGPT-written. In fact, once the accuracy measure of errors per 100 words was excluded, both Miwa and Nao were classified as "Human." This emphasizes the crucial role of incorporating an accuracy measure into the random forest model to ensure precise differentiation between human-written essays and those generated by ChatGPT.

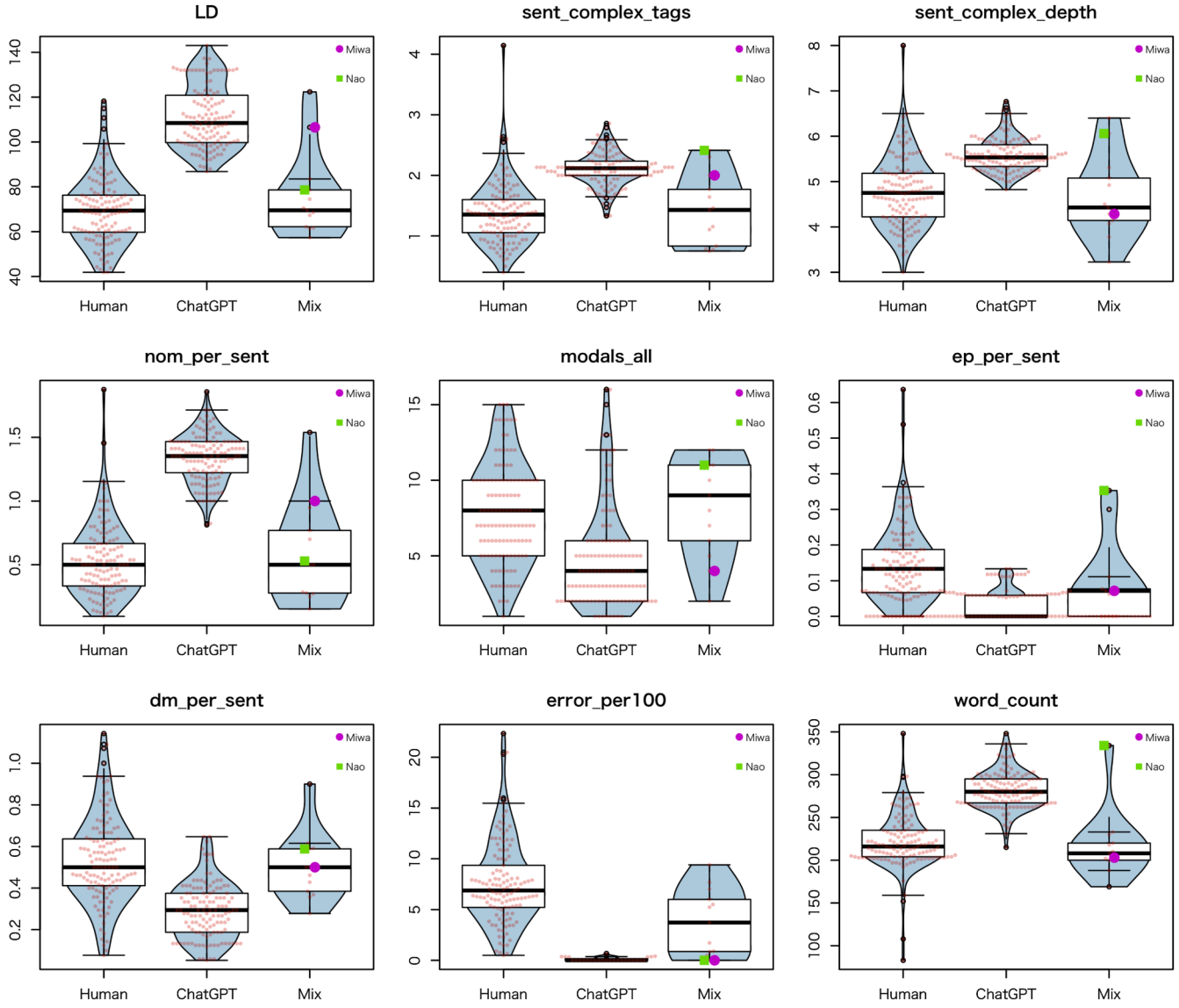

Fig. 4. Comparative analysis of linguistic features in human, ChatGPT, and mix-classified essays
Note. Human (n = 125), ChatGPT (n = 125), and Mix (n = 13). Refer to Table 2 for abbreviations.

