Potentials and Limitations of AI as an Academic Writing Assistant for Non-Native English Speakers: An Assessment Framework for ChatGPT-Generated Texts

Jaime Paster[1]

Received: September 13, 2024	Revised: February 26, 2025	Accepted: February 28, 2025

Abstract

The use of Artificial Intelligence (AI) tools like ChatGPT-4 in academic writing has gained much attention, especially among non-native English speakers. This study looks at the strengths and weaknesses of ChatGPT-4 as a writing assistant by introducing a framework that uses a scoring rubric and software tools. The results show that ChatGPT-4 can produce well-organized, accurate, and relevant academic content. However, it still struggles with improving readability for non-native speakers and ensuring sources are trustworthy. The evaluation process, which included Flesch Reading Ease (FRE) scores and similarity checks, highlights the need for careful review when using AI-generated content for academic work. The proposed framework provides a clear method to make the most of ChatGPT-4's strengths, but some challenges remain. These include the subjective nature of rubric scoring, the shallow focus of FRE scores, possible biases in software tools, and the fact that the study was done in a controlled and unchanging environment rather than reflecting real-world situations where user interactions and needs can vary. Despite these issues, the findings show that non-native English speakers can use ChatGPT-4 effectively for academic writing when combined with proper evaluation methods, ensuring their work is both high-quality and credible.

Keywords: Academic Writing; AI in Education; Generative AI; ChatGPT-Assisted Writing
Type of Article: Research Article

[1]Language Institute, Nakhon Pathom Rajabhat University
Nong Pak Long, Muang, Nakhon Pathom 73000, Thailand
Corresponding Author Email: jaimeli@webmail.npru.ac.th

Introduction

On November 30, 2022, OpenAI released ChatGPT, a chatbot that quickly became a disruptive force in various sectors, including academia. ChatGPT, a Large Language Model (LLM), can generate text by mimicking the statistical patterns of language from an extensive dataset compiled from the internet (Stokel-Walker, 2023). Within a short period, it became a subject of widespread discussion, marked by its controversial debut as a co-author of academic articles, with at least four published papers or preprints listing ChatGPT as a co-author. This rise of AI-driven text generation tools has prompted significant debate about the future of university essays and research production.

The use of LLMs like ChatGPT in academic writing requires careful consideration of several factors, including tone, validity, and reliability, to avoid ethical issues such as bias, discrimination, and misinformation. The UNESCO World Commission on the Ethics of Scientific Knowledge and Technology (2019) emphasizes the need for caution when using AI, especially in culturally and ethically sensitive domains like education and communication. This need for caution extends to AI-generated texts used for academic writing, where the potential for misleading or biased content is particularly concerning.

This article explores the potential of AI tools like ChatGPT for use by non-native English speakers as assistants in academic writing. Studies have shown that non-native English speakers often find writing software helpful for improving their academic writing skills. For instance, Fang (2010) examined the perceptions of EFL learners using a computer-assisted writing program called MyAccess in a Taiwanese college composition class. The study found that majority of the participants had a favorable attitude toward using the tool, citing its benefits in enhancing their writing skills. Similarly, Kee, Razali, Samad, and Noordin (2020) found that ESL teacher trainees showed increased motivation and improved writing performance through the use of another writing software. Although the technologies examined in the previously mentioned studies differ from ChatGPT, they share a common purpose as non-human writing assistants used to aid academic writing. These findings set a precedent for the usefulness of computer-based tools in academic writing.

In light of ChatGPT's growing prevalence and widespread accessibility, its increasing significance as a writing tool among diverse users is evident, demonstrating transformative potential across various academic sectors (Kohnke, Moorhouse, & Zou et al., 2023). However, to fully understand its implications for academic writing, it is crucial to assess and evaluate the texts it generates. Bowman (2022) emphasizes this need, citing a data scientist's observation about ChatGPT's tendency to use an authoritative tone when generating fabricated information. Therefore, careful evaluation is essential to explore both the potential benefits and limitations of ChatGPT. This article seeks to provide a comprehensive understanding of how ChatGPT can be responsibly employed as an academic writing assistant for non-native English speakers, while also outlining the precautions necessary for its ethical integration into the academic writing process.

Research Objectives

To achieve this, the study focuses on the following assessment objectives: Evaluate the effectiveness of ChatGPT-4 in generating coherent and relevant academic texts across various types of prompts. Analyze the impact of prompt specificity on the readability and accessibility of ChatGPT-4's generated texts for non-native

English speakers. Assess the improvements in readability of ChatGPT-4's outputs through iterative prompting and examine how these improvements affect user engagement and comprehension.

