JAIST Repository
https://dspace.jaist.ac.jp/

Title	Evaluating the Quality of ChatGPT 3.5-Generated Academic Essays
Author(s)	Bui, Minh Ngoc
Citation	
Issue Date	2024-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/19349
Rights	
Description	Supervisor: Kim Eunyoung, 先端科学技術研究科, 修士(知識科学)

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Master's Thesis

Evaluating the Quality of ChatGPT 3.5-Generated Academic Essays

BUI, Minh Ngoc

Supervisor: KIM, Eunyoung

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Knowledge Science)

September 2024

Acknowledgements

There is a meaningful quote that I am really into: “To overcome fear, you must go through it, not around it”. This quote truly resonated with me duirng my two years at JAIST. I always feel fortunate to have received immense love and support from those around me. In challenging moments, despite feeling scared, the encouragement from my teachers and dear friends motivated me to overcome obstacles and keep moving forward.

Throughout this journey, my heartfelt gratitude goes out to my supervisor, Prof. KIM Eunyoung, for granting me the opportunity to be where I am today and for wholeheartedly supporting me since I arrived in Japan. I am deeply thankful for her unwavering guidance and extraordinary patience. From here, Kim lab has truly become my second home, where I have been embraced with warmth and constant support from my dear lab mates. The care and encouragement from everyone have been a source of great comfort and motivation for me while away from home.

I wish to express my sincere appreciation to Prof. DAM Hieu Chi, his family, and my Vietnamese friends in Dam lab for their continual warm welcome and cozy attention. Additionally, I am deeply appreciative of Prof. HUYNH Van Nam and his family for their gracious hospitality. My special thanks are due to Prof. YUIZONO Takaya for his meticulous and insightful feedback on my minor research. Furthermore, I am profoundly thankful to Prof. MOTOYAMA Kotona for providing ample opportunities to develop my strengths, as well as giving me more motivation to strive during my studies.

Last but not least, I am truly indebted to my parents, my sister, my fiancé, and all my cherished friends at JAIST for their unwavering support. My parents and sister, through their love and care, serve as my primary motivators in attempting to pursuit my academic journey here. I extend special appreciation to my fiancé for being a steadfast companion, attentively sharing in my challenges and supporting me along every step of my educational and life journey. Lastly, to all my friends, disregarding differences in skin color, nationality, language, and culture, have stood by my side, offering invaluable assistance in navigating the challenges of life in Japan.

Abstract

Introduction This research investigates the writing quality of academic essays produced by ChatGPT and the factors influencing their effectiveness. Despite the growing use of AI-generated content in education, there is a scarcity of studies on its academic writing proficiency. The objective of this study is to address this gap by evaluating different facets of essay quality, including grammar, coherence, originality of ideas, content development, plagiarism detection, and citation precision.

Originality/Value This study combines a thorough analysis of AI-generated essays with statistical evaluations to investigate ChatGPT's writing abilities. By exploring a diverse range of essay topics and evaluating various writing components, this research offers detailed insight into ChatGPT's proficiency in academic writing. The outcomes hold considerable significance for educators, students, education policymakers, and AI developers.

Research Objectives The primary research goal is to assess the quality of ChatGPT-generated essays across different themes and writing assessment criteria. Sub-objectives include evaluating grammar, coherence, originality of ideas, content development, plagiarism prevention, and reference accuracy in AI-generated essays, as well as identifying factors influencing ChatGPT's writing quality.

Methodology The study methodology consists of four main steps. Initially, ***Data Generation*** involved creating a variety of essays using ChatGPT on diverse academic topics. Subsequently, ***Scale Development*** entailed constructing a comprehensive rubric to assess essay quality based on criteria such as grammar, coherence, originality, content development, plagiarism prevention, and reference accuracy. During the ***Data Evaluation*** phase, each essay was reviewed using this rubric to ensure consistent and reliable assessments. Finally, ***Data Analysis*** employed statistical methods, including one-way ANOVA and correlation analysis, to examine variations in writing quality across different themes and identify significant correlations among the evaluation criteria.

Results The study highlights ChatGPT's strong performance in grammar and coherence, which is particularly beneficial for non-native English speakers. However, shortcomings are observed in content development and personalized conclusions. The research identifies critical thinking abilities and writing proficiency as key components within the

writing quality assessment framework, showing a positive correlation with ChatGPT's writing quality. Additionally, consistent writing competence by ChatGPT across various subjects is noted, indicating its effectiveness regardless of the topic.

Implications for Practice The findings of the study offer significant insights for educators seeking to seamlessly incorporate AI tools into the curriculum to enhance student learning. Educational policymakers within schools can utilize these results to establish protocols ensuring the ethical and effective use of AI in educational settings. Furthermore, AI developers can capitalize on the identified shortcomings to improve the functionality and performance of forthcoming AI writing tools.

Table of Contents

List of Figures

List of Tables

1. Introduction

In recent years, the swift advancement of artificial intelligence and natural language processing technologies has led to the development of groundbreaking tools and systems that are transforming various fields, including education and academic writing (Mhlanga, 2023). A significant development in this area is ChatGPT, which was launched by OpenAI in December 2022. This AI-driven chatbot utilizes deep learning models that are trained on a broad range of datasets, encompassing sources such as books, articles, and online content, to generate responses that closely resemble human-like dialogue (Meyer et al., 2023).

The outstanding benefit of ChatGPT in education is its facilitation of personalized and interactive learning experiences. Additionally, it assists in crafting prompts for formative assessments, ensuring continuous feedback loops that enrich teaching and learning processes (Baidoo-Anu & Ansah, 2023). Moreover, ChatGPT provides precise responses to targeted inquiries and functions as a proficient content generator (Meyer et al., 2023). ChatGPT has gained recognition in higher education for its capacity to produce coherent and contextually fitting texts derived from existing datasets (Almahasees et al., 2024). It is noted that this AI chatbot exhibits the capability to compose scholarly dissertations that closely mimic human language patterns (Lund & Wang, 2023). Nevertheless, the escalating utilization of AI-generated content prompts critical inquiries into the authenticity and dependability of such texts, including concerns about plagiarism, inaccuracies, and biases in training data (Baidoo-Anu & Ansah, 2023). The autonomous nature of the composition process and the absence of direct human intervention pose distinctive challenges in assessing the quality of essays closely resembling human writing generated by ChatGPT. Consequently, there is a vital need to devise more potent evaluation methods and criteria to appraise the writing standards of academic essays produced by ChatGPT.

Although numerous scholarly articles have focused on analyzing the strengths and weaknesses of AI applications in various fields such as healthcare, finance, and creative arts (Udegbe et al., 2024; Fernandez, 2019; Oksanen et al., 2023), the research on the application of AI in education, particularly higher education, remains relatively underexplored. Current evaluations of AI-generated content often rely on general criteria tailored to human-authored texts, which may not fully capture the unique characteristics and limitations of AI writing. Furthermore, there is a lack of comprehensive frameworks specifically designed to analyze AI-generated essays' coherence, originality, and academic rigor.

The purpose of this research is to address these concerns by conducting a comprehensive evaluation of ChatGPT-generated academic essays across various themes. By employing a comprehensive approach to writing criteria such as references, grammar and syntax, plagiarism avoidance, coherence and cohesion, idea originality, and content development, this study is expected to provide an understanding of the quality of AI-generated academic writing. Additionally, the application of principal component analysis will be explored to identify key components that influence the quality of these essays. This exploration will provide valuable insights that could guide the enhancement and advancement of AI writing tools specialized for academic use.

1.1. Scope of the Research

The research examines 72 academic essays produced by ChatGPT, an AI language model by OpenAI. It focuses on essays from six categories: Economics, Education, Healthcare, Lifestyle, Work and Career, ICT, and Others, providing a thorough assessment across different academic fields. The evaluation employs six writing assessment criteria, namely references, grammar, plagiarism avoidance, coherence and cohesion, idea originality, and content development. Quantitative methods will be used to measure and analyze the essays based on these criteria. Principal component analysis will be applied to reduce data dimensionality and identify critical components influencing essay quality. The comparison aims to highlight the strengths and weaknesses of AI-generated essays relative to those written by students. The research explicitly targets the application of AI-generated essays in academic examination settings, examining their alignment with academic standards and potential impact on learning and assessment practices. By focusing on these areas, the study aims to provide a comprehensive evaluation of ChatGPT's capabilities in generating academic essays, which contributes to the broader understanding of AI's role in education and informs the development of more effective AI writing tools.

1.2. Significance of the study

One of the major challenges in implementing and utilizing AI within the academic domain lies in the deficiency of quantitative evidence to comprehensively understand its efficacy in assisting learners. Consequently, establishing a framework to assess the quality of academic writing generated by ChatGPT could introduce a new standard for evaluating AI-generated content in education. This initiative holds implications for enhancing the educational experience, elevating academic standards, and refining the practice of academia.

This study aims to identify instances of plagiarism, inaccuracies, and citation issues through a systematic evaluation of essays generated by ChatGPT. Such an assessment can establish guidelines to regulate AI writing and uphold academic integrity. Furthermore, this investigative work can empower students, particularly those in college, to enhance their writing with AI tools while maintaining their originality. The insightful discoveries will provide technology experts with valuable insights to enhance AI capabilities for superior writing assistance.

1.3. Research objectives

The main research objective (MRO) of this thesis is to evaluate the writing quality of ChatGPT-generated academic essays. This study aims to rigorously analyze how well ChatGPT can produce essays that align with the established scholarly writing standards. To achieve this goal, the research is structured around two sub-objectives. The first sub-objective (SRO1) involves devising a scale for assessing the quality of academic essays generated by ChatGPT. This entails creating a comprehensive and reliable assessment rubric that effectively evaluates various aspects of writing quality in AI-generated essays. By establishing an evaluation scale, this study expects to ensure that the assessment of ChatGPT-generated essays is systematic and objective. Subsequently, the second sub-objective (SRO2) is to investigate the factors influencing the writing quality of ChatGPT. This objective aims to identify and analyze the potential correlation that may impact the quality of the content written by ChatGPT. By understanding the influential factors, the study aims to uncover how different variables can enhance or diminish the quality of AI-generated essays, offering insights that could inform the future development and refinement of AI writing tools. Figure 1 illustrates the attainment of the MRO through the accomplishment of these sub-objectives.

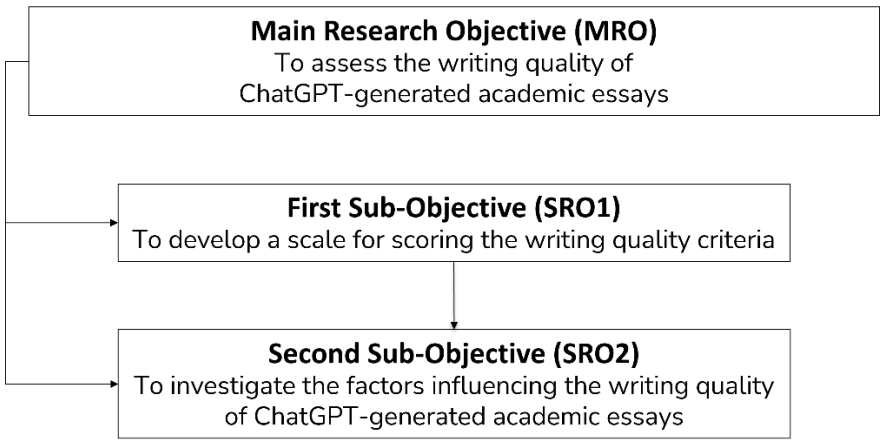

Figure 1: Research objectives of this thesis

1.4. Structure of the thesis

This thesis is comprised of the following five chapters:

Chapter 1 introduces the research, focuses on the research problem's background, outlines the research objectives, and discusses the study's significance.

Chapter 2 (Literature Review) explores the study's theoretical framework and critically evaluates literature pertaining to writing concepts and quality assessment criteria for academic writing. Additionally, this chapter offers the theoretical underpinning for applying AI-based language models in academic writing.

Chapter 3 (Research Methodology) details the methodology used in the study, outlining four key steps: data generation, scale development, data evaluation, and data analysis. These steps illustrate the data collection process, establishment of a quality assessment scale, and data evaluation and analysis based on the developed scale.

Chapter 4 (Results) reveals the research findings, including statistical analysis results and other pertinent data. This chapter comprehensively examines each criterion in the writing quality assessment scale and investigates the factors influencing the quality of academic essays generated by ChatGPT.

Chapter 5 (Discussion and Conclusion) summarizes the writing assessment criteria used to evaluate the quality of ChatGPT-composed academic writing and discusses the implications of the findings. This chapter also states the study’s contributions, limitations, implications for stakeholders, and recommendations for future research directions.

2. Literature Review

This study focuses on investigating the quality of ChatGPT's writing across various topics. This chapter then outlines the theoretical framework essential for formulating evaluation criteria for writing. Furthermore, the concept of writing quality used in this study will be explained.

The primary purposes of this chapter are to:

(1) Compare definitions and writing processes of General Writing and Academic Writing.

(2) Present the concept of quality evaluation criteria for academic writing essays.

(3) Examine previous research on the use of AI-based language models in academic writing to understand the features and abilities of AI in writing and frameworks for evaluating AI-written essays.

2.1. General Writing and Academic Writing

2.1.1. General Writing

Throughout the academic level from primary to secondary education, the indispensable skill of writing takes center stage. With a rich historical legacy, writing stands as a globally revered practice that enhances human expression and cognitive faculties (Aaron & Joshi, 2006). Writing encompasses the creation of diverse texts tailored to specific purposes and audiences, spanning genres such as fiction, non-fiction, poetry, and technical writing (Benade et al., 2021). Serving a multitude of functions, writing acts as a communication tool, a means of information preservation, and a channel for artistic expression. According to Booth and colleagues (2009), the writing style can exhibit significant variation based on the writer's voice, the intended audience, and the text's overarching objective. This flexibility empowers writers to experiment with language, structure, and narrative techniques in a creative pursuit of self-expression.

From a cognitive perspective, writing is creativity as a problem-solving endeavor that engages various mental faculties. Flower and Hayes (1981) described writing, encompassing general and academic contexts, as a recursive process involving planning, translating, and reviewing. In alignment with Flower and Hayes, Seow (2002) suggested that process writing, a classroom exercise, comprises four fundamental writing stages– planning, drafting, revising, and editing – apart from three additional stages imposed on students by instructors: responding, evaluating, and post-writing. The drafting stage serves as the introduction to the writing subject, utilizing varied approaches, including a striking statement, a concise summary, a relevant quotation, a thought-provoking question, a general declaration, an analogy, or a statement of intent (Seow, 2002). This iterative process entails continual movement between stages as

writers refine their concepts and texts (Krashen, 1984). Conversely, Elbow (1998) emphasized the expressive facet of writing, portraying it as a medium for self-expression, personal growth, and deliberate engagement with one's surroundings and experiences. He argued that writing facilitates the exploration of thoughts and emotions, serving as a potent instrument for self-discovery and contemplation. Barton and Hamilton (1998) asserted that writing practices are influenced by social dynamics and power structures, underscoring that writing is not merely an individual pursuit but interacts with the writer's social contexts and target audience. While Flower and Hayes (1981) and Seow (2002) focused on the cognitive processes inherent in writing, and Barton and Hamilton (1998) stressed the impact of social interactions and power dynamics on writing practices, there exists a discernible gap in integrating these viewpoints. A holistic comprehension of writing necessitates the fusion of cognitive and social dimensions to explore how mental operations and social frameworks converge to shape writing practices.