Discussion

The present study revealed significant differences between essays written by EFL learners and those generated by ChatGPT across all nine linguistic features examined. ChatGPT-generated essays demonstrated greater lexical diversity, higher syntactic complexity, more nominalization, substantially fewer errors, and higher word counts compared to human-written essays. Conversely, human-written essays exhibited higher usage of modals, epistemic markers, and discourse markers, which was derived from the differences in writing styles and approaches between humans and AI. These results align with those in the original study by Herbold et al. (2023).

By incorporating accuracy and fluency indicators, which were not included in the original study (Herbold et al., 2023), our research demonstrated that a random forest model could achieve 100 % accuracy in classifying essays as human-written or ChatGPT-generated when using all nine linguistic features. Even when excluding word count and error rate, the model maintained a high accuracy of 97.8 %, indicating that the remaining seven measures are primary functional linguistic features for distinguishing between human and AI-generated texts. This is reasonable, considering that the linguistic features used by advanced human writers typically include greater lexical diversity and increasing syntactic complexity (e.g., Barkaoui & Hadidi, 2020; Kyle et al., 2021; Kyle & Crossley, 2017; Mazgutova & Kormos, 2015).

The study also found that NLP techniques could effectively detect essays written entirely by ChatGPT. The random forest model accurately identified the two learners who had their entire essays written by ChatGPT, even when excluding total words and errors per 100 words from the analysis.

For learners who used ChatGPT for editing and proofreading, the study identified distinctive features. Two out of 13 learners who used ChatGPT for editing were classified as "ChatGPT" by the model, while the remaining 11 were categorized as "Human." Errors per 100 words (i.e., accuracy) emerged as the most critical indicator for differentiation, with extensively edited essays showing zero errors. Other important indicators included nominalizations, lexical diversity, and word count. Essays that underwent extensive ChatGPT editing retained some characteristics of learner-written essays while showing ChatGPT-like features in terms of accuracy and other linguistic measures.

The results regarding learners who used ChatGPT for editing and proofreading reveal a complex interplay between human and AI contributions to writing. The fact that some extensively edited essays were classified as ChatGPT-generated, primarily due to their error-free nature, raises important considerations about the appropriate use of AI tools in the writing process. It suggests that while AI can significantly improve the technical aspects of writing, it may also remove characteristics that are indicative of a learner's developmental stage in language acquisition.

One significant limitation of this study concerns the potential underreporting of the use of generative AI tools like ChatGPT by participants. Although participants were reassured that their responses would not affect their course grades and were urged to provide honest answers, the self-reported nature of the data might lead to underreporting due to social desirability bias or concerns over academic honesty. This factor could have influenced the accuracy of the findings regarding the actual use of AI in writing the essays. Future research could employ more controlled experimental designs or use indirect measures of AI use to mitigate this limitation and provide a more accurate picture of how students are using AI in academic writing.

Similar to Herbold et al. (2023), our study found that learners used more modals, epistemic markers, and discourse markers compared to ChatGPT. This is likely because learners strived to express their individual voices toward the topic that is relevant and authentic to college students—the pros and cons of doing a part-time job, whereas ChatGPT may have positioned themselves toward the topic as a third person, resulting in generating impersonal and objective tones. This is because ChatGPT is designed to generate impersonal and objective writing styles. These differences highlight the distinct nature of human versus AI-generated texts. However, it is crucial to note that using these linguistic characteristics typical of ChatGPT does not necessarily indicate "good" writing. Instead, these results offer an opportunity to understand the distinct features of AI and human writing.