Conceptual Framework

To understand the nature, quality, and usefulness of the information generated by ChatGPT, this review viewed ChatGPT-4 starting as a chatbot in the context of other AI models including the overview of generative-AI's mechanism and applications. And finally narrowed down to the assessment process and criteria used to assess the quality and validity of the texts generated by ChatGPT-4. These will allow deeper and objective evaluation.

1. ChatGPT as Artificial Intelligence

ChatGPT is a language model developed by OpenAI, and it can be used for a variety of natural language processing tasks, including proofreading and copyediting (OpenAI, 2023). Moreover, ChatGPT as a chatbot has a remarkable ability to understand and respond in a naturalistic way to language input making it useful for a variety of tasks, especially in the field of higher education. It can be trained to do specific tasks, like completing a sentence you have started or answering questions. ChatGPT can be used to help you with your writing, acting as your research assistant, offering individualized feedback, and helping you communicate with others more effectively (Atlas, 2023). In real-world use, chatbots like ChatGPT are intelligent conversational computer programs designed to simulate conversation with human users and used over the Internet. These Artificial Intelligence technologies are designed using machine learning and deep neural networks, to perform tasks such as customer service, information retrieval, and entertainment. These chatbots can interact with users through text, voice, or other means, and they can be integrated into websites, messaging platforms, or mobile apps (Jia, 2003; Bala, Kumar, Hulawale, & Pandita, 2017; Ayanouz, Abdelhakim, & Benhmed, 2020; Sojasingarayar, 2020). However, despite these models being trained on extensive amounts of human generated data from the internet to mimic human-like responses it is important to keep in mind that this is a direct result of the system's design. A design that works by maximizing the similarity between outputs and the dataset the models were trained on, allowing the potential outputs to be inaccurate, untruthful, and otherwise misleading at times (OpenAI, 2025).

2. AI Tools for Non-Native English Speakers in Academic Writing

Research shows that AI tools, including ChatGPT, can significantly aid non-native English speakers in academic writing. Non-native speakers often face challenges related to grammar, coherence, and lexical choice. AI-based writing assistants like ChatGPT can provide immediate, targeted feedback on grammar, vocabulary, and sentence structure, which helps non-native speakers to improve the clarity and accuracy of their writing (Del Giglio & Da Costa, 2023; Hwang, Lim, Lee, Matsui, Iguchi, Hiraki, & Ahn, 2023; Li, Zong, Wu, Wu, Peng, Zhao, Yang, Xie, & Shen, 2024). These tools serve as a bridge, helping users refine their writing skills through interactive dialogue and suggestions that enhance their understanding of complex linguistic nuances. However, the effectiveness of AI tools for non-native speakers is not without its challenges. While ChatGPT can enhance the clarity and coherence of academic texts, its outputs may still require careful review to ensure relevance and contextual appropriateness (Oviedo-Trespalacios, Peden, Cole-Hunter, Costantini, Haghani, Rod, Kelly,

Torkamaan, Tariq, Newton, Gallagher, Steinert, Filtness, & Reniers, 2023). Moreover, AI-generated texts sometimes lack the deeper critical analysis that academic writing often requires. And, while ChatGPT provides a valuable resource for non-native speakers, its integration into academic writing practices should be approached with a clear understanding of its limitations and potential biases, ensuring independent verification of its outputs by and expert of the topic being discussed by ChatGPT (Azaria, Azoulay, & Reches, 2024)