In terms of functions of writing, it fulfills various essential functions, including facilitating communication, maintaining records, and enabling creative expression. Britton et al. (1975) categorized writing into three primary forms: transactional, expressive, and poetic. Transactional writing is designed to inform or persuade; expressive writing refers to personal reflections, while poetic writing emphasizes artistic expression and creativity. Swales (1990) introduced the concept of discourse communities, underscoring how the writing conventions and expectations within specific groups benefit communication. He argued that grasping the genre conventions of a given discourse community is pivotal for effective writing. This perspective underscores the significance of considering context and audience when engaging in writing endeavors.

At the educational level, writing remains the main method of assessment. As Sumner and Connelly (2020) highlighted, in universities, students are expected to exhibit their comprehension of subjects through independent writing tasks. It is noted that a significant portion of academic disciplines necessitate written assignments, with many British universities mandating handwritten essays during exams at various educational stages. While the prevalence of handwritten exams in the United States and other global educational institutions is diminishing in favor of digital assessments, a considerable number of university students still opt for or are obliged to write under time constraints (Sumner & Connelly, 2020; Mogey et al., 2020). However, these studies focus primarily on writing practices and assessments in higher education, particularly in Western contexts. More studies are needed to examine writing practices in diverse educational settings, including primary and secondary education in

different cultural contexts, which could provide a more global understanding of writing education.

In East and Southeast Asian nations such as Singapore and Vietnam, writing is a fundamental skill integrated into the primary and secondary education curricula (Swandi & Netto-Shek, 2017; Bui & Hseih, 2024). Writing encompasses different genres such as narrative, descriptive, argumentative, and expository forms. In foreign language training, while the application of writing in traditional Vietnamese training programs is considered not as effective in developing comprehensive English language skills in students because rote learning and grammar instruction had been dominant (Hung, 2019), recent studies indicated a growing shift towards more interactive and student-centered methods. Le (2023) highlighted the integration of process writing and peer review in some schools, promoting drafting, revising, and collaborative feedback.

2.1.2. Academic Writing

Academic writing is defined as a diverse field that includes many different genres and purposes. In addition to being known as a means of recording information, academic writing is also considered a tool for critical thinking and sharing knowledge among the scholarly community. This is a formal writing style used in academic and educational contexts. It is characterized by a focus on evidence-based arguments, clarity, and a formal tone. This type of writing can include genres such as research articles, theses, dissertations, and academic essays whose purpose is to advance knowledge, present research results, and engage in scholarly dialogue. art. According to Swales (1990), academic writing is guided by the conventions of the academic community, which require adherence to specific formats, citation styles, and disciplinary rules. Therefore, academic writing is not aimed at a general but specialized audience, including scholars, researchers and students in a specific field. However, this paper, which generalized academic writing conventions, may overlook specific genre-based challenges. This gap suggests a need for more detailed studies that explore how evidence-based argumentation and critical thinking are applied explicitly in various academic genres.

In terms of style and structure, academic writing requires a formal tone and precise language. It avoids colloquialisms, contractions, and subjective expressions, focusing on objectivity and clarity. Birkenstein and Graff (2018) argued that academic writing must be impersonal and evidence-based, relying on a third-person perspective and formal vocabulary. This formality is essential for maintaining credibility and ensuring that arguments are presented logically and systematically. Academic writing follows a specific structure designed to

facilitate clear and logical presentation of ideas. The IMRaD format is commonly used in research papers, while essays typically include an introduction, literature review, methodology, results, discussion, and conclusion (Cusen, 2018; Gjesdal, 2013). Scientific writing is part of a broader thought-making process and involves revisions and rewrites until the final product's design (Soares, 2022). Therefore, this material arises as a result of consulting and advisory experiences for the preparation of academic papers. The process of academic writing involves multiple revisions. It rewrites before arriving at the final version of the text, making it a critical component of the broader thought-making process in academia. This structured approach helps writers present their research systematically, ensuring that their arguments are coherent and supported by evidence.

Academic writing has specific requirements that follow strict conventions concerning evidence, citation, and argumentation. Birkenstein and Graff (2018) emphasized that effective academic writing requires critical analysis, the synthesization of information, and rigorous use of evidence to support arguments. To enhance readability and ensure clarity and precision, sentences should be clear, direct, and free from unnecessary jargon or complexity. Booth et al. (2009) recommended employing techniques (e.g., active voice and straightforward sentence structures). Proper citation is crucial for acknowledging original authors and enabling readers to trace information sources accurately. Following MLA (2021) and APA (2020) guidelines ensures accurate and consistent citations. Adhering to proper citation practices prevents plagiarism and boosts the essay's credibility and reliability.

When comparing general writing with academic writing, it is clear that general writing tends to be less formal. As noted by Booth and his colleagues (2009), general writing style can vary significantly based on the writer’s perspective, target audience, and purpose of the text. This can be explained because there is no purpose in citing official scientific information; general writing may contain many subjective personal opinions of the author. In contrast, academic writing aims to share new findings, ideas, and perspectives within the academic community. According to Swales and Feak (2012), academic writing is an important means of contributing to the body of knowledge in a variety of fields. Gabi (2022) argues that academic writing allows the writer to persuade the audience with the ability to argue persuasively with a professional tone.

This study mainly focuses on examining the academic writing of AI-authored essays. Previous literature has established a comprehensive theoretical framework outlining the characteristics, format, and functions of academic writing. However, because of the specificity

of this research analyzing essays generated by a large language model, the existing research on academic writing conventions mainly focusing on human-authored texts may not fully address the unique characteristics of AI-generated essays. While works such as those by Swales and Feak (2012) and Graff and Birkenstein (2018) have provided valuable insights into traditional writing norms and structures, they may be insufficient for evaluating the quality and adherence of AI-authored academic essays. Some existing research is no longer relevant in the current context because it does not consider recent advances in AI writing technology. It can be argued that evaluating writing generated by AI is difficult and much different from evaluating writing by humans. Therefore, there is a need to investigate how AI-generated content aligns with or differs from established academic writing standards, especially in terms of evidence-based argumentation, citation practices, and structural norms.

2.2. Quality Evaluation Criteria for Academic Writing

Writing assessments are essential in educational environments as they offer insights into students' language abilities and help teachers tailor their instructional approaches. There are a variety of tools and criteria employed to gauge the quality of writing. This encompasses facets such as content development, vocabulary usage, structural coherence, and organizational proficiency. This part focuses on reviewing the methodologies, linguistic attributes, and theoretical frameworks underpinning writing assessment.

The writing process includes activities that draw upon experience and memory and inquiry strategies and techniques that empower students to explore beyond their current knowledge and experiences, as outlined by Reither (1985). This writing process requires both cognitive and metacognitive skills. Bereiter and Scardamalia (1987) distinguished between knowledge-telling and knowledge-transforming writing models. The former involves the straightforward presentation of information, while the latter requires higher-order thinking and the reorganization of knowledge to create new insights. These cognitive and metacognitive skills are deeply intertwined with the SECI model of knowledge management of Nonaka and Takeuchi’s research (1995), which described how tacit and explicit knowledge are converted and shared. In writing, socialization corresponds to collaborative activities where tacit knowledge is shared (Vygotsky, 1978), while externalization involves articulating internal thoughts into explicit text, aligning with Bereiter and Scardamalia’s knowledge-transforming model.

In the field of linguistic analysis, researchers have highlighted three essential components: lexical, syntactic, and cohesive elements, as they are effective in evaluating text

complexity and richness (McNamara et al., 2010). The significance of lexical components was underscored, including lexical diversity, density, and sophistication, positing that advanced vocabulary reflects enhanced lexical prowess and writing proficiency (Crossley, 2020). Theoretical perspectives suggest that factors, namely word frequency, associative learning, automatization, abstraction, and representations of word forms and meanings, collectively contribute to lexical acquisition (Ellis, 2002; Langacker, 2007; Goldberg, 2006). Similarly, writing quality is closely associated with syntactic abilities, with proficient writers displaying more refined sentence construction skills than beginners (Nippold, 2000; Scott, 2004). The development of syntactic complexity progresses significantly from elementary school through college, reflecting an increasing ability to structure sentences in more sophisticated ways (Nippold, 2000; Scott, 2004). For instance, in a study by Beers and Nagy (2009), freshman college writers produced a greater number of syntactically complex sentences measured by the number of modifiers per noun phrase) than ninth-grade writers, highlighting a clear developmental trajectory in syntactic sophistication. Similarly, Nippold (2000) found that more advanced writers produced longer sentences and longer clauses, indicating syntactic growth over time. This growth is not limited to sentence length but also includes the variety and complexity of sentence structures used.

Another critical aspect of assessment is the development of a writing piece, which is how the ideas are communicated in an essay rather than the sentence-level structure. Content development can be evaluated through modifiers, concrete nouns, quantitative indicators, and examples, although this method offers a broad overview and may not intricately capture syntactic complexity (Crossley, 2020). Furthermore, Bae and his co-authors (2016) investigated five factors, including content, coherence, originality, grammar, and text length, to find out the relationship between content development and other factors. Three content models were evaluated using structural equation modeling, with the model that considers content as influenced by the five writing elements providing a reasonable explanation of the data. The results suggested that content is directly or indirectly impacted by these elements, highlighting their importance in assessing writing. Furthermore, Booth and his partners (2009) underscored the importance of structuring arguments to persuade readers and establish the validity of one's research, focusing on the structured development of claims, reasons, evidence, and the addressing of counterarguments. They highlighted that an argument commences with a claim that asserts the accuracy of a statement; reasons provide justification for the claim, and strong arguments necessitate evidence to bolster the reasons. Evidence comprises factual data,

information, and credible sources that support the reasons. Although the logical connection of ideas indicated the high ability to utilize more complex syntactic forms, such as relative clauses and subordinate clauses, which contribute to a more nuanced expression of ideas, as research by Scott (2004), Ruegg and Sugiyama (2010) suggested that content development in writing can be evaluated through organization scores and essay length, rather than aspects such as main ideas, logical connections, support, and development.

Creativity plays a vital role in writing, focusing on originality, imagination, and individual expression. Scholars such as Buzan (2017) and Gardner (1983) highlighted that creative writing involves generating unique ideas and using language in innovative ways. It is essential in shaping a writer's distinct voice and style. Creative writing goes beyond just generating new ideas; it also requires presenting them in a captivating way that captivates and connects with the audience. Sofia (2023) noted that distinctive metaphors, vibrant imagery, and a variety of sentence structures are often used in creative storytelling to evoke emotions in their readers. Achieving proficiency in this style of writing necessitates a profound grasp of language and a readiness to explore different writing techniques. Nevertheless, Gardner's paper (1983) predominantly emphasized the "purely cognitive components" without illustrating the application of creativity in diverse domains. Conversely, Sofia's paper overlooked the potential influence of educational background and training on cultivating creative intelligence and its utilization in creative writing.

Additionally, various scales have been developed to evaluate students' writing quality. The Writing Quality Scale serves as a user-friendly tool for assessing writing quality in higher education. It considers aspects such as content development, organization, vocabulary usage, sentence structure, punctuation, and spelling (Stuart & Barnett, 2023). Conversely, another group of authors utilized a correlational approach to explore the relationship between extensive reading and writing fluency in EFL learning. Their study focused on argumentation, evidence presentation, refutation, rebuttal, language structure, and conclusion formulation among students in Palangka Raya (Fitriansyah & Miftah, 2020). While both studies emphasized writing structure and lexical choices, Fitriansyah and Miftah's research emphasized examining argument logic and connections. Nevertheless, the existing writing quality assessment scales primarily focus on writing techniques and language use, with minimal consideration for evaluating problem-solving depth and idea development.

In general, the aforementioned articles encompassed a variety of fundamental criteria for assessing writing, primarily focusing on aspects such as language precision, vocabulary

diversity, and idea elaboration in writing. Because of the novelty of artificial intelligence, few studies have delineated the requisite criteria for evaluating articles generated by AI. Leveraging machine learning algorithms such as deep neural networks and artificial intelligence systems have the capacity to render judgments akin to those made by human experts. However, these advanced capabilities are accompanied by transparency-related challenges and the potential for biases (Bathaee, 2020). Consequently, evaluating articles produced by AI necessitates a paradigm shift in assessment rubrics to align with the distinctive characteristics of AI, potentially involving scrutinizing the incorporation of external sources, verifying the accuracy and presence of citations within an article, and evaluating the rationale and decision-making processes applied to problem-solving scenarios.

2.3. The application of AI in Academic Writing

2.3.1. Characteristics and Capabilities of AI in Writing

AI-powered tools, such as OpenAI's GPT models, have brought about a considerable transformation in automated content creation. These tools utilize natural language processing methods to produce text that closely resembles human writing when given a prompt. The impact of these AI models has been notable in journalism, marketing, and creative writing, as emphasized by Lobajova (2023). AI-based tools, specifically Grammarly, Hemingway, and ProWritingAid provide instant automated writing support by checking grammar, offering style tips, and improving readability in real-time. By employing natural language processing algorithms to evaluate text, these tools give writers feedback on grammar, syntax, and style, helping them create more refined drafts efficiently. AI technologies also enhance the quality of writing through semantic analysis and style correction. Writefull and AI Writer use advanced deep learning models to suggest better word choices, sentence structures, and coherence. These sophisticated tools serve as invaluable assets for individuals honing their English proficiency as a secondary language, offering insightful recommendations to enhance communication fluency. In terms of nurturing creativity, advanced AI platforms such as GPT-3 and OpenAI's Codex stand out as exceptional resources, adept at crafting text that is not only cohesive but also contextually precise (Brown et al., 2020). These models are utilized for drafting research papers, formulating hypotheses, and even composing literature reviews, significantly reducing the time needed for creating content. However, the quality and originality of AI-generated content remain ongoing subjects of research and discussion (Baidoo-Anu & Ansah, 2023).

The utilization of Artificial Intelligence (AI) in academic writing stands out significantly due to its substantial advantages for both educators and students. AI tools play a pivotal role in

simplifying the writing process, enabling researchers to dedicate their efforts towards more advanced cognitive tasks such as idea generation and argument enhancement (Aljuaid, 2024). Besides, AI-driven feedback mechanisms improve the quality of writing by providing consistent and objective suggestions. These tools help writers adhere to academic writing standards, enhancing their work's clarity, coherence, and overall quality, especially by providing non-native English speakers with immediate feedback on grammar, syntax, and style (Schmidt-Fajlik, 2023; Li & Zhang, 2020). Moreover, AI systems can be customized to adhere to specific style guides and formatting requirements, further enhancing the consistency of academic writing. This support helps bridge language barriers, allowing international researchers to contribute more effectively to the global academic community.

However, previous authors have also been concerned about the quality and ethical implications of AI-generated content. A significant concern is the likelihood that ChatGPT will produce false or fabricated information (Baidoo-Anu& Ansah, 2023). Therefore, AI-generated content started to threaten creativity and credibility in academia. Regarding the application of AI in detecting unethical actions in academia, although Turnitin and Dupli Checker are widely used AI-powered plagiarism detection tools in academia and publishing, they were shown to be incapable of efficiently detecting a paraphrased passage (Bhuyar& Deshmukh, 2023). Issues related to bias, privacy, and algorithmic transparency raise concerns regarding the fairness and accountability of AI systems (Akinrinola et al., 2024). Moreover, the proliferation of AI-generated content blurs the line between human and machine-authored work, challenging traditional notions of authorship and creativity (Abbott& Rothman, 2023).

2.3.2. Assessment Frameworks for AI-Authored Essays

Current frameworks for evaluating writing quality, such as those used in human-authored essays, may not be fully applicable to AI-generated content. The essays composed by AI require the development of specific metrics that account for the unique characteristics of AI writing, including syntactic complexity, coherence, and the degree of novelty. Additionally, the role of self-regulation and cognitive processes in writing was emphasized, suggesting that assessment tools should also consider these dimensions in evaluating AI-generated essays (Graham & Harris, 2000).

