ISSN 2959-6122

A Comparative Study on ERNIE Bot 4.0 Turbo and ChatGPT 4o's Performance in Evaluating First-Year Undergraduate Persuasive Essays

Xinran Yue[1, *]

[1]Department of Translation and Linguistics, City University of Hong Kong, Hong Kong, 999077, China
*Corresponding author: xinranyue3-c@my.cityu.edu.hk

Abstract:

This study compares the performance of two prominent AI language models, ERNIE Bot 4.0 Turbo and ChatGPT 4o, in evaluating first-year undergraduate persuasive essays within the social sciences domain. Drawing from the Louvain Corpus of Native English Essays, a comprehensive collection of academic writings by British and American university students, this study aims to examine the models' capabilities in assessing the grammatical correctness, vocabulary usage, coherence, content depth, and writing style of the essays. This study adopts a structured evaluation framework based on IELTS writing criteria to assess the models' performance. A 40 persuasive essays from the Louvain Corpus were evaluated by both AI models and compared with human raters' evaluations to ensure validity. The findings reveal distinct differences in the assessment styles of the two models. ChatGPT 4o exhibits a more critical approach, pinpointing areas for improvement, such as lack of argument development, coherence issues, and grammatical errors. Conversely, ERNIE Bot 4.0 Turbo offers a more balanced assessment, acknowledging essays' strengths and suggesting improvement areas. Notably, ERNIE Bot's evaluation highlights potential biases in AI-based assessment systems, particularly in its unequal emphasis on viewpoints. This comparative examination offers valuable perspectives on the advantages and constraints of AI models in assessing scholarly compositions, underscoring the significance of amalgamating varied AI functionalities to establish more all-inclusive and efficient feedback systems for learners. By understanding these differences, researchers and educators can better utilize AI-assisted essay evaluation systems to enhance student learning experiences.

Keywords: AI-assisted essay evaluation; ERNIE Bot 4.0 Turbo; ChatGPT 4o; Corpus Analysis.

1. Introduction

In the past few years, there has been a notable surge in the incorporation of artificial intelligence (AI) across different aspects of education. Among its diverse applications, AI-assisted essay evaluation emerges as a prominent exemplification. These systems expedite the grading process and provide students with personalized feedback on their writing, enhancing their learning experience [1]. However, variations in the performance of AI models during the training process may arise across different countries [2].

The Louvain Corpus of Native English Essays established by the Centre for English Corpus Linguistics (CECL), Université Catholique de Louvain, Belgium, which encompasses a vast range of academic writings by British and American university students, serves as an invaluable resource for examining patterns in student writing and assessing proficiency levels. By selecting 40 representative samples from this corpus, this study aims to conduct an analysis of two prominent AI models: ERNIE Bot 4.0 Turbo and ChatGPT 4o. These models have been trained in distinct countries with diverse Internet backgrounds, which gives them a distinctive ability to offer extensive understandings of the intricacies involved in language processing and assessment. Specifically, first-year undergraduate persuasive essays within the social sciences domain are selected.

The novelty of this study lies in its comparative analysis of two distinct AI models, ERNIE Bot 4.0 Turbo and ChatGPT 4o, in evaluating English essays written by university students. While the use of AI for essay evaluation has become increasingly prevalent, only a few studies have explored the performance of models trained in cross-lingual settings. By leveraging the Louvain Corpus of Native English Essays, the objective of this research is to illuminate the similarities and differences in the assess-

ment capacities of ERNIE Bot 4.0 Turbo and ChatGPT 4o, particularly concerning assessing grammar, syntax, and vocabulary usage in English essays. The unique focus on the intersection between AI-assisted essay evaluation and cross-lingual language processing provides a novel perspective that has significant potential to advance our understanding of both limitations and strengths inherent within current AI models used in educational settings.

In summary, this study provides a comparative analysis of two prominent AI models in AI-assisted essay evaluation, revealing valuable perspectives on their performance and suitability across various evaluation criteria. By elucidating the strengths and limitations inherent in these models, this study aims to inform the development of more efficacious and tailored AI-assisted essay evaluation systems that can better cater to the requirements of educators and students alike.

2. Literature Review

A notable surge in interest has emerged surrounding artificial intelligence-powered writing tools, primarily attributed to their efficacy in aiding student writing endeavors, offering grammatical and stylistic guidance, and fostering creative content development. Research conducted by Gayed et al. underscores the instrumental role these tools play in enhancing students' writing skills and bolstering their confidence levels [3]. Artificial intelligence (AI) tools present immense value to learners, particularly in offering prompt feedback and enhancing writing proficiency. Nonetheless, Makarius et al. underscore the necessity for further advancements in these tools to deepen their comprehension and efficacy across diverse academic disciplines [4]. The integration of AI technologies within the realm of academic paper writing has sparked a discourse surrounding the responsibilities and ethical considerations of educators. A study conducted by Su et al. delves into the implications of AI on educational practices, ultimately affirming that while AI tools offer valuable feedback, the guidance of teachers remains paramount in nurturing critical thinking skills and creativity [5]. Chaudhry et al. delve into the ethical implications of artificial intelligence in plagiarism detection, emphasizing the paramount importance of well-defined protocols and extensive student education in acknowledging the limitations and proper application of AI technologies [6].