3. The Need for Structured Assessment of ChatGPT-generated Information

It has been documented in studies that ChatGPT is designed to interact in a conversational way by answering follow-up questions, and challenging incorrect premises, with high context sensibility (Atlas, 2023; Bansal, Chamola, Hussain, Guizani, & Niyato, 2024). Another study by Gilson, Safranek, Huang, Socrates, Chi, Taylor, and Chartash (2023) evaluated the performance of ChatGPT on questions within the scope of the United States Medical Licensing Examination (USMLE). The results showed that ChatGPT was able to generate conversation-style responses that were relevant and helpful in improving medical student learning about neurological localization. In addition to this, the quality of outputs generated by ChatGPT has been assessed by Kung, Cheatham, Medenilla, Sillos, Leon, Elepaño, Madriaga, Aggabao, Diaz-Candido, Maningo, and Tseng (2023), and stated that ChatGPT displayed comprehensible reasoning and valid clinical insights, lending increased confidence to trust and explainability. A study that suggests large language models such as ChatGPT may potentially assist human learners in a medical education setting, as a prelude to future integration into clinical decision-making. In contrast, Teresa Kubacka, a data scientist based in Zurich, Switzerland, found that this language model lies with confidence. Despite its authoritative tone, there have been instances in which ChatGPT won't tell you when it does not have the answer. She experimented with the language model by asking it about a made-up physical phenomenon (Bowman, 2022). These instances are not foreign to the creators of this generative-AI, as documented in the limitations declared in ChatGPT's user page is the clear warning that this AI may occasionally generate incorrect information, thus may occasionally produce harmful instructions or biased content and a limited knowledge of the world and events after 2021 (OpenAI, 2023). And on this note, it might be necessary to examine the research of Giray (2023) that stresses the role of inputting the most appropriate prompts to improve the quality of output generated by a Large Language Models like ChatGPT. The works of Nazari and Saadi (2024) also stressed that, while ChatGPT can produce clear and grammatically correct output, it is crucial to provide the appropriate prompts and verify the accuracy of the information it generates and consider other verification methods. Therefore, with this mechanism and possibilities laid out, this study will observe proper prompting by giving concise and the most accurate prompts possible and pursue with a structured assessment the generated texts outputs.

4. Criterion-based Assessment of ChatGPT-generated Texts

To systematically evaluate ChatGPT-generated texts for academic writing, this study employs a combination of software tools that measure readability and similarity to existing databases, alongside an analytical rubric (see Table 1) grounded in academic writing assessment literature. The key criteria adapted for this evaluation include readability, technical accuracy, coherence, relevance, and information validity: (1) Readability: Assessed using the Flesch Reading Ease (FRE) Score, which measures how easy a text is to read. Higher scores indicate better readability and accessibility for a broader audience (Flesch, 1948). (2) Technical

Accuracy: Focuses on spelling, punctuation, and grammar, which are critical for effective writing and perceived competence (Daffern, Mackenzie, & Hemmings, 2017; Romano, 2019; Pan, Rickard, & Bjork, 2021). (3) Coherence: Evaluates the logical flow and organization of ideas, ensuring the text is easy to follow (Johns, 1986; Aminovna, 2022). (4) Relevance: Ensures that the content directly relates to the central focus or argument, avoiding irrelevant or misleading information (Singhal, 2004; Khazaal, 2019). (5) Information Validity: Focuses on maintaining academic integrity and credibility by ensuring that the information is accurate and verifiable (Metzger, Flanagin, & Lara Zwarun, 2003; Ahmed & Ishtiaq, 2021).

By focusing on these criteria, a systematic and grounded evaluation is enabled for proper assessment of the quality and suitability of ChatGPT-generated texts for academic use. This framework ensures that the content is readable, technically accurate, coherent, relevant, and based on valid information, aligning with the standards of academic writing, which prioritize clarity, precision, and rigor (Berkeley, King-Sears, Vilbas, & Conklin, 2016; Corcoran & Ahmad, 2016; Xu, Ellis, & Umphrey, 2019; Adhariani & du Toit, 2020). This literature review highlights the potential of ChatGPT as an aid in academic writing, particularly for non-native English speakers, while acknowledging the limitations inherent in generative AI technologies. Therefore, this study proposes an assessment framework (see Figure 1), focusing on readability, technical accuracy, coherence, relevance, and information validity, that aims to provide a structured approach in assessing ChatGPT-generated texts. Such a framework is crucial for guiding responsible integration of AI tools in academic writing and ensuring that user perceptions are informed by evidence-based assessments.

Methodology

This study employs a mixed-methods approach to assess the quality and usability of texts generated by ChatGPT-4 for academic purposes. The methodology integrates both quantitative and qualitative techniques to provide a comprehensive evaluation of text output, ensuring its technical accuracy, coherence, relevance, information validity, readability, and originality. The combination of an analytical rubric, readability metrics, and plagiarism detection software offers a robust framework for analyzing the academic suitability of AI-generated content.

1. Rubric-based Assessment Tool

The primary qualitative tool for this study is an analytical rubric (see Table 1) designed to evaluate text output based on four key criteria: technical accuracy, coherence, relevance, and information validity. Each criterion is rated on a scale from 0 to 5, with detailed descriptors provided for each level to ensure consistent evaluation. The rubric is applied to assess multiple samples of text generated by ChatGPT-4, with the aim of providing an objective evaluation that informs user perceptions of the generated content.