The research conducted by Yeadon et al. (2024) analyzed 300 short-form physics essays. Half of these were written by students prior to the introduction of ChatGPT, while the other half were generated by OpenAI's GPT-4. The evaluation process replicated the assessment method used in the “Physics in Society” module at Durham University, with all assessors being

experienced in this field. Essays were graded based on the standard United Kingdom university criteria on a 100-point scale. The study also examined the effectiveness of five software tools – ZeroGPT, QuillBot, Hive Moderation, Sapling, and Radar – in identifying essay authorship. ZeroGPT exhibited the highest accuracy, achieving a 98% accuracy rate and a precision score of 1.0 when simplified to binary outcomes. While the research raises questions about the impact of AI integration in academic writing, it does not explore academics' viewpoints on the appropriate level of AI involvement in human-written work, highlighting a gap in understanding the evolving role of AI in academic writing assistance.

The study by Katar et al. (2023) focused on the peer review process within academic publishing, examining research that evaluates the benefits and drawbacks of employing GPT-3 for crafting research papers. Except for the abstract and conclusion part of this article, all parts were provided by queries, they found out a 5% plagiarism rate of ChatGPT when generating content, the article we retrieved had a relatively low degree of resemblance. They also looked through and checked 32 references generated by GPT-3 for the article (i.e, article title, authors, journal, and DOI). Another main drawback related to the essays generated by GPT-3 is citations with wrong APA format and not validated sources. However, this paper lacks specific quantifiable evidence to support this. Relying solely on comments from human evaluators could lead to a lack of transparency and bias in how AI evaluates essays.

Safrai and Orwig's (2024) study aimed to assess ChatGPT-4's capability in creating a biomedical review article about fertility preservation. Their findings revealed that ChatGPT-4 can generate a scientific review on this topic with minimal plagiarism. Although accurate in content, the study noted factual and contextual inaccuracies, and inconsistent reference reliability. These limitations suggest that ChatGPT-4 should not be the sole tool for scientific writing but could be beneficial as a writing aid. The experts evaluated the article and references produced by ChatGPT for accuracy, plagiarism using online tools, and rated its relevance, depth, and timeliness on a scale from 0 to 5. One drawback of this research is its focus solely on medical topics, which may not be representative of other academic subjects.

Due to the novelty and rapid development of artificial intelligence, a common feature of the research articles reviewed here is that they often provide general comments on the characteristics and challenges of AI-based chatbots without offering specific solutions. For instance, recent studies have highlighted the potential of AI in generating human-like text but have also pointed out limitations in creativity, context understanding, and ethical considerations (Baidoo-Anu& Ansah, 2023; Bhuyar& Deshmukh, 2023; Akinrinola et al.,

2024). Quantitative evidence shows that while AI-generated content can achieve high levels of grammatical correctness, it often lacks the nuanced understanding necessary for complex academic tasks (Yeadon et al., 2024). On the other hand, previous articles have mentioned ChatGPT's contributions to academic writing. However, these discussions have primarily revolved around specific domains and binary evaluations, such as right/wrong answers or predefined reasoning tasks (Brown et al., 2020). These studies typically measure success based on the AI's ability to generate correct responses to well-known questions or to follow structured prompts rather than evaluating its creative or critical thinking capabilities. For example, studies by Katar et al. (2023) and Safrai and Orwig (2024) have focused on analyzing the accuracy of article citations, grammar, and spelling in AI-generated texts. These studies provide quantitative data showing that while ChatGPT can produce grammatically correct sentences with accurate citations, it often struggles with the nuances of academic argumentation and critical analysis. Furthermore, these articles have criticized the inaccuracy in ChatGPT's assessment of plagiarism when providing answers to existing questions, highlighting the need for more sophisticated plagiarism detection and content originality measures.

Evaluating an AI-based chatbot based on its ability to develop content, demonstrate flexibility in application, and provide personal opinions is an important dimension that has been somewhat overlooked. Current research often fails to prove how AI is assessed for its ability to engage in open-ended, high-level reasoning tasks that require deep understanding and original thought. For example, when AI systems are challenged with open questions that require social reasoning, formality, and literary analysis, they frequently fall short in delivering contextually relevant and creatively insightful responses. Moreover, assessing the quality of AI-generated academic essays should involve tasks that challenge the chatbot's ability to formulate coherent arguments, demonstrate depth of knowledge, and adapt to various academic genres and styles. This includes evaluating the AI's performance on tasks that require synthesis of information, critical thinking, and personal interpretation, such as formulating thesis statements, constructing logical arguments, and engaging with counterarguments.

Therefore, this paper aims to explore these dimensions by systematically analyzing the quality of academic essays generated by ChatGPT, emphasizing content development, flexibility in wording, and the ability to provide personal insights. In addition, factors such as plagiarism avoidance, reference, and grammar will be added to investigate the stability in the quality of ChatGPT when generating social-related essays to satisfy the requirement of an international exam such as IELTS. The interpretation of Writing Quality Criteria used to

evaluate ChatGPT-generated academic essays, which is the result of inheriting the conclusions from previous research and presenting new features to adapt the academic writing assessment performed by AI, will be presented in Table 1.

Table 1: Writing assessment criteria to evaluate ChatGPT-authored essays

Writing Quality Criteria	Definition
Grammar	The correct use of subject-verb agreement, correct use of tenses, and proper sentence structure (Strunk, 2000)
Plagiarism Avoidance	The uncredited use of another person's words, ideas, or creative expressions (Carroll, 2007).
Reference	Citing the sources of information, ideas, and data in previous literature involves providing details that allow readers to locate and verify these sources (Booth et al., 2009; Katar et al., 2023).
Content development	Identifying components of an argument, determining the validity of claims, supporting arguments with evidence, and evaluating reasoning (Bae et al., 2016; Booth et al., 2009; Stuart & Barnett, 2023)
Coherence and Cohesion	Connecting sentences and paragraphs using explicit linguistic elements such as substitution, ellipses, conjunctions, and linking words (Halliday & Hasan, 1976; Todd et al., 2007).
Idea Originality	Advancing from known knowledge to new understanding, including rational thinking based on facts or beliefs (Halliday & Hasan, 1976; Traxler & Gernsbacher, 1992).

3. Methodology

3.1. A Summary of Research Design

The main research objective (MRO) of this thesis is to evaluate the writing quality of academic essays generated by ChatGPT. The methodology is structured into four essential steps: Data Generation, Scale Development, Data Evaluation, and Data Analysis.

To achieve MRO, this study encompasses two sub-objectives. The initial sub-objective (SRO1) aims to establish a scoring scale for assessing writing quality criteria, while the subsequent sub-objective (SRO2) focuses on investigating the factors influencing the writing quality of ChatGPT-generated academic essays. SRO1 and SRO2 are interconnected, where SRO1 sets the standards for assessing ChatGPT's writing skills, and SRO2 applies these standards to explore various elements influencing the quality of essays. These sub-goals rely on each other; the metrics developed in SRO1 provide crucial evaluation criteria and shape the investigation in SRO2. Conversely, findings from SRO2 can enrich and confirm the scoring system established in SRO1.

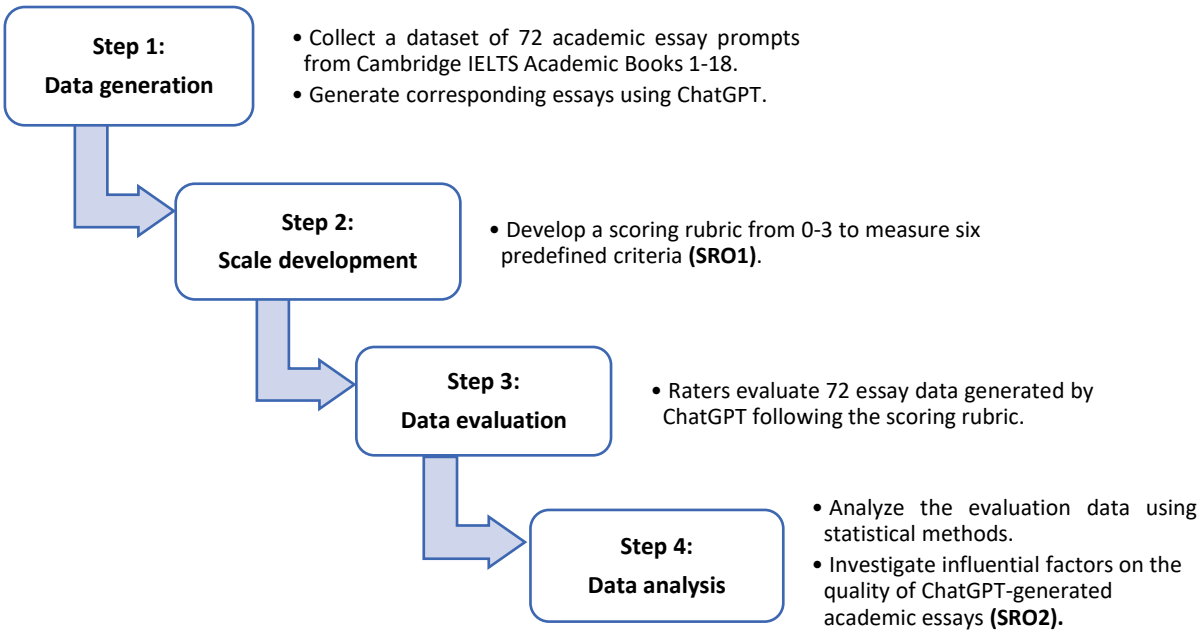

Figure 2: Research Design

3.2. Data generation

The first step in the methodology involves collecting ChatGPT-generated essays. A dataset was meticulously compiled, comprising 72 academic essay prompts and the corresponding responses generated by ChatGPT. The essay prompts originated from 72 questions extracted from Writing Task 2 in Cambridge IELTS Academic Books. This evaluation process spans from the first book to the eighteenth book of Cambridge IELTS. This series of books is the official preparation materials for the examination of International English Language Testing System (IELTS) developed by Cambridge University Press and collaborators such as the British Council, IDP, and Cambridge English Language Assessment. In the realm of IELTS Writing Task 2, test-takers are challenged to eloquently present a range of perspectives, debates, and challenges. This writing task necessitates clear organization, logical progression, and the skill to substantiate arguments with pertinent examples and evidence in a formal academic style. Covering a wide array of topics spanning education, environment, health, technology, and social concerns, the IELTS Writing Task 2 aims to gauge candidates' aptitude in expressing opinions, analyzing subjects, and constructing coherent arguments supported by evidence.

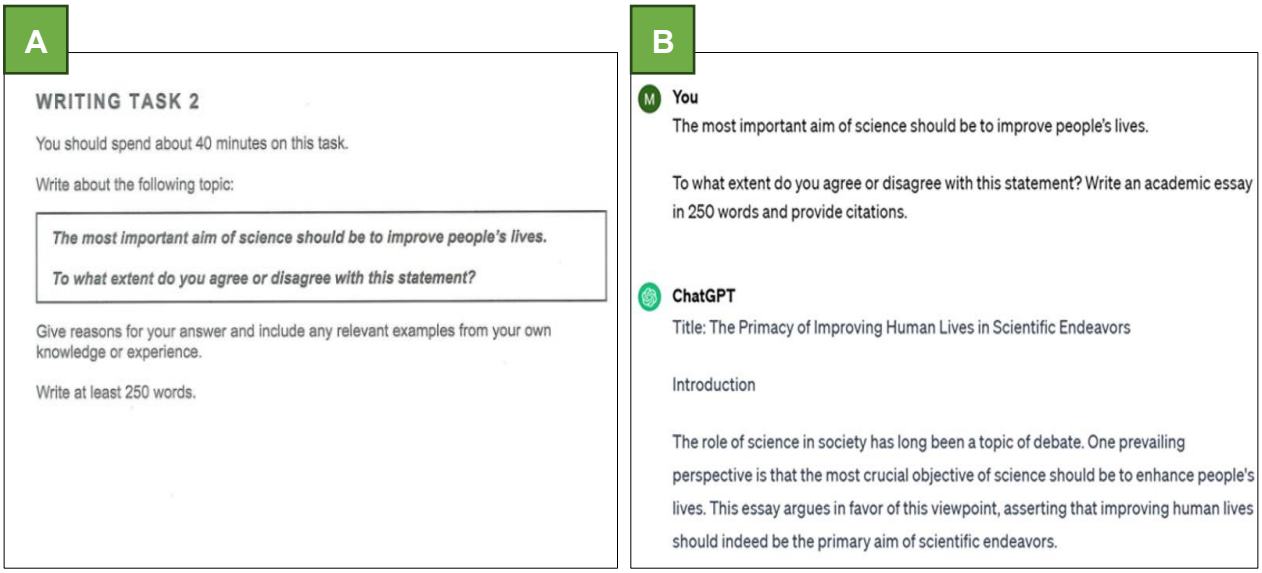

Figure 3: (A)-Sample of a question in IELTS Writing Task 2; (B)- A prompt and a response in GPT 3.5

This study assessed academic essays generated by ChatGPT version GPT3.5, an advanced AI language model, using prompts sourced from the Writing Task 2 section of the Cambridge English IELTS Academic Books series 1 to 18. The evaluation focused on ChatGPT's capability to produce original, logically structured, and grammatically accurate content without plagiarism, reflecting proficiency in critical thinking and constructing well-

supported arguments. The study included examples of an essay question in Cambridge IELTS Book 12 and a Q&A between a user and ChatGPT based on the provided prompt, as illustrated in Figure 3. The data collection process took place over a span of three weeks, commencing on January 7, 2024 and concluding on January 30, 2024.

3.3. Scale Development

A detailed scoring rubric (SRO1) was developed to evaluate the quality of the ChatGPT-generated essays. This rubric assesses the essays based on six writing assessment items, including grammar, plagiarism avoidance, the accuracy of references, content development, coherence and cohesion, and the originality of ideas (see Appendix A). Each item is rated on the basis of a four-point Likert scale, where 0 means “Poor Quality” and 3 means “High Quality”. A score of 0 signifies frequent errors that significantly impede comprehension or denote serious mistakes, whereas a score of 3 indicates an absence of errors. The intermediate scores represent occasional errors that may sometimes disrupt understanding (1 point) and a few mistakes that seldom hinder comprehension (2 points). This scale facilitates a thorough assessment encompassing both the essays' technical elements and creative substance.

Regarding the writing quality scale items, *Grammar* refers to grammatical correctness and adherence to proper syntax, in which essays will receive ratings on a scale emphasizing minimal grammatical errors and appropriate sentence structure. The aspect of *Plagiarism Avoidance* assesses the textual originality of essays in comparison to external sources. Evaluations for this aspect are assessed based on the percentage of text similarity shown in the Grammarly application, a software program designed to support proofreading and grammar learning in the context of their own writing. The essays with higher scores are assigned to essays demonstrating lower resemblance to other works. The criterion of *Reference* scrutinizes the accuracy of citations regarding authors, titles, and journals. For each component that is found to exist, ChatGPT's *Reference* score will be added by 1. Essays are rated on their precision in the mentioned components’ names, with higher scores indicative of correct and consistent citation practices.

In addition, *Content Development* evaluates the progression of ideas and their substantiation with evidence. Essays are evaluated based on the layers of the arguments, supporting ideas, and the provision of relevant examples. The higher scores reflect a well-developed content framework and persuasiveness in argumentation. Conversely, articles that lack strong arguments or adequate evidence highlight ChatGPT's limitations in presenting convincing points. *Coherence and Cohesion* gauge the logical flow within essays, assessing

the coherence between sentences and paragraphs. Scores are assigned based on the use of linking words between sentences and paragraphs and the clarity of those connective devices between ideas, with higher scores indicating an effective use of cohesive elements. *Idea Originality* aims to assess the presence of ChatGPT's own conclusions with unique insights, reasonable explanations, and good termination of ideas. Essays are rated on the individualism and ownership of ideas, with higher scores awarded to compositions showcasing its own opinions and perspectives.