Concerns about the generic nature of ChatGPT-generated texts, including the loss of authorial voice, have been noted by learners, emphasizing a significant challenge in using ChatGPT for L2 writing (Zou & Huang, 2023). While model essays can be useful (Nguyen & Le, 2022), texts produced by ChatGPT should only be seen as examples, lacking the unique voice crucial for learners' essays, especially in the context of constructing an argument. Writing instructors must recognize this limitation and refine teaching methods and evaluation of written assignments to ensure each writer can express their unique voice while enhancing the complexity, accuracy, and fluency of their written work. This is especially crucial in an era where AI-assisted writing is becoming increasingly prevalent. Additionally, the evolving criteria of "what constitutes good writing" (Yasuda, 2024) in this context need careful consideration.

To address these issues, teachers should engage students in analyzing the relationship between linguistic choices and their rhetorical impacts on the overall quality of the writer's argument. It might be especially beneficial for novice writers to raise their awareness of the contextual factors that may affect the use of authorial voice markers or the use of more voiceless, neutral, and objective linguistic features. Teachers should also engage students with examples of appropriate AI uses, such as brainstorming, editing, and proofreading, to prevent plagiarism and ensure academic integrity. Emphasizing that AI contributions are detectable through NLP techniques, as reported in this study, may deter misuse. It is also important to shift the focus from the final written product to the writing process, particularly for learners with lower proficiency or motivation, to discourage mere copying. For example, in this study, essays extensively edited to eliminate all errors, like those of Miwa and Nao, were classified as ChatGPT-generated when evaluated as a single product. While this approach may not be acceptable in a single-product assessment, focusing on the process through methods like portfolios, which facilitate the noticing of corrections, could potentially enhance writing skills.

Furthermore, incorporating languaging (Suzuki & Storch, 2020) into revisions can be beneficial. This involves learners explaining and discussing their revisions with ChatGPT, including their rationale and any disagreements with suggested changes. Integrating these discussions into their submissions encourages active engagement rather than passive reliance on AI, fostering greater learner autonomy and self-regulation.

While debates about whether ChatGPT is a "friend or foe" in language learning and teaching continue (Evmenova et al., 2024), its integration into language education, like that of electronic dictionaries, Google, and machine translation technologies such as DeepL, seems inevitable. Therefore, we should focus on developing better ways to use it and equip learners with its proper, effective use. As Godwin-Jones (2024) notes, "the power and versatility of AI tools are likely to turn them into constant companions in many people's lives, creating a close connection that goes beyond simple tool use" (p. 5). Moreover, considering ecological validity, access to such resources is crucial (Pusey & Butler, 2023).

Since institutional policies may not perfectly align with language learning and teaching practices regarding ChatGPT use, it is essential for language teachers and learners to share a common understanding. This shared understanding is crucial when incorporating GenAI like ChatGPT into language education. Teachers should proactively communicate policies to learners upfront and model appropriate use to avoid undesirable outcomes such as learning outsourcing or academic misconduct. Rather than seeking answers from others, classroom teachers should take the initiative in considering usage methods, sharing policies with

learners, and allowing practical application.

Finally, to implement these practices in the classroom, enhancing teacher AI literacy is essential (Ding et al., 2024). Language teacher education should include opportunities to develop AI literacy (Moorhouse & Kohnke, 2024), and teachers need to explicitly instruct learners in this area. Such instruction in metacognitive resource use (Mizumoto, 2023), including ChatGPT, can be implemented and is predicted to be effective. Reflecting on this, Godwin-Jones (2024) emphasizes the importance of "critical AI literacy," through which both learners and teachers gain agency. This shared agency, developed through the use of GenAI by both teachers and learners, is likely to become the de facto standard in modern language education, enabling both parties to thrive in this new technological landscape.