Table 1

Assessment Rubrics

Criteria	Level	Description
Technical	5	No spelling, punctuation, or grammar errors. Demonstrates superior writing skills and effective communication.
Accuracy	4	Minor spelling, punctuation, or grammar errors that do not impede clarity.
	3	Noticeable errors in spelling, punctuation, or grammar that occasionally affect clarity.

Criteria	Level	Description
	2	Frequent errors in spelling, punctuation, or grammar that often hinder clarity.
	1	Many errors in spelling, punctuation, or grammar that make understanding difficult.
	0	Pervasive errors that render the text unintelligible.
Coherence	5	Text is exceptionally well-organized with a clear logical flow, making it easy to follow and comprehend.
	4	Text is mostly well-organized with minor disruptions in logical flow.
	3	Some organizational issues and logical disruptions, though the main points are still understandable.
	2	Poor organization and frequent disruptions in logical flow, affecting understanding.
	1	Disconnected ideas and lack of logical flow make the text difficult to follow.
	0	No coherence, the text is confusing and lacks logical order.
Relevance	5	All content is directly relevant to the central focus, enhancing the main idea.
	4	Most content is relevant, with minor deviations that do not distract from the main idea.
	3	Some content is relevant, but there are noticeable deviations that partially distract from the main idea.
	2	Much of the content is irrelevant or tangentially related, which confuses the main idea.
	1	Most content is irrelevant to the central topic or argument, severely affecting the main idea.
	0	Content is entirely irrelevant and fails to support the main idea.
Information Validity	5	All information is accurate and well-supported by credible sources, maintaining high academic integrity.
	4	Most information is accurate with minor issues in source credibility.
	3	Some information is accurate, but there are several inaccuracies or less reliable sources.
	2	Frequent inaccuracies or questionable sources that undermine the validity of the information.
	1	Predominantly inaccurate information with unreliable sources, compromising academic integrity.
	0	Completely inaccurate information with no credible sources, violating academic integrity.

To ensure the content validity of the rubric, expert judgment was employed with input from two university lecturers—one with a master's in linguistics and the other in English teaching. These experts assessed the rubric's relevance and comprehensiveness in capturing the necessary components of academic writing. In addition to their expertise in English writing, both lecturers have extensive experience using ChatGPT.

2. Software-based Assessment Tool

To objectively assess the readability of the texts generated by ChatGPT, the Flesch Reading Ease (FRE) scores were calculated from Microsoft Word 2021, a software used in many other studies involving text readability due to its accessibility and convenience (Bothun, Feeder, & Poland, 2021; Rori, Olii, & Rettob, 2021). This analytical tool evaluates the readability of the text by examining several key metrics. Specifically, it calculates the number of words and sentences, as well as the average number of syllables per word (Flesch, 1948). By analyzing these elements, the readability analyzer determines how easily a text can be understood by the average reader. Higher FRE scores indicate texts that are easier to read, while lower scores suggest more complex and challenging content. This step is crucial for ensuring that the generated text meets the desired readability standards and is accessible to the intended audience (Courtis & Hassan, 2002; Mcinnes & Haglund, 2011). Lastly, Turnitin software was used to generate the similarity report. To verify the sources of information used by ChatGPT in producing the text outputs, the study employs Turnitin plagiarism detection software and utilizes its similarity report. Generated texts are submitted to Turnitin, and a similarity report is generated. This report highlights any matching text from various sources, giving the researcher the opportunity to check for possible plagiarism and at the same time allow manual location of the linked sources for further verification.

3. Dynamic Information Retrieval

The study also uses the WebChatGPT extension to get current information from the web. This browser extension allows ChatGPT to access up-to-date information online, making the interactions more relevant to recent events and needs. The sources listed are then verified by the researcher.

4. Prompting Strategy

Informed by the practice that effective generative-AI use requires well-designed prompts (Giray, 2023; Nazari & Saadi, 2024). This study uses a systematic approach in structuring the prompts used by incorporating the key elements, such as: context, task, and specificity. This structure includes providing background information, outlining the desired outcome, and specifying any constraints or style requirements. Following this strategy ensures that the AI produces clear, coherent, and contextually relevant responses, making it particularly useful for non-native English speakers looking to enhance their academic writing.

The texts used in prompting Output numbers 1-6 were selected through a convenience sampling method. The decision to select these particular outputs was based on their relevance to the objectives of the study, which sought to explore how ChatGPT can assist non-native English speakers in enhancing their academic writing. Each output represents a distinct step in the process of improving writing, such as simplifying vocabulary or paraphrasing content. The use of these outputs was purposeful to demonstrate a range of writing improvements, aligned with the study's focus on providing clear, accessible writing for non-native speakers.