3.4. Data Evaluation

During this step, the focus transitions to the practical implementation of the developed scoring rubric for assessing the essays generated by ChatGPT. To ensure a high standard of evaluation process, two essay evaluators were thoughtfully selected based on their qualifications and proficiency in academic writing. The chosen evaluators underwent a careful selection process to guarantee the credibility and consistency of the evaluation. Both evaluators had achieved a certification score of 7.0 or higher on the IELTS test, corresponding to a C1 level as per the Common European Framework of Reference for Languages (CEFR) standards. Additionally, they are Master's students well-versed in academic paper composition and analysis, equipping them with the requisite expertise to assess the essays critically.

A comprehensive training program was designed to equip the evaluators with a clear understanding of the evaluation criteria and the specific tasks they were to perform. The training process involved a verbal briefing session and an assessment guide. In the Verbal Briefing Session, the training commenced with a 30-minute verbal briefing session. During this session, the evaluators were introduced to the assigned tasks and the evaluation criteria. This interactive session allowed for the clarification of any doubts and ensured that the evaluators had a unified understanding of the evaluation process. Following the verbal briefing, the evaluators were provided with an Assessment Guide in PDF format. This guide contained detailed information about the scale, the scoring methodologies, and illustrative examples to aid in the evaluation process. The guide served as a comprehensive reference tool, ensuring that the evaluators could consistently apply the scoring criteria.

Once the training was completed, the evaluators proceeded to assess 72 essays within a designated timeframe. The evaluation process spanned two periods: the initial phase in February 2024, when two evaluators evaluated 20 academic essays to pilot the results generated by ChatGPT, and the subsequent phase in April 2024, involving the evaluation of the remaining 52 essays. Each essay was meticulously examined based on six pre-established criteria:

grammar, plagiarism, reference accuracy, content development, coherence and cohesion, and idea originality, with scores assigned according to a predefined rubric from Step 2. The evaluators diligently analyzed each essay to provide a comprehensive evaluation of the writing quality of the ChatGPT-generated essays.

3.5. Data Analysis

The final step in this research methodology involves a comprehensive analysis of the data collected from the evaluators' assessments to extract valuable insights regarding the quality of essays generated by ChatGPT. This step involves a variety of statistical techniques to ensure a comprehensive analysis of the collected data and the revelation of significant findings.

Initially, a descriptive analysis is conducted to identify the overall quality of the essays generated by ChatGPT. This analysis aims to pinpoint which elements of the essays are of the highest and lowest quality. By examining the scores across the six criteria, grammar and syntax, plagiarism, reference accuracy, content development, coherence and cohesion, and idea originality, which can be identified through patterns and trends in the data. This step provides a detailed overview of the strengths and weaknesses in the writing quality of ChatGPT-generated essays.

To ensure the reliability of the evaluation process, the inter-rater agreement between the two evaluators is measured. This assessment involves using Cohen's Kappa coefficient, a statistical measure that assesses the level of agreement between raters on categorical items. A substantial Cohen's Kappa value indicates strong agreement between the evaluators, confirming the coherence and credibility of the scoring process (Cohen, 1988). Another statistical method used to measure the internal consistency of a set of items is Cronbach’s alpha value (Cronbach, 1951). In interpreting Cronbach’s alpha test results, Nunnally (1967) initially proposed that values as low as 0.50 are suitable for exploratory research. In contrast, Hair et al. (2010) indicated that while a value of 0.70 is commonly accepted as satisfactory, figures as low as 0.60 may suffice for exploratory research purposes. On the other side, Cortina (1993) argued against basing scale adequacy solely on Cronbach’s alpha level, emphasizing that the acceptable reliability threshold depends on the intended use of the scale.

Following the descriptive analysis and inter-rater agreement assessment, the study employs Principal Component Analysis to identify underlying factors that contribute to the variance in the writing quality assessment scale. By reducing the dimensionality of the data, PCA simplifies intricate relationships among evaluation criteria, facilitating the interpretation of primary factors affecting writing quality.

Finally, correlation analysis is conducted to explore the relationships between various aspects of the essays and their overall quality scores. Specifically, this analysis investigates two key correlations: the correlation between themes of essays and average quality scores and the correlation between individual criteria scores and average quality scores. Exploring the connection between essay themes and the quality of ChatGPT-generated essays aims to uncover potential variations in writing quality based on different topics. Understanding this correlation can offer valuable insights into how essay prompts influence the quality of ChatGPT responses. On the other hand, identifying the influencing level of individual criteria scores on the general quality of writing could be beneficial in determining which aspects of writing are most critical to achieving high-quality essays and pinpointing the limitations of ChatGPT in generating academic essays.

4. Results

4.1. Overview

4.1.1. Themes of Essays

Table 2 demonstrates the descriptive data of essay themes among 72 questions from the Cambridge IELTS Book and their corresponding quality based on their thematic focus. The breakdown of essay themes reveals a notable emphasis on specific subjects.

Table 2: Descriptive Statistic of Essay Themes

		Frequency of Essays	Average Score	Percent of Frequency	Percent of Average Score
Valid	Economics	8	2.52	11.1	15.7
	Education	17	2.39	23.6	14.9
	Health	5	2.07	6.9	12.8
	ICT	8	2.25	11.1	14.0
	Lifestyle	18	2.23	25.0	13.9
	Work and Career	5	2.17	6.9	13.5
	Others	11	2.47	15.3	15.3
	Total	72	16.10	100	100

Despite not constituting the largest percentage of quantity, the *economics-related* theme garners the highest average score of 2.52, denoting outstanding performance within this category. This result shows good uniformity among articles on the topic of economics, reflected in 15.7% of the average score in all topics. *Lifestyle* stands out as the most frequent theme; however, its average score of 2.23 remains moderate, contributing 13.9% to the total average score. The same pattern was witnessed in *Education-related* essays, which emerged as the second most prevalent theme following *Lifestyle-related* essays. Although the number of articles on education and lifestyle topics accounts for twice as many economics-related articles, essays within these themes do not exhibit extraordinary average scores. The least common topic is *Healthcare*, which is worth noting that the essays on this topic also displayed the lowest average score of 2.07.

4.1.2. Writing Quality Criteria

Table 3 presents the data outcomes, utilizing SPSS 29.0 for a descriptive analysis of the average scores concerning the writing assessment criteria in ChatGPT-generated academic essays.

Overall, the analysis reveals that while ChatGPT can produce essays with solid grammar, syntax, and originality, there are notable weaknesses in coherence, cohesion, and content development.

With a mean score of 2.25 in terms of *Reference*, it is suggested that the essays generally provide adequate references, but there is quite a large variability, indicating an inconsistency in citation practices across different essays. Some of the citations might be fabricated in at least one of the three aspects, including the names of the authors, the titles of the articles, or the journal. It is worth noting that both *Coherence* and *Content Development* scored lowest, averaging 1.89 and 1.85, respectively. However, a wide score disparity was witnessed, which indicates a striking difference in the quality of logical flow and explicit connections between sentences or paragraphs. This impedes a prevalent challenge in maintaining logical progression and coherence between ideas within ChatGPT-generated essays. Moreover, the lower score in *Content Development* suggests a potential deficiency in elaboration and depth of argumentation, notwithstanding the creativity of ideas. This metric evaluates the quality of reasoning, necessitating sound rationale and concrete examples. While ChatGPT presents a coherent argument, it often lacks specific evidence to substantiate the central thesis or fails to delve deeply into the exploration and expansion of key concepts.

Table 3: Average Score of Writing Assessment Criteria

	N	Minimum	Maximum	Mean	Std. Deviation
Reference	72	1	3	2.25	.622
Coherence and Cohesion	72	0	3	1.89	.848
Idea Originality	72	0	3	2.24	.593
Content Development	72	1	3	1.85	.620
Grammar	72	2	3	2.85	.362
Plagiarism Avoidance	72	2	3	2.88	.333
Valid N (listwise)	72				

Conversely, the *Idea Originality* criterion achieves an average score of 2.24, indicating a noteworthy capability in articulating a distinct personal viewpoint on a given topic. In the realm of AI chatbots, the ability to generate original ideas signifies progress in the ongoing AI training endeavors. The substantially highest average *Grammar* and *Plagiarism Avoidance* scores were observed, highlighting ChatGPT's adeptness in underscoring its proficiency in

generating text with minimal grammatical errors and a strong emphasis on plagiarism avoidance. On the other hand, high scores on the plagiarism assessment criteria, with an average score of 2.88, can be inferred from the poor sensitivity of the plagiarism detector tools (Safrai & Orwig, 2024). The highest mean score of 2.85 on the *Grammar* criterion shows that the essays are generally well-written in terms of verb tenses, subject-verb agreement, and other grammatical types.

4.1.3. Credibility of Writing Quality Scale

The alpha coefficients of reliability for the writing quality criteria items ranged between 0.33 and 0.85, with an overall Cronbach's alpha of 0.67, indicating internal consistency across all six criteria (refer to Table 4). Furthermore, this section will explore the inter-rater reliability value to investigate the consistency between the assessment of the two raters.

The *Coherence and Cohesion* criterion exhibited high internal consistency with Cronbach's alpha of 0.85, suggesting that the 72 items are well-correlated and reliably assess this aspect of writing. The *Reference, Content Development*, and *Idea Originality* criteria, respectively, demonstrate moderate internal consistency with Cronbach's alpha values of 0.62, 0.60, and 0.62. Conversely, the Grammar criterion accounts for a low Cronbach's alpha of 0.36, and the Plagiarism Avoidance criterion occupies a similarly low alpha of 0.33. These low scores are attributable to ChatGPT's occasional subtle grammatical errors and minor instances of plagiarism. The low Cronbach's alpha value observed in assessing *Grammar* and *Plagiarism Avoidance* implies two potential implications. Firstly, it indicates that existing plagiarism detection tools may not effectively identify instances of plagiarism in content generated by ChatGPT. Secondly, it raises the possibility that the evaluation items employed may not consistently measure these constructs.

Table 4: Cronbach's alpha value of writing criteria items

Writing Assessment Criteria	Cronbach's alpha	Number of items
Reference	.62	72
Coherence and Cohesion	.85	72
Idea Originality	.60	72
Content Development	.62	72
Grammar	.36	72
Plagiarism Avoidance	.33	72
A total of six criteria	**.67**	**72**

Regarding the consistency in the assessment of the two evaluators, the inter-rater reliability score of 0.75 was established for the 72 total items, which indicates substantial agreement between the two raters assessing the ChatGPT-generated academic essays. This substantial level of agreement underscores the reliability and consistency of the evaluations, affirming the credibility of the scoring process.

4.2. Writing Quality of ChatGPT-generated academic essays

4.2.1. Quality of Reference

The criterion *Reference* assesses the accuracy of the references used in the essay. The existence of the names of authors, titles, and journals in ChatGPT's citations is evaluated. Referencing is considered a system of formal acknowledgment of the sources of other writers' words and thoughts. Moreover, the proper application of citations shows a set of skills requiring the understanding of other writers' work, being able to restate that understanding, having the intellectual confidence to admit another's precedence, and finally, mastering the control of a variety of tools for the proper display of this recognition (Borg, 2000).

The assessment of *Reference* in 72 essays produced by ChatGPT presented in Figure 5 reveals that a considerable portion of the essays, 40 in total, can be classified as moderate in quality. This finding suggests that while the references in these essays are satisfactory, enhancement needs to be made. Notably, a substantial number of essays, 25 in total, were rated as high quality, showcasing the effective and appropriate utilization of references. Conversely, essays in the poor and low-quality category of reference highlight the difficulties they might face in accurately integrating and citing sources. These results also suggested that several citations are fabricated by at least one of the three criteria: author's name, article title's name, or journal's name. These lower-rated essays, numbering 7 in total, underscore the challenges in ensuring consistent reference quality across all outputs.

In assessing ChatGPT's capability to cite information sources, three sub-aspects were suggested including author's name, article title, and journal name. These fundamental criteria are essential components within any research paper across various citation styles. Authorship recognition, as per APA guidelines (2020), acknowledges individuals who substantially contribute and take responsibility for a published work. Proper citation practices enhance transparency and academic integrity. Meanwhile, the title of an article serves as a succinct encapsulation of its subject matter and findings. Jamali and Nikzad (2011) elaborated on how the clarity and type of article titles can influence the visibility and citation rates of scholarly works, underscoring the significance of selecting informative and pertinent titles. Furthermore,

including the journal's name offers crucial contextual information and lends credibility to the referenced research. A comparative analysis of diverse citation tracking methods sheds light on how a journal's prestige and impact factor can impact citation metrics and scholarly discourse (Bakkalbasi et al., 2006).

The example provided in Figure 4 illustrates the process of verifying citations generated by ChatGPT. Initially, citations are inputted into Google in their complete citation form. The

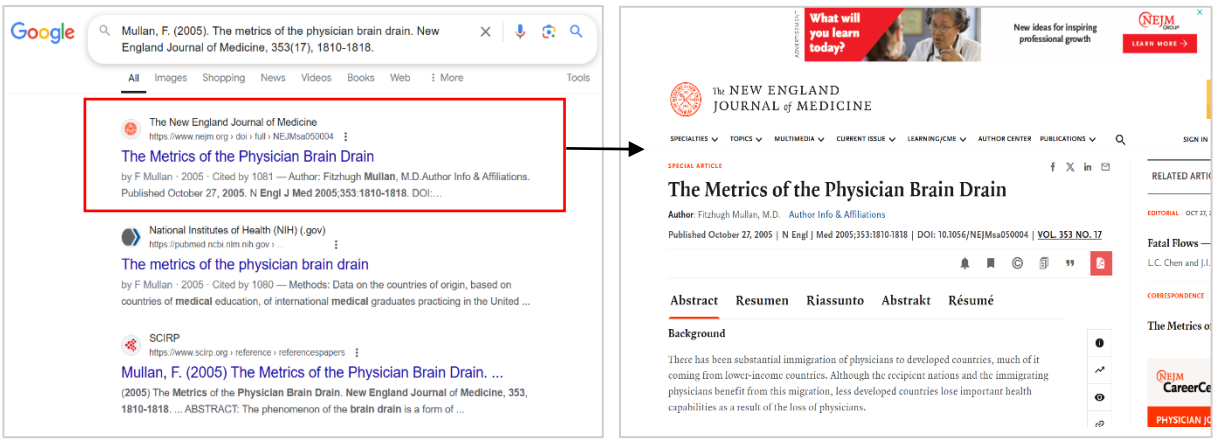

Figure 4: Example of verifying one citation produced by ChatGPT

reference list provided by ChatGPT was utilized to search for citations across various online platforms, including well-known resources such as Google, Google Scholar, Scopus, or SpiSpace. In cases where certain articles are not located during the initial search, they will be segmented into smaller components for a more targeted search approach. For instance, the citation proceeded by abbreviating the author's name and publication details, focusing solely on the article title. This process will be repeated until a match is found, emphasizing accuracy concerning the author's name, article title, and journal name. Conversely, failure in searching results that deviated from the intended keywords was witnessed. In such scenarios, evaluators may delve deeper by utilizing platforms such as Google Scholar and Scopus or refining their search query to continue searching. As the final search results reveal no citation matching the one mentioned in ChatGPT's essay, this suggests a high probability that ChatGPT might generate fabricated names for certain components, potentially undermining the credibility of the academic paper. If no citation is found, the search process will persist by either abbreviating the citation details or exploring alternative platforms for retrieval.