CRediT authorship contribution statement

Atsushi Mizumoto: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sachiko Yasuda:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Conceptualization. **Yu Tamura:** Writing – review & editing, Writing – original draft, Validation, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The author whose name is listed immediately below certify that he has NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Acknowledgment

In the preparation of this manuscript, we employed ChatGPT (GPT-4) and Claude 3.5 Sonnet to enhance the clarity and coherence of the language, ensuring it adheres to the standards expected in scholarly journals. While ChatGPT played a role in refining the language, it did not contribute to the generation of any original ideas. The authors alone are responsible for any inaccuracies present in the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.acorp.2024.100106.

References

Barkaoui K., & Hadidi A. (2020). Assessing change in second language writing performance. Routledge.

Barrot, J.S., 2023. Using ChatGPT for second language writing: pitfalls and potentials. Assess. Writ. 57, 100745. https://doi.org/10.1016/j.asw.2023.100745.

Berber Sardinha, T., 2024. AI-generated vs human-authored texts: a multidimensional comparison. Appl. Corpus Linguist. 4 (1), 100083. https://doi.org/10.1016/j.acorp.2023.100083.

Berriche, L., Larabi-Marie-Sainte, S., 2024. Unveiling ChatGPT text using writing style. Heliyon 10 (12), e32976. https://doi.org/10.1016/j.heliyon.2024.e32976.

Casal, J.E., Kessler, M., 2023. Can linguists distinguish between ChatGPT/AI and human writing?: a study of research ethics and academic publishing. Res. Methods Appl. Linguist. 2 (3), 100068. https://doi.org/10.1016/j.rmal.2023.100068.

Crossley, S.A., Kyle, K., Allen, L.K., Guo, L., McNamara, D.S., 2014. Linguistic microfeatures to predict L2 writing proficiency: a case study in automated writing evaluation. J. Writ. Assess. 7 (1). https://escholarship.org/uc/item/06n1v820.

Crosthwaite, P., Baisa, V., 2023. Generative AI and the end of corpus-assisted data-driven learning? Not so fast. Appl. Corpus Linguist. 3 (3), 100066. https://doi.org/10.1016/j.acorp.2023.100066.

Derakhshan, A., Ghiasvand, F., 2024. Is ChatGPT an evil or an angel for second language education and research? A phenomenographic study of research-active EFL teachers' perceptions. Int. J. Appl. Linguist. 12561. https://doi.org/10.1111/ijal.12561.

Desaire, H., Chua, A.E., Isom, M., Jarosova, R., Hua, D., 2023. Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. Cell Rep. Phys. Sci. 4 (6), 101426. https://doi.org/10.1016/j.xcrp.2023.101426.

Ding, A.C.E., Shi, L., Yang, H., Choi, I., 2024. Enhancing teacher AI literacy and integration through different types of cases in teacher professional development. Comput. Educ. Open 6, 100178. https://doi.org/10.1016/j.caeo.2024.100178.

Evmenova, A.S., Borup, J., Shin, J.K., 2024. Harnessing the power of generative AI to support all learners. TechTrends. https://doi.org/10.1007/s11528-024-00966-x.

Ferris, D.R., 1994. Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. TESOL Q. 28 (2), 414. https://doi.org/10.2307/3587446.

Ferris, D.R., 2010. Second language writing research and written corrective feedback in SLA: intersections and practical applications. Stud. Second. Lang. Acquis. 32 (2), 181–201. https://doi.org/10.1017/S0272263109990490.

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S.D., Köller, O., Möller, J., 2024. Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. Comput. Educ. Artif. Intell. 100209. https://doi.org/10.1016/j.caeai.2024.100209.

Flower, L., Hayes, J.R., 1981. A cognitive process theory of writing. Coll. Compos. Commun. 32 (4), 365–387. https://doi.org/10.2307/356600.

Foung, D., Lin, L., Chen, J., 2024. Reinventing assessments with ChatGPT and other online tools: opportunities for GenAI-empowered assessment practices. Comput. Educ. Artif. Intell. 6, 100250. https://doi.org/10.1016/j.caeai.2024.100250.