Table 2

Prompts Used in Generating Outputs in ChatGPT-4

Output No.	Prompts
1	As an academic writer write a concise paragraph about academic writing.
2	Expound: +the previously generated outputs
3	Rewrite this output using simpler vocabulary: +the previously generated outputs
4	Paraphrase concisely using a simpler vocabulary: +the previously generated outputs
5	Rewrite using very simple vocabulary that a non-native English speaker can read with ease: +the previously generated outputs
6	Rewrite to make it readable to the level of a grade school student: +the previously generated outputs

*ChatGPT, personal communication, August 9, 2024

The data collection for this study was conducted in a single day on the 9th of August 2024 to ensure consistency and maintain the logical flow of the outputs generated by ChatGPT-4. Conducting the data collection in one session reduced potential variations in the model's responses that could arise from external updates or changes in its algorithm over time, thereby enhancing the reliability of the analysis.

5. The Framework for Assessing ChatGPT-generated Texts for Academic Writing

Using a combination of analytical rubric and software. This framework is designed to systematically assess key dimensions of text quality, including readability, technical accuracy, coherence, relevance, and information validity.Drawing on established metrics and scholarly criteria from academic writing literature, the framework aims to provide a structured and objective method for evaluating the utility of AI-generated content, particularly in academic contexts. By focusing on these critical dimensions, the framework not only ensures that the generated texts meet the high standards required in academic writing but also facilitates a more informed and responsible integration of AI tools like ChatGPT into academic practices. This approach is essential for determining the extent to which AI can support non-native English speakers and other users in producing clear, accurate, and contextually appropriate academic texts.

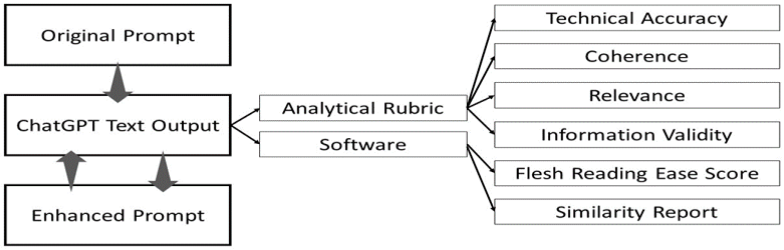

Figure 1 Assessment Framework

6. Data Analysis

The data analysis involves both quantitative and qualitative techniques:

(1) Qualitative Data: Texts are evaluated using the analytical rubric. The results are synthesized to provide an overall evaluation of the text quality generated by ChatGPT-4.

(2) Quantitative Data: Readability scores (FRE) and Turnitin similarity scores are analyzed using descriptive statistics to determine the overall readability and originality of the generated text.

Results

The assessment of ChatGPT-4 outputs using an analytical rubric provides valuable insights into the quality, readability, and credibility of the AI-generated content. The outputs, which responded to a range of academic prompts, were evaluated based on four main criteria: technical accuracy, coherence, relevance, and information validity. For non-native English speakers, these factors play a crucial role in shaping their academic writing experiences and outcomes.

Table 3

Flesch Reading Ease Scores and Text Statistics of ChatGPT-3.5 Outputs

Output No.	Word Count	Average Word per Sentence	FRE Score	Description
1	92	15.1	13.8	Very Difficult
2	235	26.1	5.3	Very Difficult
3	212	19.2	23.6	Very Difficult
4	119	14.8	22.5	Very Difficult
5	124	15.5	36.2	Difficult
6	124	15.5	38.9	Difficult

Table 4

*Quality Assessment of ChatGPT-4 Outputs**

Output No.	Technical Accuracy	Coherence	Relevance	Information Validity
1	5	5	5	4
2	5	5	5	4
3	5	5	5	4
4	5	5	5	4
5	5	5	5	4
6	5	5	5	4

Table 5

WebChatGPT Referenced Sources and Turnitin Similarity Report

WebChatGPT Referenced Sources	Turnitin Similarity Report
2 are university websites.	Overall Similarity Report: 16%
3 commercial websites	Top sources found in the following databases:
1 online knowledge sharing platform.	9% Internet Database
Note: None of the following were cited as references by WebChatGPT extension, published book, research article, and academic paper.	16% Submitted Works Database
	0% Publications Database
	Note: Overlapping sources and flagged at 80% to 90% AI generated

1. Readability – The readability of the generated text was measured using the Flesch Reading Ease (FRE) software, which revealed variability across outputs (Table 3). Outputs 1 to 4 were categorized as "Very Difficult," with FRE scores between 5.3 and 23.6, indicating highly complex text. In contrast, Outputs 5 and 6, which were generated with simpler vocabulary prompts, showed improved readability with FRE scores of 36.2 and 38.9, categorized as "Difficult." While ChatGPT-4 can adjust its language to enhance accessibility, the resulting text may still be challenging for lower-proficiency readers, such as many non-native English speakers. This suggests that while the AI can produce readable content but needs further simplification.