4.2.2. Quality of Coherence and Cohesion

Achieving cohesion in writing involves employing linguistic devices that connect ideas throughout a text, which is crucial for creating coherent and well-developed writing (Struthers

et al., 2013). As shown in Figure 5, the evaluation of the essays indicates a predominance of moderate to high-quality scores for coherence and cohesion. The majority of essays are categorized as moderate quality (n=26), while a notable amount achieved high quality (n=20). This suggests that ChatGPT-generated essays are generally characterized by well-structured and logically organized content, facilitating the clear articulation of ideas. However, it is noteworthy that the quantity of articles scored from 1 to 3 is relatively similar, with each category comprising around 20 articles. This finding highlights an inconsistency in the seamless flow and coherence of the ChatGPT articles.

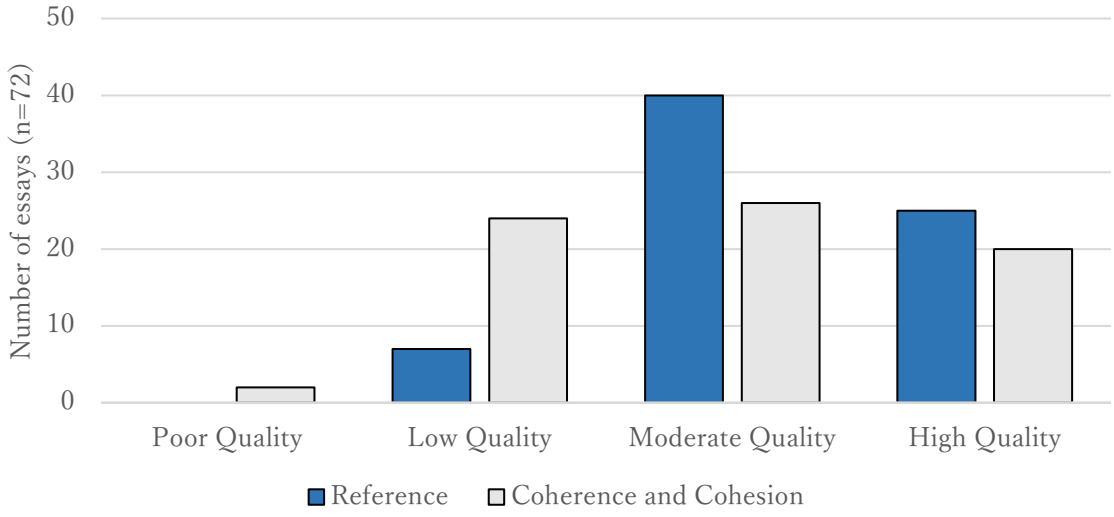

Figure 5: Quality of Reference and Quality of Coherence and Cohesion

Within the low-quality classification, a noticeable gap exists concerning references and interconnectedness. While five essays are classified as having substandard references, a notably higher count of 24 essays is categorized as low-quality in coherence and cohesion. This disparity suggests that while referencing quality may not be a significant issue, maintaining coherence and cohesion is a more prevalent challenge. Even essays with satisfactory references may encounter difficulties in establishing a logical flow and effectively connecting ideas. The majority of essays were assessed as having moderate quality concerning coherence and cohesion. This suggests that a significant portion of ChatGPT-generated essays excels in smoothly linking ideas within and between paragraphs. This detailed organization boosts the essays' argumentative strength and idea progression, guaranteeing clarity, readability, and effective communication of concepts.

Ensuring coherence and cohesion in a text leads to a logically structured and easily understandable piece, enabling readers to comprehend the connections between ideas and their impact on the main message or argument. The assessment of ChatGPT's quality of coherence and cohesion focuses on evaluating the logical connections within the text and analyzing how

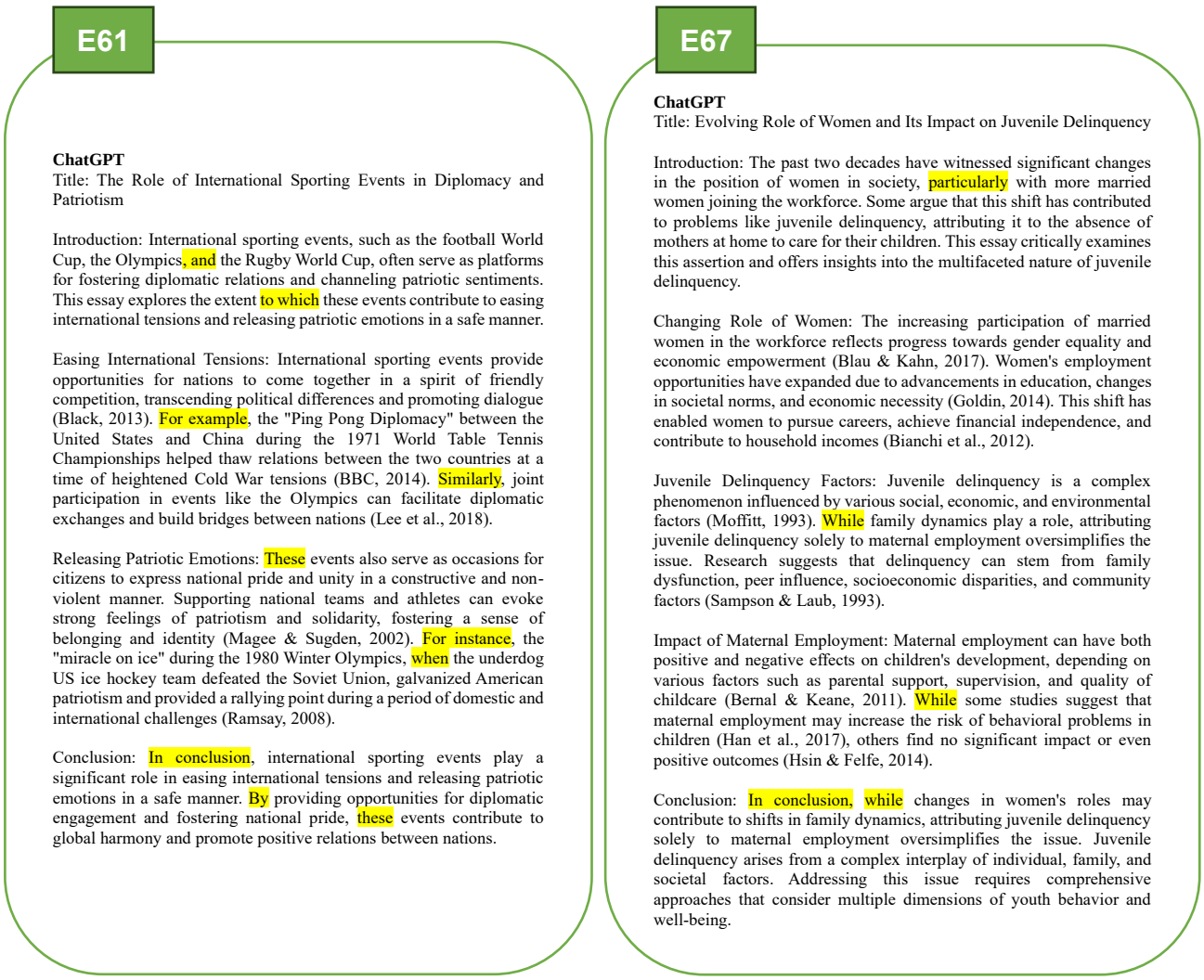

Figure 6: Example of identifying coherence and cohesion devices in ChatGPT-authored essays; (E61)-An essay with 3-score coherence and cohesion; (E67)-An essay with 1-score coherence and cohesion

ideas, sentences, and paragraphs flow together. The presence of cohesive devices plays a vital role in linking sentences and paragraphs. Figure 6 displays a good example of identifying the cohesion and coherence devices in written content. In Figure 6-E67, the use of linking words within sentences enhances intra-paragraph connectivity. However, some deficiencies are noted in the essay, such as the repetitive use of the conjunction "while" and the absence of connecting words between paragraphs. This highlights the limitations in linking the concepts and the inflexibility in connecting core ideas of ChatGPT, resulting in a perception of rigidity that can

be challenging for readers to comprehend. In contrast, Figure 6-E61 demonstrates a smooth and natural progression through local transitions. Notably, conjunctions such as "and," "for example," and "for instance" facilitate the expression of addition. The text also employs grammatical cohesive devices, including "these" and "which" for backreferencing and drawing comparisons between ideas or statements. The effective integration of these coherent devices induces a sense of seamless continuity in the text, enhancing its overall fluidity and cohesiveness.

4.2.3. Quality of Content Development

An evaluation scale to assess ChatGPT's content development capabilities was developed, focusing on three critical criteria: clarity of argument, incorporation of appropriate supporting ideas, and evidence substantiation. It is required that essays rated "high quality" of content development should encompass all three aspects within a single paragraph; failure to express each will result in a reduced score.

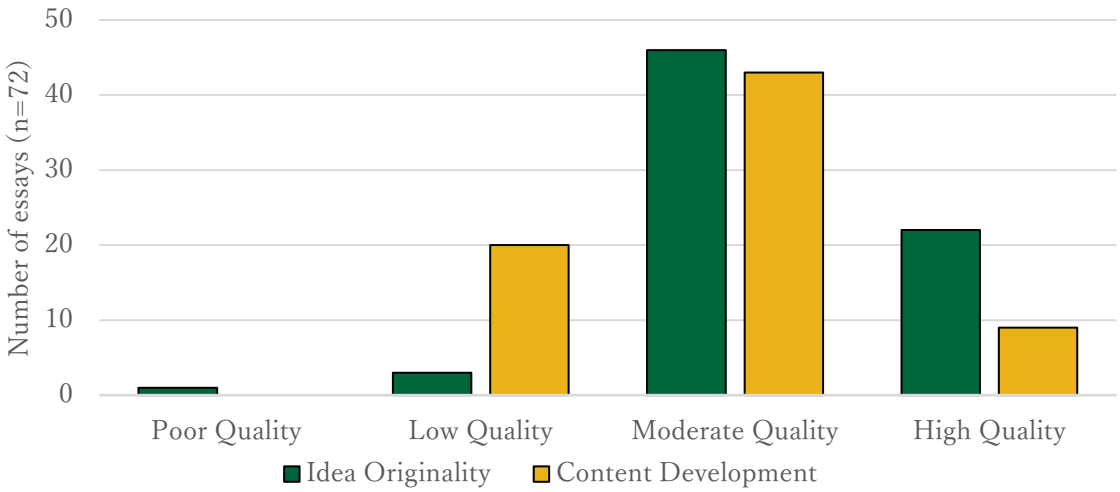

Figure 7: Quality of Idea Originality and Quality of Content Development

Figure 7 reveals a significant portion of essays (n=20) fell under the “low quality” category, suggesting that nearly 30% of the examined essays lack the requisite level of detailed elaboration and thoroughness expected in academic discourse. This result also highlights that despite ChatGPT's articles featuring clear arguments, they often lack the necessary supplementary ideas and compelling evidence to bolster the main concepts presented. On the other hand, the majority of the essays (n=43) were found to be of “moderate quality”, demonstrating clear main ideas, though they frequently did not excel in providing rich, insightful, or thoroughly developed arguments. A smaller fraction of essays (n=9) achieved

"high quality" status, showcasing well-organized, detailed, and insightful content development. This distribution highlights the variability in performance, revealing that while ChatGPT demonstrates a commendable capacity to craft paragraphs enriched with coherent arguments, supporting points, and corroborating evidence, a notable portion of articles do not consistently meet this standard.

Passage 1: Advancements in Medicine: Scientific research has played a pivotal role in improving healthcare **(1 Argument).** Milestones like the development of antibiotics, vaccines, and medical imaging technologies **(1 Example)** have significantly increased life expectancy and reduced mortality rates (World Health Organization, 2021). Such advancements directly enhance people's lives by promoting well-being and extending longevity **(1 Supportive Idea).**

Passage 1: 1 Argument + 1 Example + 1 Supportive Idea = 3 points

Passage 2: Environmental Sustainability: Science has a vital role in addressing environmental challenges such as climate change and resource depletion **(1 Argument).** Solutions derived from scientific research not only safeguard the planet but also ensure a better quality of life for current and future generations (Intergovernmental Panel on Climate Change, 2018). This demonstrates science's capacity to improve lives through ecological stewardship. **(2 Supportive Ideas).**

Passage 2: 1 Argument + 0 Example + 2 Supportive Ideas = 2 points

Figure 8: Example of evaluating Content Development quality

Figure 8 exhibits an example of assessing the quality of content development within an academic essay. Each essay underwent an evaluation based on the clarity and coherence of arguments, the inclusion of supporting ideas, and examples. An argument, defined as the primary claim or series of statements within the essay, is awarded one point per identifiable argument. A good example, which earns one point, should serve as concrete evidence that bolsters the main arguments' credibility. Supporting ideas, which garner one point, offer further elaboration on arguments or examples, providing additional context and depth. The first passage in Figure 8 met all the criteria, therefore scoring 3 points. Conversely, the second paragraph, lacking an example yet containing two supporting ideas, received a score of 2 points.

4.2.4. Quality of Idea Originality

The assessment of idea originality within 72 academic essays produced by ChatGPT was centered on ChatGPT's capacity to articulate its subjective viewpoint with cogent reasoning in the concluding segment, which was aimed at addressing the specified query comprehensively.

Based on the data analysis, only one essay is characterized as possessing "poor quality" in terms of idea originality, indicating that the majority of essays succeeded in incorporating a

degree of personal insight. A small yet discernible subset of three essays was classified as “low quality,” suggesting a deficiency in providing relevant reasoning to support their personal opinions. The largest portion, comprising 46 essays, was placed in the “moderate quality” bracket, indicating there is the presence of personal opinions in the conclusion, but they are either unclear or presented dilemmas. The articles classified as having moderate quality of idea originality often have ambiguous conclusions, neglecting to provide a direct response to the required question. Figure 9 presents an example of 2-point idea originality as a response to a prompt: “It is important for people to take risks, both in their professional lives and their personal lives. Do you think the advantages of taking risks outweigh the disadvantages?”. Given the requirement posed by the question, it is expected that ChatGPT will provide a clear personal perspective at the conclusion of the essay on whether taking risks brings disadvantages or advantages to human life. However, rather than presenting definitive perspectives, it tends to use vague terms such as "balance" and "both," thus avoiding a decisive resolution.

Conclusion

In conclusion, the advantages and disadvantages of taking risks are intertwined, and **the balance between** them depends on the context and individual circumstances. **While** risk-taking promotes innovation, learning, and enhanced decision-making, **it also** entails the potential for failure, stress, and social repercussions. Ultimately, the judicious assessment of risks, coupled with a willingness to learn and adapt, is essential for individuals seeking to navigate the complex interplay between risk and reward in **both their professional and personal lives**.

Unclear and dilemma opinion => 2 points

Figure 9: Example of moderate quality in Idea Originality

On the other hand, a considerable number of essays (n=22) were ranked as “high quality” concerning idea originality. They showcased well-structured, insightful, and notably original personal viewpoints that substantially enriched the overall quality of the essays. This breakdown accentuates ChatGPT's proficiency in generating a considerable percentage of essays with distinctive and well-founded conclusions.

4.2.5. Quality of Plagiarism Avoidance

Plagiarism evaluation of ChatGPT-authored essays aligns with the updated regulations in India (Kadam, 2018; Vandana & Nagaveni, 2019). According to these guidelines, ChatGPT-generated essays with a Grammarly-detected plagiarism index of 10% or lower will be classified as demonstrating a high level (3 points) of plagiarism avoidance. Essays falling within the 10%-40% range will be rated with 2 points, while those above 40% to 60% range will receive 1 point. Essays exceeding 60% plagiarism will be deemed of "poor quality" and awarded 0 points.