Gebril, A., Plakans, L., 2013. Toward a transparent construct of reading-to-write tasks: the interface between discourse features and proficiency. Lang. Assess. Q. 10 (1), 9–27. https://doi.org/10.1080/15434303.2011.642040.

Godwin-Jones, R., 2024. Distributed agency in second language learning and teaching through generative AI. Lang. Learn. Technol. 28 (2). https://hdl.handle.net/10125/73570.

Goh, T.T., Sun, H., Yang, B., 2020. Microfeatures influencing writing quality: the case of Chinese students' SAT essays. Comput. Assist. Lang. Learn. 33 (4), 455–481. https://doi.org/10.1080/09588221.2019.1572017.

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., Trautsch, A., 2023. A large-scale comparison of human-written versus ChatGPT-generated essays. Sci. Rep. 13 (1), 18617. https://doi.org/10.1038/s41598-023-45644-9.

Ibrahim, K., 2023. Using AI-based detectors to control AI-assisted plagiarism in ESL writing: "The terminator versus the machines. Lang. Test. Asia 13 (1), 46. https://doi.org/10.1186/s40468-023-00260-2.

In'nami, Y., Mizumoto, A., Plonsky, L., Koizumi, R., 2022. Promoting computationally reproducible research in applied linguistics: recommended practices and considerations. Res. Methods Appl. Linguist. 1 (3), 100030. https://doi.org/10.1016/j.rmal.2022.100030.

Kar, S.K., Bansal, T., Modi, S., Singh, A., 2024. How sensitive are the free AI-detector tools in detecting AI-generated texts? A comparison of popular AI-detector tools. Indian J. Psychol. Med. https://doi.org/10.1177/02537176241247934.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Kasneci, G., 2023. ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individ. Differ. 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274.

Kyle, K., Crossley, S.A., 2017. Assessing syntactic sophistication in L2 writing: a usage-based approach. Lang. Test. 34 (4), 513–535. https://doi.org/10.1177/0265532217712554.

Kyle, K., Crossley, S., Verspoor, M., 2021. Measuring longitudinal writing development using indices of syntactic complexity and sophistication. Stud. Second Lang. Acquis. 43 (4), 781–812. https://doi.org/10.1017/S0272263120000546.

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

Mazgutova, D., Kormos, J., 2015. Syntactic and lexical development in an intensive english for academic purposes programme. J. Second Lang. Writ. 29, 3–15. https://doi.org/10.1016/j.jslw.2015.06.004.

McCarthy, P.M., Jarvis, S., 2010. MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. Behav. Res. Methods 42 (2), 381–392. https://doi.org/10.3758/BRM.42.2.381.

Mizumoto, A., 2023. Data-driven learning meets generative AI: introducing the framework of metacognitive resource use. Appl. Corpus Linguist. 3 (3), 100074. https://doi.org/10.1016/j.acorp.2023.100074.

Mizumoto, A., Eguchi, M., 2023. Exploring the potential of using an AI language model for automated essay scoring. Res. Methods Appl. Linguist. 2 (2), 100050. https://doi.org/10.1016/j.rmal.2023.100050.

Mizumoto, A., Shintani, N., Sasaki, M., Teng, M.F., 2024. Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. Res. Methods Appl. Linguist. 3 (2), 100116. https://doi.org/10.1016/j.rmal.2024.100116.

Moorhouse, B.L., Kohnke, L., 2024. The effects of generative AI on initial language teacher education: the perceptions of teacher educators. System 122, 103290. https://doi.org/10.1016/j.system.2024.103290.

Nguyen, L.Q., Le, H.V., 2022. Improving L2 learners' IELTS task 2 writing: the role of model essays and noticing hypothesis. Lang. Test. Asia 12 (1), 58. https://doi.org/10.1186/s40468-022-00206-0.

Pecorari, D., 2023. Generative AI: same same but different? J. Second. Lang. Writ. 62, 101067. https://doi.org/10.1016/j.jslw.2023.101067.