2. Technical Accuracy and Coherence – ChatGPT-4 demonstrated consistently high performance in terms of technical accuracy and coherence across all outputs. These were assessments focusing on grammar, spelling, punctuation, and logical text structure. ChatGPT-4 achieved top scores these criteria, producing error-free and logically structured text (Table 4). For non-native English speakers, this capability is particularly beneficial as it provides a model of grammatically correct and well-organized academic writing.

3. Source Credibility and Information Validity – The analytical rubric also evaluated the validity and credibility of the information provided, focusing on the quality and reliability of the sources cited. ChatGPT-4 consistently scored a four (4) in this category, indicating that although the information was generally accurate, the credibility of the sources was a concern (Table 5). The outputs predominantly cited internet and commercial websites rather than peer-reviewed academic papers, published books, or other scholarly sources. This reliance on less authoritative sources can affect the validity of information.

4. Similarity Index – The overall similarity index of 16%, suggest very low concern for plagiarism. For academic writers however this is not an exemption for proper citation practices to avoid unintentional plagiarism. At the same time should raise concern on verifying the validity of information used since it did not indicate published or peer-reviewed articles. Rather they are sourced from internet websites and works in progress, which could be a concern for reliability.

Discussion

1. Insights from Assessment Rubrics

The assessment process, guided by a detailed analytical rubric and various software tools, offered a structured evaluation of ChatGPT-4's outputs. For non-native English speakers, understanding how these tools assess text can help them leverage AI effectively to improve their academic writing. The technical accuracy criterion evaluated grammar, spelling, punctuation, and adherence to writing conventions. Consistent with

previous studies (e.g., Giray, 2023; Nazari & Saadi, 2024), ChatGPT-4 demonstrated proficiency in producing error-free outputs. This level of accuracy provides non-native speakers with a reliable model for understanding correct language usage and structure. Coherence was another crucial component, focusing on logical flow and clarity. In line with findings from Kohnke et al. (2023), ChatGPT-4 exhibited strong organizational skills, producing outputs with clear and structured content. By analyzing these examples, non-native speakers can learn how to better structure their academic writing for improved clarity and cohesion. The relevance criterion assessed how closely the content aligned with the prompt. ChatGPT-4 consistently maintained focus and provided appropriate information without digressions, similar to findings in Bowman (2022). This highlights the importance of maintaining topical relevance in academic writing, a key skill for non-native English speakers. Information validity evaluated the accuracy of content and the credibility of sources. While ChatGPT-4 generally provided accurate information, it occasionally relied on less authoritative sources—a limitation also noted by previous research (Kohnke et al., 2023). Non-native learners should be guided to cross-check AI-generated references and develop skills in evaluating source reliability.

2. Insights from Software Metrics

Text readability was evaluated using the Flesch Reading Ease (FRE) score in Microsoft Word. Outputs generated with prompts encouraging simpler language demonstrated improved readability, although they still posed challenges for lower-level readers. This finding supports Giray's (2023) argument that intentional simplification strategies are necessary to make AI-generated content more accessible. For non-native English speakers, learning to adjust their writing style based on readability metrics can be an essential skill for crafting audience-appropriate texts. Similarity checks were conducted using Turnitin software to assess the uniqueness of ChatGPT-4 outputs and identify sources with similar content. Turnitin flagged between 80% to 90% of the generated outputs as AI-written, raising serious concerns about originality in academic writing. This significant detection rate underscores ongoing debates regarding the acceptability of AI-assisted writing, as institutions continue to vary in their stance on authorship and originality standards (Nazari & Saadi, 2024). For non-native English speakers, understanding proper citation practices and maintaining originality are critical competencies (Liu, Lin, Kou, & Wang, 2016). The high AI similarity rate reinforces the need for users to recognize when and how to credit AI-generated content to avoid issues of academic integrity or diminished credibility. While ChatGPT-4 demonstrates strong capabilities in generating technically accurate, coherent, and relevant content, challenges remain in ensuring source credibility, enhancing readability, and meeting originality standards. The findings emphasize the need for responsible and informed use of generative AI tools in academic settings, balancing their potential benefits with adherence to ethical writing practices.