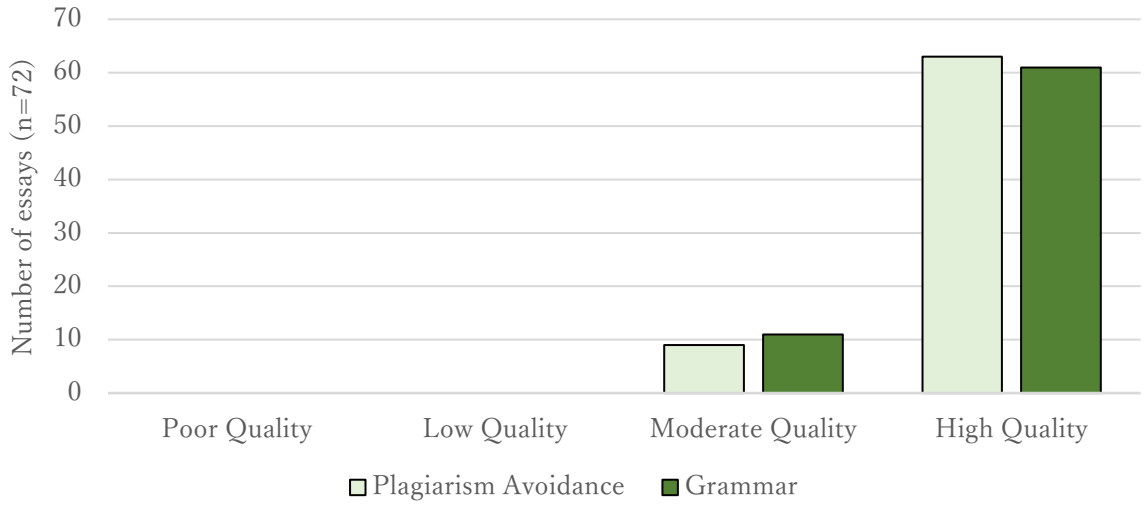

Figure 10: Quality of Plagiarism Avoidance and Quality of Grammar

The data shown in Figure 10 suggests that the vast majority of the essays (n=63) were categorized as “high quality” in plagiarism avoidance. Only 9 articles were rated “moderate quality” with 2 plagiarism points, and none were rated “low quality” or “poor quality”, which contradicts previous observations about the plagiarism issues of ChatGPT’s writing (Baidoo-Anu & Ansah, 2023; Akinrinola et al., 2024; Abbott& Rothman, 2023). However, this also raises another concern regarding the efficacy of plagiarism detection tools when assessing essays generated by machines for potential instances of plagiarism. Upon reviewing Table 5, several instances of plagiarism detection within ChatGPT-generated essays assessed by Grammarly were shown. The detection rate for each essay stands at around 10%. However, it is essential to recognize that most of the potential plagiarism segments are components of a sentence rather than complete sentences. As a result, these identified clusters of text cannot be definitively categorized as either factual assertions or opinions. Consequently, the presence of these segments alone does not suffice to assert a violation of plagiarism.

Table 5: Plagiarism detection report by Grammarly

Essay code	% of plagiarism	Plagiarized sentences	Reference title
E21	9%	"Living in a country where one must speak a foreign language can"	TOEFL Writing- Leverage Edu. https://leverageedu.com/learn/toefl-writing-topic-it-is-beneficial-for-people-to-spend-some-time-living-in-a-country-where-they-must-speak-a-foreign-language/
		"within communities, leading to feelings of isolation and"	Where Does Mercy Come From? - Time News Global. https://timenewsglobal.com/business/where-does-mercy-come-from/
E25	12%	"whether children should be allowed to make their own choices on everyday matters is"	IELTS Writing Task 2 Sample– pteielts.com. https://www.pteielts.com/some-people-believe-that-allowing-children-to-make-their-own-choices-on-everyday-matters/
		"healthy relationships built on trust and mutual respect."	Bazylak, D. (2002). A study of factors contributing to the success of female Aboriginal students in an inner city high school. https://core.ac.uk/download/226159047.pdf
		"By striking a balance between guidance and freedom, parents"	Adolescence Problems (Understanding & Solutions). https://tagvault.org/blog/adolescence-problems/
E44	11%	"parents are the primary caregivers and role models for their children,"	What are the crucial role of parents in the child development?. https://www.thinkingineducating.com/what-are-the-crucial-role-of-parents-in-the-child-development/
		"with the knowledge and skills necessary to participate in democratic processes,"	20 Reasons Why Education Is Important - Education Guidez. https://educationguidez.com/20-reasons-why-education-is-important/
		"can create a supportive and enriching environment that fosters the"	Employing an Early Years Apprentice in Your Nursery. https://scopetraining.co.uk/employing-an-early-years-apprentice-in-your-nursery/

4.2.6. Quality of Grammar

The assessment of grammar quality in the 72 academic essays indicates a strong adherence to grammatical correctness and proper syntax. The data in Figure 10 shows that most essays (n=61) were rated as "high quality" in terms of grammar, reflecting the model's proficiency in producing well-structured and error-free sentences. These essays consistently demonstrated proper syntax, punctuation, and overall linguistic accuracy. A smaller subset of essays (n=11) was categorized as "moderate quality", indicating the presence of occasional grammatical errors or less-than-optimal sentence structures that did not significantly hinder readability but were noticeable. Importantly, no essays were found in the "poor quality" or "low quality" categories, highlighting ChatGPT's capability to generate grammatically academic content consistently.

Combining the conclusions from the evaluations of plagiarism avoidance and grammar

quality, it is evident that ChatGPT excels in generating academic essays with high standards of originality and linguistic accuracy. Approximately 90% achieved “high quality” ratings in both plagiarism avoidance and grammar, demonstrating a strong ability to produce content with minimal grammatical errors. The consistently high scores in grammar align with previous statements about ChatGPT's ability to assist non-native English students in learning proper grammar usage and syntax, improving their writing skills over time (Schmidt-Fajlik, 2023; Li & Zhang, 2020). This finding also suggests that integrating AI writing tools such as ChatGPT into language learning curricula could enhance the learning experience for ESL students.

4.3. Influential Factors on the Quality of ChatGPT-generated Academic Essays

4.3.1. Underlying Factors in Writing Quality Assessment Scale

A principal component analysis (PCA) of the six assessment criteria was undertaken to establish whether the Writing Assessment Scale could assess one or more components of writing quality. The PCA was run on the six criteria in the Writing Assessment Scale for 72 academic essays generated by ChatGPT. The PCA result revealed two components that had eigenvalues greater than one. The two components explained 38.431% and 26% of the total variance, respectively.

As shown in Table 6, the first component captures a significant portion of the variability in the data and encompasses variables related to four criteria, including Coherence and Cohesion, Originality, Content Development, and Reference. This explained that the quality of the essay can be attributed to how well the content is organized, the individualism of ideas presented, and the effectiveness of supporting those ideas with references. In terms of the definition of each criterion, coherence is inherently linked to logical skills as it requires the writer to organize ideas in a manner that makes sense to the reader. Coherent writing displays clear, logical progression from one idea to another, which is essential for effective communication in academic and professional contexts (Todd et al., 2007). Another criterion is Reference, which refers to the action of acknowledging and applying existing research to new ones. Referencing, hence, signifies the ability to utilize pre-existing knowledge to address novel challenges. This process entails the skill of using information in new situations, which fits the feature of applying skill as suggested in Bloom’s Taxonomy (1956). On the other hand, Idea Originality requires individualism in concepts and embodies creative personal thought, a key element in effective writing. Guilford's model of creativity (1967) emphasized the significance of divergent thinking, where individuals generate multiple solutions to a given issue. This form of thinking is vital for producing fresh ideas and viewpoints that can lead to

groundbreaking solutions. Finally, content development refers to identifying components of an argument, determining the validity of claims, supporting arguments with evidence, and evaluating reasoning (Marttunen, 1992). Therefore, all of these criteria belong to the category of "Idea Depth and Structural Integrity", which encompass a set of higher-order thinking processes that include information-processing skills, reasoning skills, enquiry skills, creative thinking skills and evaluation skills (Bloom, 1956; QCA, 2000).

Table 6: Component Matrix for PCA of a two-component solution

Writing Assessment Criteria	Component 1	Component 2
Coherence and Cohesion	.720	
Idea Originality	.713	
Content Development	.662	
Reference	.660	
Grammar		.828
Plagiarism Avoidance		.746

Extraction Method: Principal Component Analysis.[a]
a. 2 components extracted.

The second component, including *Grammar and Plagiarism Avoidance*, attaining an eigenvalue of 1.564, accounts for 26% of the total variance. This component is categorized as "Writing Technique and Manner", represented by grammatical accuracy and the ability to avoid plagiarism. This dimension emphasized the importance of linguistic precision and adherence to academic standards in determining the quality of the essays. The result could be explained as both grammatical correctness and plagiarism avoidance are fundamental in crafting lucid, trustworthy, and ethical academic content. Grammar is typically one of the initial stages in gaining a foundational understanding of a language. It acts as the essential component that enables the coherence and clarity of written text, allowing readers to engage seamlessly with the content (Strunk & White, 2000). Conversely, disregarding proper grammar could lead to confusion, hindering effective communication. Moreover, maintaining integrity in writing involves steering clear of plagiarism, a fundamental principle that upholds the originality and credibility of one's work. Upholding academic standards at a commendable level ensures not only quality but also fosters a culture of unwavering integrity (Brown & Janssen, 2017). The fusion of these standards underscores the harmonious interplay between exact technicality and ethical principles essential for creating exceptional academic works.

The cumulative variance explained by these two components is 64.505%, indicating that these two dimensions provide a comprehensive overview of the factors contributing to essay quality. The high loadings on the respective components suggest clear separations between structural integrity and writing techniques and reveal the importance of these two key factors in the quality of academic essays written by ChatGPT.

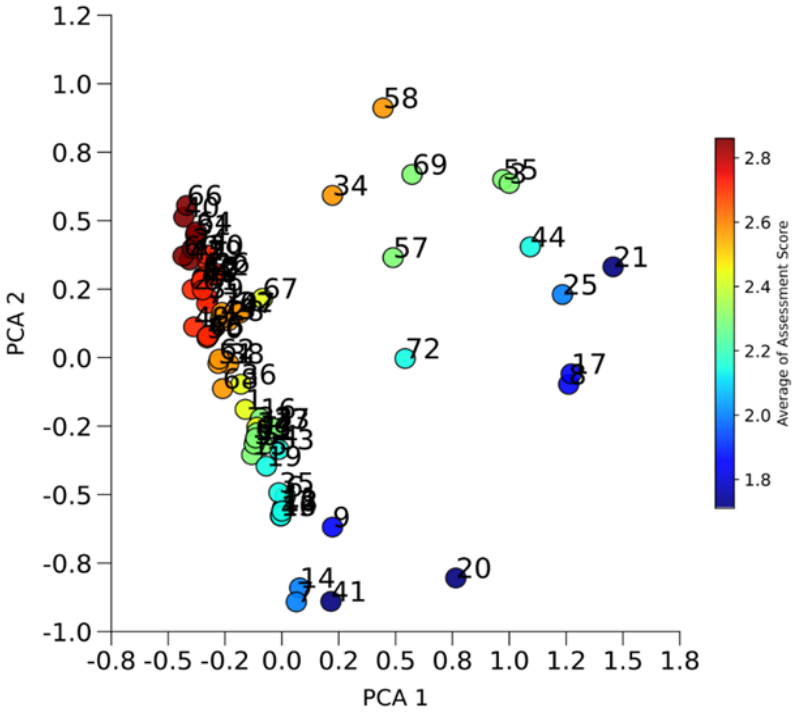

Figure 11: Distribution of Assessment Scores for ChatGPT-Generated Academic Essays

The PCA biplot presented in Figure 11 illustrates the distribution of the 72 ChatGPT-generated academic essays based on the two principal components extracted from the analysis. The horizontal axis PCA 1 predominantly captures the variance associated with the "Idea Depth and Structural Integrity" dimension, while the vertical axis PCA 2 captures the variance related to "writing technique and manner". Each point on the plot shows the average score of a particular essay, with the numbers representing the essay identifier. The color gradient, shifting from blue to red, reflects the average assessment scores, where blue represents lower scores and red represents higher scores. This color coding aids in presenting the overall quality of the essays based on the established criteria.

Essays clustered towards the left side of the plot (negative PCA 1 values) tend to have lower scores, suggesting they are weaker in generating thinking skills-required content. Conversely, essays positioned towards the right side (positive PCA 1 values) generally exhibit higher scores, indicating stronger content. Similarly, essays higher up on the plot (positive PCA

2 values) demonstrate better technical correctness and manner, while those lower down (negative PCA 2 values) are weaker in these aspects. The densest cluster is located around (-0.3, 0.6) to (-0.1, -0.3) in PCA space. This indicates that a significant number of essays have similar feature patterns. There are isolated groups of essays, particularly around (1.2, 0.2), (0.8, -0.8), and (0.4, 0.9), which suggests these essays have unique feature sets distinguishing them from the main cluster. As can be seen from Figure 11, essays with ID numbers 20, 21, 17, 41, and 9 are extreme points in the PCA space, suggesting they have unique characteristics not shared by the majority. They could be significant outliers with very low scores and distinct features, suggesting potential issues such as lack of content, poor structure, or other significant deviations from scoring criteria. Interestingly, these points are predominantly blue, indicating lower assessment scores. The clustering pattern reveals that a majority of the essays are concentrated towards the center-left of the plot, indicating a trend where most essays exhibit moderate content quality but varying degrees of technical correctness. Notably, some outliers, such as essays 58 and 69, are positioned towards the top-right, highlighting them as exemplary essays with both high content quality and technical correctness. Essay 72, located at (0.6, 0.0), and Essay 69 at (0.5, 0.7) are somewhat isolated but have relatively higher scores.

To provide a comprehensive analysis, Essays 20, 21, and 66 were specifically selected based on their distinguished positions in the PCA scatter plot to illustrate the variance in the essays' features visually (see Figure 12). The PCA depiction underlines the considerable separation among these essays, indicating significant differences in their attributes and characteristics. This divergence implies varying levels of performance and features. This separation suggests diverse performance levels and assessment criteria, making them ideal examples for understanding the range of essay quality and the underlying factors contributing to their scores. Examining essays that are far apart in the PCA space aims to uncover insights into the various strengths and weaknesses that influence assessment outcomes, thereby providing a comprehensive view of the factors that impact essay quality.

Essay 66 stands out as a high-quality composition that consistently excels across various criteria, positioning it among the top scorers on the PCA scale, as presented in Figure 11. In contrast, essays 20 and 21 place close to two different extremes of the chart, indicating notable score variations between these two papers. Among the analyzed essays, Essay 20 received the lowest score of 0 points in the criterion of Coherence and Cohesion, indicating notable issues with logical progression and idea linkage that can impede comprehension of the essay. Conversely, Essay 66 and Essay 21 achieved higher scores of 2 points, reflecting a more

effectively organized structure that enhances clarity.