Plakans, L., Gebril, A., Bilki, Z., 2019. Shaping a score: complexity, accuracy, and fluency in integrated writing performances. Lang. Test. 36 (2), 161–179. https://doi.org/10.1177/0265532216669537.

Plonsky, L., Oswald, F.L., 2014. How big is "big"? Interpreting effect sizes in L2 research. Lang. Learn. 64 (4), 878–912. https://doi.org/10.1111/lang.12079.

Polio, C., Shea, M.C., 2014. An investigation into current measures of linguistic accuracy in second language writing research. J. Second Lang. Writ. 26, 10–27. https://doi.org/10.1016/j.jslw.2014.09.003.

Pusey, K., Butler, Y.G., 2023. Investigating the ecological validity of second language writing assessment tasks. System 119, 103174. https://doi.org/10.1016/j.system.2023.103174.

Ranade, N., Eyman, D., 2024. Introduction: composing with generative AI. Comput. Compos. 71, 102834. https://doi.org/10.1016/j.compcom.2024.102834.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., Olson, C.B., 2024. Comparing the quality of human and ChatGPT feedback of students' writing. Learn. Instr. 91, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894.

Su, Y., Lin, Y., Lai, C., 2023. Collaborating with ChatGPT in argumentative writing classrooms. Assess. Writ. 57, 100752. https://doi.org/10.1016/j.asw.2023.100752.

Suzuki W., & Storch N. (Eds.). (2020). Languaging in language learning and teaching: a collection of empirical studies (Vol. 55). John Benjamins. 10.1075/lllt.55.

Tate, T.P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., Warschauer, M., 2024. Can AI provide useful holistic essay scoring? Comput. Educ. Artif. Intell. 100255. https://doi.org/10.1016/j.caeai.2024.100255.

UNESCO, 2023. Guidance for generative AI in education and research. UNESCO. https://doi.org/10.54675/EWZM9535.

Wang, Q., Gayed, J.M., 2024. Effectiveness of large language models in automated evaluation of argumentative essays: finetuning vs. zero-shot prompting. Comput. Assist. Lang. Learn. 1–29. https://doi.org/10.1080/09588221.2024.2371395.

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., Tate, T., 2023. The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. J. Second Lang. Writ. 62, 101071. https://doi.org/10.1016/j.jslw.2023.101071.

Williams, J.D., 2003. Preparing to Teach writing: Research, theory, and Practice, 3rd ed. Routledge. https://doi.org/10.4324/9781410607461.

Woo, D.J., Susanto, H., Yeung, C.H., Guo, K., Fung, A.K.Y., 2024. Exploring AI-Generated text in student writing: how does AI help? Lang. Learn. Technol. 28 (2), 183–209. https://hdl.handle.net/10125/73577.

Yamashita, T., 2024. An application of many-facet Rasch measurement to evaluate automated essay scoring: a case of ChatGPT-4.0. Res. Methods Appl. Linguist. 3 (3), 100133. https://doi.org/10.1016/j.rmal.2024.100133.

Yang, L., Li, R., 2024. ChatGPT for L2 learning: current status and implications. System 124, 103351. https://doi.org/10.1016/j.system.2024.103351.

Yasuda, S., 2024. Does "more complexity" equal "better writing"? Investigating the relationship between form-based complexity and meaning-based complexity in high school EFL learners' argumentative writing. Assess. Writ. 61, 100867. https://doi.org/10.1016/j.asw.2024.100867.

Zhou, T., Cao, S., Zhou, S., Zhang, Y., He, A., 2023. Chinese intermediate English learners outdid ChatGPT in deep cohesion: evidence from English narrative writing. System 118, 103141. https://doi.org/10.1016/j.system.2023.103141.

Zou, M., Huang, L., 2023. The impact of ChatGPT on L2 writing and expected responses: voice from doctoral students. Educ. Inf. Technol. https://doi.org/10.1007/s10639-023-12397-x.