In conclusion, the rubric-based assessment, combined with software tools, demonstrates that ChatGPT-4 is effective in generating well-structured, technically accurate, and relevant academic content. However, challenges remain in improving readability for non-native English speakers and ensuring the credibility of sources. The structured evaluation process used an analytical rubric to generate FRE scores and Similarity report highlights the need for careful assessment and validation when using AI-generated content for academic purposes. By leveraging the strengths of ChatGPT-4 while addressing its limitations through the use of rubrics

and software tools, non-native English speakers can better navigate the complexities of academic writing, ensuring both the quality and integrity of their work.

Recommendations

Based on the identified limitations, the following recommendations are proposed to enhance the assessment of AI-generated texts for academic use:

1. Incorporate Multiple Evaluators – To improve the reliability of qualitative assessments, future studies should involve multiple evaluators to mitigate individual biases.

2. Utilize Diverse Readability Metrics – While the Flesch Reading Ease (FRE) score is useful, combining various readability measures can help capture different aspects of text quality.

3. Expand Methodological Tools – Diversify the software tools used for evaluation to include a range of readability and plagiarism detection tools. This will provide a more robust analysis of quality.

4. Compare Different AI Models – To gain a broader understanding of AI's impact on academic writing, future research should compare outputs from different versions of ChatGPT and other AI models. This comparison will offer insights into the strengths and limitations of various AI technologies.

5. Assess Dynamic Contexts – Conduct evaluations in more dynamic, real-world settings that reflect actual user interactions and requirements.

Implementing these recommendations will help refine the assessment framework and provide more accurate, applicable insights into the effectiveness of AI tools like ChatGPT in academic writing.

References

Adhariani, D., & du Toit, E. (2020). Readability of sustainability reports: Evidence from Indonesia. *Journal of Accounting in Emerging Economies, 10*(4), 621-636.

Ahmed, I., & Ishtiaq, S. (2021). Reliability and validity: Importance in medical research. *Journal of the Pakistan Medical Association, 71*(10), 2401-2406.

Aminovna, B. D. (2022). Importance of coherence and cohesion in writing. *Eurasian Research Bulletin, 4,* 83-89.

Atlas, S. (2023). *ChatGPT for higher education and professional development: A guide to conversational AI*. Retrieved from https://digitalcommons.uri.edu/cba_facpubs/548

Ayanouz, S., Abdelhakim, B. A., & Benhmed, M. (2020). *A smart chatbot architecture based NLP and machine learning for health care assistance*. Retrieved from https://dl.acm.org/doi/pdf/10.1145/3386723.3387897

Azaria, A., Azoulay, R., & Reches, S. (2024). ChatGPT is a remarkable tool – For experts. *Data Intelligence, 6*(1), 240-296.

Bala, K., Kumar, M., Hulawale, S., & Pandita, S. (2017). Chat-bot for college management system using A.I. *International Research Journal of Engineering and Technology, 4*(11), 2030-2033.

Bansal, G., Chamola, V., Hussain, A., Guizani, M., & Niyato, D. (2024). Transforming conversations with AI – A comprehensive study of ChatGPT. *Cognitive Computation, 16,* 2487-2510.

Berkeley, S., King-Sears, M. E., Vilbas, J., & Conklin, S. (2016). Textbook characteristics that support or thwart comprehension: The current state of social studies texts. *Reading & Writing Quarterly, 32*(3), 247-272.

Bothun, L. S., Feeder, S. E., & Poland, G. A. (2021). Readability of participant informed consent forms and informational documents: From phase 3 COVID-19 vaccine clinical trials in the United States. *Mayo Clinic Proceedings, 96*(8), 2095-2101.

Bowman, E. (2022). *A new AI chatbot might do your homework for you, but it's still not an A+ student*. Retrieved from https://www.npr.org/2022/12/19/1143912956/chatgpt-ai-chatbot-homework-academia

Corcoran, N., & Ahmad, F. (2016). The readability and suitability of sexual health promotion leaflets. *Patient Education and Counseling, 99*(2), 284-286.

Courtis, J. K., & Hassan, S. (2002). Reading ease of bilingual annual reports. *The Journal of Business Communication, 39*(4), 394-413.

Daffern, T., Mackenzie, N. M., & Hemmings, B. (2017). Predictors of writing success: How important are spelling, grammar and punctuation?. *Australian Journal of Education, 61*(1), 75-87.

Del Giglio, A., & Da Costa, M. U. P. (2023). The use of artificial intelligence to improve the scientific writing of non-native english speakers. *Revista Da Associacao Medica Brasileira, 69*(9), e20230560.