In terms of Idea Originality, Essay 66 attained the highest ratings in idea originality (3 points) and content development (3 points), showcasing its capacity to offer its own perspective

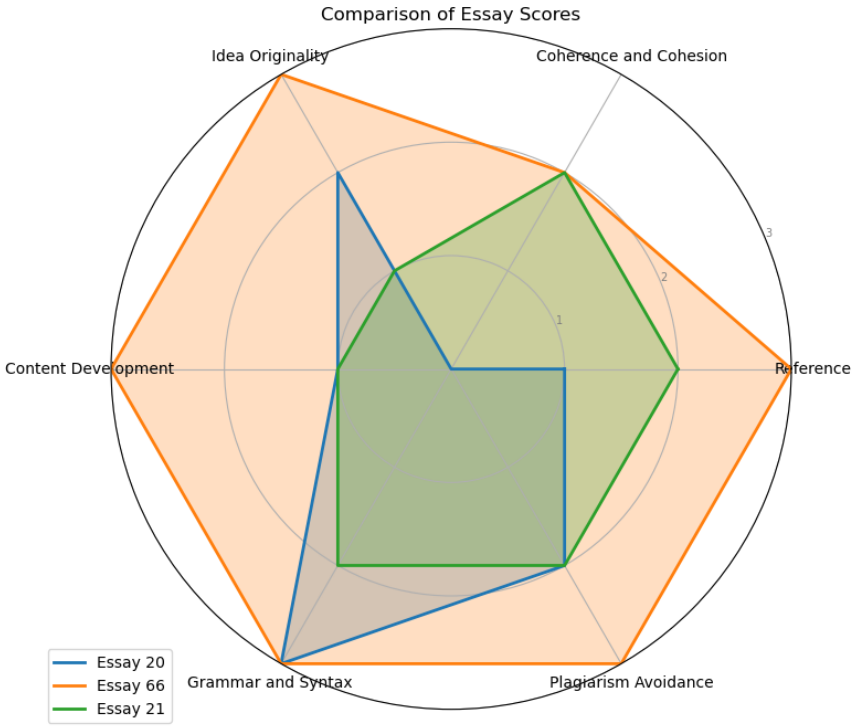

Figure 12: Comparison of score among distinguished essays

when wrapping ideas and elaborating on its content thoroughly through supportive ideas and examples. In contrast, Essay 20 and Essay 21 scored lower in content development (1 point each) and diverged in terms of idea originality, with Essay 20 scoring 2 points and Essay 21 scoring 1 point. This discrepancy suggests that these essays may lack depth in layers of argumentation and creativity in articulating concepts. Additionally, regarding writing technique and manner, all three essays appear to attain high scores for each criterion.

In conclusion, Essay 66 can be considered an exemplary case because of its comprehensive and well-developed content, originality, and linguistic proficiency, resulting in the highest overall score. Essay 21, while better organized than Essay 20, lacks ChatGPT's perspective in conclusion and depth of content. Essay 20, possessing the weakest quality compared to the others, faces considerable challenges in structure and content, contributing to its lower performance.

4.3.2. Factors influencing the writing quality of ChatGPT-generated

academic essays

In this section, three primary factors that may influence the quality of ChatGPT-generated academic essays are examined, including essay themes, thinking dimensions, and writing technique and manner dimension. Each factor is analyzed to understand its impact on overall writing quality.

4.3.2.1. Themes of essays

The study utilized a one-way analysis of variance (ANOVA) to evaluate the writing quality of ChatGPT across seven distinct topics. Initially, Levene's Test for Homogeneity of Variances was employed to confirm uniformity among the groups. The results indicated no significant differences in variances, with all test values exceeding 0.05. Subsequently, the ANOVA test was conducted, revealing an F-value of 1.711 and a p-value of 0.133. Given that the p-value was greater than 0.05, the null hypothesis could not be rejected. Therefore, no statistically significant variations in mean essay scores were observed across the seven topics.

Table 7: ANOVA test result between themes of essays

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.272	6	.212	1.711	.133
Within Groups	8.057	65	.124		
Total	9.330	71			

Overall, the homogeneity of variances test indicates that the variance in essay quality is consistent across different themes. The ANOVA results further suggest no significant differences in the average scores of essays among the various themes. This implies that the quality of ChatGPT-generated essays is relatively uniform regardless of the thematic category.

4.3.2.2. Idea Depth and Structural Integrity dimension

An examination was conducted on the correlations between different cognitive skills and the overall quality of writing. Table 8 shows the outcome of the correlation analysis, which demonstrated that all assessed criteria, including the use of reference, coherence and cohesion, idea originality, and content development, positively correlate with the average score of writing quality.

Table 8: Correlation between Idea Depth and Structural Integrity Dimension and Writing Quality

	Reference	Coherence	Idea	Content	Average

			&Cohesion	Originality	Development	Score
Reference	Pearson Correlation	1	.480**	.410**	.319**	.677**
	Sig. (2-tailed)		<.001	<.001	.006	<.001
	N	72	72	72	72	72
Coherence &Cohesion	Pearson Correlation	.480**	1	.333**	.422**	.776**
	Sig. (2-tailed)	<.001		.004	<.001	<.001
	N	72	72	72	72	72
Idea Originality	Pearson Correlation	.410**	.333**	1	.367**	.686**
	Sig. (2-tailed)	<.001	.004		.002	<.001
	N	72	72	72	72	72
Content Development	Pearson Correlation	.319**	.422**	.367**	1	.673**
	Sig. (2-tailed)	.006	<.001	.002		<.001
	N	72	72	72	72	72
Average Score	Pearson Correlation	.677**	.776**	.686**	.673**	1
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	
	N	72	72	72	72	72

**. Correlation is significant at the 0.01 level (2-tailed).

Among the analyzed criteria, *Coherence and Cohesion* emerged as the factors displaying the strongest positive correlation with a coefficient of r=0.776. This indicates that an increase in an essay's coherence paralleled an improvement in its overall quality score. The statistical significance level of $p<0.001$ confirms the credibility of this relationship, indicating that it is highly improbable to have occurred by chance. The substantial correlation underscores the pivotal role of coherence in determining the quality of writing. Well-structured and logically organized essays tend to garner superior assessments. The strong positive correlation observed between referencing frequency and the mean score ($r = 0.677$, $p < 0.001$) suggests that this criterion played a pivotal role in influencing the caliber of academic essays produced through ChatGPT. This finding implies that the integration of precise and reliable references markedly improved the quality of writing. The adept utilization of references not only enhances the credibility of arguments but also showcases the writer's proficiency in interacting with established literature, a fundamental element in academic writing.

The originality of ideas presents a notable correlation with the average score ($r = 0.686$, $p < 0.001$), indicating that ChatGPT-authored essays featuring their own concepts tended to achieve higher scores. This highlights the significance of creativity and analytical thinking in scholarly writing. Unique ideas captivate readers and engage readers, which can contribute the advancement of knowledge, reflecting the importance of originality in writing assessments (Flower & Hayes, 1981).

Content development is also positively correlated with the average score ($r = 0.673$, $p < 0.001$), indicating that essays with well-developed content with layers of argumentation receive higher evaluations. This adjoins the finding of previous study by Cumming et al. (2002) that well-developed content demonstrates a thorough understanding and exploration of the topic, which is essential for high-quality writing.

The finding also witnesses significant correlations among the criteria themselves. For example, *Coherence* and *Reference* ($r = 0.48$, $p < 0.001$); *Idea Originality* and *Reference* ($r = 0.41$, $p < 0.001$); *Content Development* and *Idea Originality* ($r = 0.422$, $p < 0.001$). These correlations indicate that improvements in one area of writing quality are often associated with enhancements in other areas, suggesting an interconnected framework of writing skills. For instance, a well-referenced essay is likely to exhibit better coherence, as the incorporation of credible sources can enhance the logical flow and support of arguments. Similarly, clear personal perspectives contribute to the overall content development, indicating that creativity and depth of thought are crucial for comprehensive essay writing.

4.3.2.3. Writing Technique and Manner Dimension

The findings displayed in Table 9 highlight notable correlations, demonstrating that both grammatical accuracy ($r = 0.346$, $p < 0.001$) and the reduction of plagiarism ($r = 0.437$, $p < 0.001$) are positively associated with overall writing quality despite its low degree. The observed moderate correlation between grammar and overall writing scores indicates that adherence to grammatical norms enhances the clarity and coherence of written material. Essays with a strong grammatical foundation are generally more understandable, leading to higher overall evaluations. Conversely, minimizing plagiarism is crucial for improving writing quality. Original content reflects a thorough grasp of the subject and a commitment to academic honesty. Additionally, the analysis shows a substantial positive correlation between grammar and plagiarism ($r = 0.657$, $p < 0.001$), suggesting that effective grammar usage is often linked with reduced plagiarism. This correlation implies that well-structured, grammatically accurate

essays are likely to show lower levels of plagiarism, indicating greater originality and adherence to academic standards. This connection emphasizes the importance of grammatical precision as both a mark of skilled writing and ethical academic practice.

Table 9: Correlation between Writing Technique and Manner Dimension and Writing Quality

		Grammar	Plagiarism	Average Score
Grammar	Pearson Correlation	1	.657**	.346**
	Sig. (2-tailed)		<.001	.003
	N	72	72	72
Plagiarism	Pearson Correlation	.657**	1	.437**
	Sig. (2-tailed)	<.001		<.001
	N	72	72	72
Average Score	Pearson Correlation	.346**	.437**	1
	Sig. (2-tailed)	.003	<.001	
	N	72	72	72

**. Correlation is significant at the 0.01 level (2-tailed).

5. Discussion and Conclusion

The main goal of the present study is to evaluate the quality of ChatGPT's academic writing through two sub-objectives: develop a writing assessment scale to measure the quality of writing and analyze factors affecting the quality of writing. there. In this chapter, key findings in chapter four will be discussed in line with the limitations of the study and suggestions for future research.

5.1. Writing Quality Criteria

5.1.1. Grammar and Coherence

The assessment of grammar quality in the ChatGPT-generated Essays reveals a high level of grammatical accuracy, with 61 out of 72 essays rated as “high quality”. This finding aligns with previous research highlighting ChatGPT's proficiency in grammatical accuracy, confirming its ability to support English learners' fluency development and individualize language acquisition (Meniado, 2023; Zhang, 2024; Imran & Lashari, 2023). On the other hand, the role of coherence and cohesion in academic essays generated by ChatGPT, measured through coherence and cohesion scores, is equally critical in enhancing essay quality. Essays with high coherence scores demonstrate logical flow and connectivity and contribute significantly to the readability and overall impact. In this study, the number of articles achieving fair quality or higher in the coherence and cohesion section accounted for 64% of the total number of articles, demonstrating weaknesses in the cohesion and coherence of the content of the essays. Some essays not on this list often lack connection and depth between paragraphs and simply present previously trained arguments. With this same view, in the study comparing narrative writing performance by ChatGPT and Chinese immediate English learners (2023), Zhou and his partners indicated that ChatGPT performed better than the Chinese students in word concreteness and referential cohesion but worse in syntactic simplicity and depth coherence.

5.1.2. Idea Originality and Content Development

The relationship between the originality of ideas and content development stands as a pivotal aspect of academic writing. The assessment in this study revealed that a significant portion of the essays exhibited a moderate quality in articulating personal viewpoints on various topics. Nonetheless, it became apparent that while ChatGPT can furnish a comprehensive essay conclusion, it lacks the capacity to offer a subjective stance on issues necessitating a choice between alternatives. Essentially, this prevalent pattern emerged in ChatGPT-generated essays, revealing a deficiency in creativity and individualism necessary to arrive at a cogent and

reasoned conclusion. ChatGPT underwent training on a vast array of textual sources from the internet, encompassing literature from books, articles, and websites and spanning diverse subjects such as news, Wikipedia entries, and works of fiction. Unlike humans, ChatGPT lacks the innate ability to engage in natural language acquisition and true creativity (Chomsky et al., 2023). Similarly, the utility of ChatGPT within the realm of AI-assisted academic writing cannot originate ideas, as noted by Mahama et al. (2023).

The analysis of content development in the ChatGPT-generated essays reveals a notable limitation in the depth and layering of argumentation. Specifically, 63 out of 72 essays were found to lack one of the two critical components of argumentation: supportive ideas and examples. This deficiency indicates a significant gap in the robustness of the essays' arguments, as effective argumentation typically requires both evidence and detailed elaboration to convincingly support main points. This issue is consistent with findings from other studies that highlight the challenges AI-generated content faces in producing nuanced and deeply reasoned arguments. For instance, Bender et al. (2021) discussed the "stochastic parrot" problem, where AI systems similar to GPT-3 generate text that superficially resembles human writing but often lacks the depth and contextual understanding needed for substantive argumentation. While the essays may successfully present main arguments, the connections between these points are often weak or nonexistent, leading to a disjointed structure. The lack of internal logic and seamless transitions between arguments further detract from the persuasiveness and clarity of the writing.

5.1.3. Plagiarism Avoidance and Reference

ChatGPT operates based on user instructions to provide engaging information tailored to individual queries. This study highlights the limitations of referencing in ChatGPT-generated content, which aligns with the findings of Safrai and Orwig (2024). Their paper revealed that only 9 out of 25 references produced by ChatGPT were accurate. Conversely, regarding the plagiarism level, this research demonstrate that ChatGPT is effective at reducing plagiarism, with more than 90% of essays being classified as "high quality" in terms of plagiarism prevention, showing less than 10% similarity to external sources. Supporting evidence from various studies, such as those by Khalil (2023), Hsu et al. (2023), and Safrai and Orwig (2024), aligns with these findings, confirming ChatGPT's low levels of plagiarism as identified by platforms namely Grammarly, Turnitin, iThenticate, and Quetext. However, this impedes a concern regarding the reliability of current plagiarism detection software. While adept at identifying direct copying, these tools may struggle with subtler forms of plagiarism, such as

inadequately attributed paraphrasing. The emergence of AI-generated content necessitates reevaluating existing plagiarism detection frameworks to accommodate the intricate ways in which AI can emulate ideas without explicit replication (Safrai & Orwig, 2024). Consequently, a more comprehensive approach to upholding academic integrity is imperative to address the evolving challenges posed by AI technology.

5.2. Limitations of this study and recommendations for future research

A significant challenge within this research lies in the rapid evolution of AI technology. The emergence of new iterations of ChatGPT heralds discernible improvements in writing proficiency. OpenAI continuously refines its models through enhanced algorithms, expanded datasets, and refined training techniques. These enhancements have the potential to markedly enhance the grammatical accuracy, coherence, and overall quality of AI-generated content. Consequently, the insights gleaned today may not endure as these advancements influence forthcoming iterations of the model. For instance, issues identified in this study, such as occasional lapses in coherence or limitations in originality, could be mitigated in future versions. As a result, the performance metrics and correlations observed in this research might differ when newer, more advanced versions of ChatGPT are evaluated. Future research should explore longitudinal studies to track the development of AI-generated writing over time and conduct comparative analyses with other AI models and human writers across various academic disciplines. This will assist in understanding how improvements in AI technology impact the quality of academic writing and whether the identified influential factors remain consistent.

Another limitation of this study is its exclusive focus on English-language essays. While this focus provides valuable insights into the capabilities of ChatGPT in generating academic content in English, it overlooks the model's performance across other languages. By concentrating solely on English essays, this study does not account for the linguistic and cultural nuances that might influence writing quality in other languages. Every language has its own set of unique grammatical rules, idioms, and argument structures that impact the quality of AI-generated text. To tackle this issue, future studies should focus on a wider array of languages to evaluate AI-generated essays. By assessing performance across different languages, researchers can gain deeper insights into how ChatGPT functions within diverse linguistic environments.

The evaluation framework used in this research highlighted certain gaps, particularly in

the areas of grammar and the avoidance of plagiarism. These are key factors in assessing academic writing, but they can pose challenges when dealing with AI-generated content. While ChatGPT generally performs well in terms of grammar, there are instances where it may generate sentences that, although grammatically correct, feel awkward or misplaced. Traditional grammar checkers for human-written essays might not capture these subtleties well, potentially leading to an inflated view of the model's grammatical skill. Moreover, existing tools excel at identifying direct plagiarism by comparing text with known sources but struggle to detect subtler forms of academic dishonesty, for example, minor paraphrasing or failing to credit ideas properly. Given that ChatGPT learns from a wide array of online texts, it could unintentionally generate content resembling specific sources, complicating the verification of originality. Future studies should explore developing more sophisticated assessment techniques to effectively tackle grammar and originality concerns in AI-generated content.