Fang, Y. (2010). Perceptions of the computer-assisted writing program among EFL college learners. *Educational Technology & Society, 13*(3), 246-256.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221-233.

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education, 9,* e45312.

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering, 51,* 2629-2633.

Hwang, S. I., Lim, J. S., Lee, R. W., Matsui, Y., Iguchi, T., Hiraki, T., & Ahn, H. (2023). Is ChatGPT a "fire of prometheus" for non-native English-speaking researchers in academic writing?. *Korean Journal of Radiology, 24*(10), 952-959.

Jia, J. (2003). *The study of the application of a keywords-based chatbot system on the teaching of foreign languages*. Retrieved from https://arxiv.org/pdf/cs/0310018

Johns, A. M. (1986). Coherence and academic writing: Some definitions and suggestions for teaching. *TESOL Quarterly, 20*(2), 247-265.

Kee, L. L, Razali, A. B., Samad, A. A., & Noordin, N. (2020). Effects of digital writing software as a tool for process approach to writing on teacher trainees' academic writing performance. *The Journal of Asia TEFL, 17*(4), 1158-1546.

Khazaal, E. N. (2019). Improving postgraduates' academic writing skills with summarizing strategy. *Arab World English Journal, 10*(3), 413-428.

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537-550.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., Leon, L. D., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health, 2*(2), e0000198.

Li, J., Zong, H., Wu, E., Wu, R., Peng, Z., Zhao, J., Yang, L., Xie, H., & Shen, B. (2024). Exploring the potential of artificial intelligence to enhance the writing of english academic papers by non-native english-speaking medical students – The educational application of ChatGPT. *BMC Medical Education, 24,* 736.

Liu, G., Lin, V., Kou, X., & Wang, H. (2016). Best practices in L2 English source use pedagogy: A thematic review and synthesis of empirical studies. *Educational Research Review, 19,* 36-57.

Mcinnes, N., & Haglund, B. J. A. (2011). Readability of online health information: Implications for health literacy. *Informatics for Health and Social Care, 36*(4), 173-189.

Metzger, M. J., Flanagin, A. J., & Lara Zwarun, L. (2003). College student web use, perceptions of information credibility, and verification behavior. *Computers & Education, 41*(3), 271-290.

Nazari, M., & Saadi, G. (2024). Developing effective prompts to improve communication with ChatGPT: A formula for higher education stakeholders. *Discover Education, 3,* 45.

OpenAI. (2023). *ChatGPT (version 3.5) [Conversational AI]*. Retrieved from https://chat.openai.com

OpenAI. (2025). *What is ChatGPT?*. Retrieved from https://help.openai.com/en/articles/6783457-what-is-chatgpt

Oviedo-Trespalacios, O., Peden, A. E., Cole-Hunter, T., Costantini, A., Haghani, M., Rod, J., Kelly, S., Torkamaan, H., Tariq, A., Newton, J. D. A., Gallagher, T., Steinert, S., Filtness, A. J., & Reniers, G. (2023). The risks of using ChatGPT to obtain common safety-related information and advice. *Safety Science, 167,* 106244.

Pan, S. C., Rickard, T. C., & Bjork, R. A. (2021). Does spelling still matter – And if so, how should it be taught? Perspectives from contemporary and historical research. *Educational Psychology Review, 33,* 1523-1552.

Romano, F. (2019). Grammatical accuracy in EAP writing. *Journal of English for Academic Purposes, 41,* 100773.

Rori, R. O., Olii, S. T., & Rettob, A. (2021). Assessing readability of reading text "bright," an English course for junior high school students. *Journal of English Language and Literature Teaching, 6*(1), 19-35.

Singhal, M. (2004). Academic writing and generation 1.5: Pedagogical goals and instructional issues in the college composition classroom. *The Reading Matrix, 4*(3), 1-13.

Sojasingarayar, A. (2020). *Seq2Seq AI chatbot with attention mechanism*. Retrieved from https://arxiv.org/abs/2006.02767

Stokel-Walker, C. (2023). *ChatGPT listed as author on research papers: Many scientists disapprove*. Retrieved from https://www.nature.com/articles/d41586-023-00107-z

World Commission on the Ethics of Scientific Knowledge and Technology. (2019). *Preliminary Study on the ethics of artificial intelligence*. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000367823

Xu, Z., Ellis, L., & Umphrey, L. R. (2019). The easier the better? Comparing the readability and engagement of online pro- and anti-vaccination articles. *Health Education & Behavior, 46*(5), 790-797.