In conclusion, despite those limitations, the findings signify an important step in evaluating the writing proficiency of ChatGPT within an academic context. This study provides another aspect of AI application in education, furnishing valuable insights that can assist educators and students in better integrating these tools into their teaching and learning strategies. Furthermore, this study has implications for school policies that encourage the responsible use of AI tools, ensuring that these technologies are used to enhance, rather than replace, traditional learning methods.

Bibliography

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). https://doi.org/10.1037/0000165-000

Aaron, P. G., & Joshi, R. M. (2006). Written language is as natural as spoken language: A biolinguistic perspective. *Reading Psychology*, *27*(4), 263-311.

Abbott, R., & Rothman, E. (2023). Disrupting creativity: Copyright law in the age of generative artificial intelligence. *Fla. L. Rev.*, *75*, 1141.

Akinrinola, O., Okoye, C. C., Ofodile, O. C., & Ugochukwu, C. E. (2024). Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability. *GSC Advanced Research and Reviews, 18*(3), 050-058.

Aljuaid, H. (2024). The Impact of Artificial Intelligence Tools on Academic Writing Instruction in Higher Education: A Systematic Review. *Arab World English Journal (AWEJ) Special Issue on ChatGPT*.). The Impact of Artificial Intelligence Tools on Academic Writing Instruction in Higher Education: A Systematic Review. *Arab World English Journal (AWEJ) Special Issue on ChatGPT*.

Bae, J., Bentler, P. M., & Lee, Y. S. (2016). On the Role of Content in Writing Assessment. *Language Assessment Quarterly*, *13*(4), 302–328. https://doi.org/10.1080/15434303.2016.1246552

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52-62.

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries, 3*, 7. https://doi.org/10.1186/1742-5581-3-7

Barton, D., & Hamilton, M. (2012). *Local literacies: Reading and writing in one community*. routledge.

Bathaee, Y. (2020). Artificial intelligence opinion liability. *Berkeley Technology Law Journal, 35*(1), 113-170.

Benade, L., Stewart, G. T., & Devine, N. (2021). Writing for Various Academic Purposes and Genres. *Writing for Publication: Liminal Reflections for Academics*, 1-15.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21).

Bereiter, C., & Scardamalia, M. (1987). *The Psychology of Written Composition*. Lawrence Erlbaum Associates.

Bhuyar, V., & Deshmukh, S. N. (2023). Analysis of Support Tools for Plagiarism Detection. In *International Conference on Applications of Machine Intelligence and Data*

Analytics (ICAMIDA 2022) (pp. 38-46). Atlantis Press.

Birkenstein, C., & Graff, G. (2018). *They say/I say: The moves that matter in academic writing*. WW Norton & Company.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals*. Handbook I: Cognitive Domain. Longmans, Green.

Booth, W. C., Colomb, G. G., & Williams, J. M. (2009). *The craft of research*. University of Chicago press.

Borg, E. (2000). Citation practices in academic writing. In P. Thompson (Ed.), *Patterns and perspectives: Insights into EAP writing practice* (pp. 26-42). Reading, UK: Centre for Applied Language Studies.

Britton, J., Martin, N., Burgess, T., McLeod, A., & Rosen, H. (1975). *The Development of Writing Abilities (11–18)*. Macmillan.

Brown, N., & Janssen, R. (2017). Preventing plagiarism and fostering academic integrity: a practical approach. *Journal of Perspectives in Applied Academic Practice*, *5*(3), 102-109.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Bui, L. A. P., & Hsieh, I. H. (2024). Creative Writing Instruction For Primary Students: An In-Depth Analysis In The Context Of Curriculum Reform In Vietnam. *International Online Journal of Primary Education*, *13*(2), 109-121. https://doi.org/10.55020/iojpe.1467861

Buzan, T. (2017). *The Power of Creative Intelligence: 10 ways to tap into your creative genius*. HarperCollins UK.

Carroll, J. (2007). A Handbook for Deterring Plagiarism in Higher Education. Oxford Centre for Staff and Learning Development.

Chomsky, N., Roberts, I., & Watumull, J. (2023, March 8). Opinion | Noam Chomsky: The False Promise of ChatGPT. The New York Times. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd Edition ed.). Routledge.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*(1), 98.

Cotton, Debby R. E., Cotton, Peter A., & Shipway, J. Reuben. 2023. Chatting and Cheating: Ensuring Academic Integrity in The Era of ChatGPT Chatting and Cheating: Ensuring

Academic Integrity in The Era of ChatGPT. Innovations in Education and Teaching International. Doi: 10.1080/14703297.2023.2190148.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297-334.

Crossley, S.A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research, 11*(3), 415-443. https://doi.org/10.17239/jowr-2020.11.03.

Cumming, A., Kantor, R., Powers, D. E., Santos, T., & Taylor, C. (2002). Scoring TOEFL essays and speaking responses: The influence of task and rater characteristics. *ETS Research Report Series, 2002(1)*, i-36.

Cusen, G. (2018). "Borders" in the writing of academic texts: Investigating informativeness in academic journal abstracts. *Acta Universitatis Sapientiae Philologica, 10* (2), 141- 154.

Elbow, P. (1998). *Writing Without Teachers*. Oxford University Press.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, 24(2), 143-188.

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron & R. J. Sternberg (Eds.), *Teaching Thinking Skills: Theory and Practice*. W.H. Freeman.

Fitriansyah, N., & Miftah, M. Z. (2020). Positive connection of extensive reading and writing fluency in EFL learning. *LET: Linguistics, Literature and English Teaching Journal*, 10(2), 44. https://doi.org/10.18592/let.v10i2.4137

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365-387.

Gabi, C. (2022). *Academic Writing*. Altralogue.

Gardner, H. E. (2011). *Frames of mind: The theory of multiple intelligences*. Basic books.

Gjesdal, A. M. (2013). The influence of genre constraints on author representation in medical research articles. The French indefinite pronoun on in IMRAD research articles. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (12).

Goldberg, A. E. (2006). Constructions at work: The nature of generalization in language. *Oxford University Press*.

Guilford, J. P. (1967). Creativity and learning. *Brain function*, *4*, 307-326.

Guilford, J. P. (1967). *The Nature of Human Intelligence*. McGraw-Hill.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*: Pearson College Division.

Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. London: Longman

Hsu, T. W., Tsai, S. J., Ko, C. H., Thompson, T., Hsu, C. W., Yang, F. C., ... & Su, K. P. (2023). Plagiarism, quality, and correctness of ChatGPT-generated vs human-written abstract for research paper. *Available at SSRN 4429014*.

Hung, B.P. (2019). Meaningful learning and its implications for language education in Vietnam. *Journal of Language and Education* 5(1): 98–102. doi: 10.17323/2411-7390-2019-5-1-98-102.

Imran, A. A., & Lashari, A. A. (2023). Exploring the world of Artificial Intelligence: The perception of the university students about ChatGPT for academic purpose. *Global Social Sciences Review, VIII*.

Jamali, H. R., & Nikzad, M. (2011). Article title type and its relation with the number of downloads and citations. *Scientometrics, 88*(2), 653-661. https://doi.org/10.1007/s11192-011-0412-z

Jarrah, A. M., Wardat, Y., & Fidalgo, P. (2023). Using ChatGPT in academic writing is (not) a form of plagiarism: What does the literature say?. Online Journal of Communication and Media Technologies, 13(4), e202346, pp.1- 20. Doi:10.30935/ojcmt/13572.

Katar, O., Özkan, D., Yıldırım, Ö., & Acharya, U. R. (2023). Evaluation of GPT-3 AI language model in research paper writing. *Turkish Journal of Science and Technology*, *18*(2), 311-318.

Khalil, M., Er, E. (2023). Will ChatGPT Get You Caught? Rethinking of Plagiarism Detection. In: Zaphiris, P., Ioannou, A. (eds) Learning and Collaboration Technologies. HCII 2023. Lecture Notes in Computer Science, vol 14040. Springer, Cham. https://doi.org/10.1007/978-3-031-34411-4_32

Krashen, S. D. (1984). *Writing: Research, theory and applications*. Oxford: Pergamon Institute of English.

Labajová, L. (2023). The state of AI : Exploring the perceptions, credibility, and trustworthiness of the users towards AI-Generated Content (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:mau:diva-61215

Langacker, R. W. (2007). Cognitive grammar. *Cognitive Science*, 31(5), 733-755.

Le, P. T. N. (2023). The Effectiveness of and Students' Perceptions of Peer Feedback: A Vietnam Action Research Project. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, *15*(1).

Li, Y., & Zhang, L. (2020). *Enhancing Writing Skills with AI: Opportunities and Challenges*. Journal of Second Language Writing, 48, 100722.

Mahama, I., Baidoo-Anu, D., Eshun, P., Ayimbire, B., & Eggley, V. E. (2023). Chatgpt in academic writing: a threat to human creativity and academic integrity? An exploratory study. *Indonesian Journal of Innovation and Applied Sciences (IJIAS)*, *3*(3), 228-239.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written communication*, *27*(1), 57-86.

Meniado, J. C. (2023). The Impact of ChatGPT on English Language Teaching, Learning, and Assessment: A Rapid Review of Literature. *Arab World English Journal*, *14*(4).

MLA. (2021). *MLA Handbook* (9th ed.). Modern Language Association.

Mogey, N., Paterson, J., Burk, J., & Purcell, M. (2010). Typing compared with handwriting for essay examinations at university: letting the students choose. *ALT-J*, *18*(1), 29-47.

Nippold, M. A. (2000). Language development during the adolescent years: Aspects of pragmatics, syntax, and semantics. *Topics in Language Disorders, 20*(2), 15-28.

Nonaka, I., & Takeuchi, H. (1995). *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press.

Nunnally, J. C. (1967). *Psychometric Theory*. New York: McGraw-Hill.

Park, C. (2003). In Other (People's) Words: Plagiarism by university students--literature and lessons. *Assessment & Evaluation in Higher Education*, 28(5), 471-488.

Qualifications And Curriculum Authority (2000). English: the National Curriculum for England, Key Stages 1–4. London: QCA.

Reither, J. A. (1985). Writing and Knowing: Toward Redefining the Writing Process. *College English*, *47*(6), 620–628. https://doi.org/10.2307/377164

Ruegg, R., & Sugiyama, Y. (2010). Do analytic measures of content predict scores assigned for content in timed writing. *Melbourne Papers in Language Testing*, *15*(1), 70-91.

Safrai, M., & Orwig, K. E. (2024). Utilizing artificial intelligence in academic writing: an in-depth evaluation of a scientific review on fertility preservation written by ChatGPT-4. *Journal of Assisted Reproduction and Genetics*, 1-10.

Schmidt-Fajlik, R. (2023). ChatGPT as a grammar checker for Japanese English language learners: A comparison with Grammarly and ProWritingAid. *AsiaCALL Online Journal*, *14*(1), 105-119.

Scott, C. M. (2004). Syntactic contributions to literacy learning. In C. A. Stone, E. R. Silliman, B. J. Ehren, & K. Apel (Eds.), *Handbook of Language and Literacy: Development and Disorders* (pp. 340-362). Guilford Press.ae

Seow, A. (2002). The writing process and process writing. *Methodology in language teaching: An anthology of current practice*, *315*, 320.

Soares, B. B. C. (2022). Academic Writing: Practical Tips For Writing Scientific Works. *Revista Gênero e Interdisciplinaridade*, *3*(03), 228-235.

Sofia, T. N. (2023). Creative Intelligence And Creative Writing. *Educatia Plus*, *34*(2), 26-42.

Strunk, W., & White, E. B. (2000). *The Elements of Style*. Allyn & Bacon.

Struthers, L., Lapadat, J. C., & MacMillan, P. D. (2013). Assessing cohesion in children's writing: Development of a checklist. *Assessing Writing*, *18*(3), 187-201.

Stuart, M., & Barnett, A. (2023). The Writing Quality Scale: Development and Validation. *Journal of Educational Measurement*, 60(1), 45-67.

Sumner, E., & Connelly, V. (2020). Writing and revision strategies of students with and without dyslexia. *Journal of Learning Disabilities*, *53*(3), 189-198.

Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.

Swandi, I. S. B., & Netto-Shek, J. A. (2017). Teaching writing at the primary levels. *Indonesian Journal of Applied Linguistics*, *7*(1), 1-10.

Todd, R. W., Khongput, S., & Darasawang, P. (2007). Coherence, cohesion and comments on students' academic essays. *Assessing writing*, *12*(1), 10-25.

Traxler, M. J., & Gernsbacher, M. A. (1992). Improving written communication through minimal feedback. Quarterly *Journal of Experimental Psychology, 45*(1), 45-61. https://doi.org/10.1080/14640749208401302

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Yeadon, W., Agra, E., Inyang, O. O., Mackay, P., & Mizouri, A. (2024). Evaluating AI and Human Authorship Quality in Academic Writing through Physics Essays. *arXiv preprint arXiv:2403.05458*.

Zhang, Z. (2024). New Communicative Language Teaching Methods: How ChatGPT is Used in English Teaching and Its Impacts. *Journal of Education, Humanities and Social Sciences*, *32*, 74-78.

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.

Appendices

Appendix A: Writing Quality Assessment Scale

Criteria	Description	0 point	1 point	2 points	3 points
Grammar	The grammatical correctness and adherence to proper syntax.	Grammatical errors are frequent and/or severely interfere with the syntax.	Grammatical errors are occasional and/or sometimes interfere with the syntax.	Grammatical errors are few.	No errors in grammar and syntax.
Plagiarism Avoidance	Similarity with other sources' words, ideas, or work without proper acknowledgment or citation presents them as one's own original creation.	The similarity of word use compared to other sources is above 60%.	The similarity of word use compared to other sources is above 40% to 60%.	The similarity of word use compared to other sources is above 10% to 40%.	The similarity of word use compared to other sources is up to 10%. [1] [2]
Reference	The accuracy in the names of authorship, titles, and journals/publications.	0 out of 3 criteria (names of authorship, titles, and journals/publications) is correct.	1 out of 3 criteria (names of authorship, titles, and journals/publications) is correct.	2 out of 3 criteria (names of authorship, titles, and journals/publications) are correct.	3 criteria (names of authorship, titles, and journals/publications) are correct.
Content Development	The connection between developing ideas and supporting ideas in a paragraph.	State an unclear argument with no supportive idea and no example. (1-0-0)	State a clear argument with one supportive idea or one example. (1-1-0)	State a clear argument with more than one supportive idea or more than one example. (1-2-0)	State a clear argument with at least one supportive idea and at least one example. (1-1-1)
Coherence and Cohesion	The flow of the text and how well ideas, sentences, and paragraphs are connected logically.	No cohesive device is used to connect ideas logically.	Cohesive devices are repetitively or scarcely used.	Cohesive devices are effectively but inadequately used.	Cohesive devices are effectively used
Idea Originality	Personal opinion with appropriate reasoning to answer the question of the topic.	No personal opinion to answer the question of the topic.	Personal opinions are stated without relevance to the topic.	Personal opinions are stated but unclear or in dilemmas.	Personal opinions are strongly stated with appropriate evidence.

[1] Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulations, 2018 (lasted accessed 27th March 20202). Available on https://www.ugc.ac.in/pdfnews/7771545_academic-integrity-Regulation2018.pdf.

[2] Kadam, D. (2018). Academic integrity and plagiarism: The new regulations in India. *Indian Journal of Plastic Surgery*, *51*(02), 109-110.