In light of past recognitions regarding the challenges and opportunities ahead, Theodosiou and Read contribute to the field by introducing methodologies designed to improve transparency and comprehension in the realm of AI-generated writings, thereby addressing a persistent issue [7]. Expanding the horizon, Mazzone, Elgammal, Dwivedi, and colleagues have investigated how AI can empower creative paper writing, offering fresh perspectives on the transformative potential of AI that extends beyond mere facilitation [8,9]. In conclusion, studies on the integration of AI in academic writing reveal its significant transformative effects on education. The availability of AI-based writing tools presents a valuable resource for students and educators, yet further investigation and development are necessary to address challenges associated with contextual awareness, bias elimination, and ethical utilization. By fostering a harmonious blend of AI-driven automation and human-guided intervention, this research endeavor endeavors to tap into AI's full potential, enriching educational experiences and fostering academic excellence.

The following research questions guided the present study:

1. What are the similarities and differences in the performance of these two AI models when evaluating first-year undergraduate persuasive essays within the social sciences domain?
2. How do these performance differences affect teachers or students utilizing AI models for essay evaluation?

3. Research Design

The current research design aims to conduct a comparative study on the performance of ERNIE Bot 4.0 Turbo and ChatGPT 4o in evaluating first-year undergraduate persuasive essays within the social sciences domain. This study builds upon the recent integration of AI into various aspects of education, particularly in AI-assisted essay evaluation.

3.1 Data Collection

To conduct a comparative study on the performance of ERNIE Bot 4.0 Turbo and ChatGPT 4o in evaluating university students' essays, 40 representative samples of essays from the Louvain Corpus of Native English Essays are selected. This corpus comprises a diverse array of academic writings by British and American university students, offering a rich resource for studying student writing patterns and proficiency levels. Specifically, first-year undergraduate persuasive essays within the social sciences domain are selected. The strategic selection of the samples for this comparative study balances specificity with practicality. This focus not only adheres to a standardized structure and purpose common in such curricula, making them an ideal foundation for evaluating AI models' performance, but it also ensures a consistent level of academic maturity across the sample. Furthermore, the essays' origin from a corpus representative of diverse yet comparable writings from two major English-speaking countries

offers insights into how AI models trained under different cultural and linguistic contexts interpret and evaluate persuasive arguments. Ultimately, this analysis of first-year essays provides valuable insights into how AI can support students at the outset of their academic journeys. In these pinpointing areas, personalized feedback may hold the greatest potential for enhancing writing development.

3.2 AI Models Selection

The selection of ERNIE Bot 4.0 Turbo and ChatGPT 4o as the AI models for this comparative study stems from their distinct training corpora and their respective state-of-the-art language processing and evaluation capabilities. ERNIE Bot 4.0 Turbo, a model that has potentially been trained with a heavier emphasis on a Chinese Internet background, is anticipated to demonstrate exceptional proficiency in capturing the intricate nuances of language structure, syntax, and grammar specific to the Chinese linguistic system. This unique training paradigm equips ERNIE Bot 4.0 Turbo with a nuanced understanding of the complexities inherent in Chinese writing, making it a valuable tool for exploring how AI can interpret and evaluate texts within a culturally and linguistically distinct context.

In contrast, ChatGPT 4o represents a more general-purpose language model trained on an extensive and diverse array of data, spanning multiple languages, including English and Chinese. This broad exposure to various linguistic patterns and styles endows ChatGPT 4o with a comprehensive understanding of language across cultures and domains. As a result, ChatGPT 4o is expected to demonstrate versatility and robustness in evaluating persuasive essays, drawing upon its extensive knowledge base to provide insightful feedback that transcends specific linguistic or cultural boundaries.

The juxtaposition of these two models, each with its distinct training background and capabilities, offers a unique opportunity to delve into the intricacies of language processing and evaluation from multiple perspectives.

3.3 Evaluation Framework

The performance of these two models was assessed using the structured writing assessment framework developed by the IELTS test [10,11]. This framework encompassed several metrics:

Firstly, the models evaluated the essays' grammatical correctness, identifying syntactic errors or inconsistencies. This assessment helped identify areas where the writing could have been refined for clarity and accuracy.

Secondly, the vocabulary usage in the essays was scrutinized. The AI models assessed word choices' range, appropriateness, and correctness,, providing feedback on how to enrich the language and make it more precise.

Thirdly, the coherence and structure of the essays were evaluated. The models checked for logical flow, paragraph organization, and the overall structure of the argument, suggesting improvements where necessary.

Fourthly, the models assessed the essays' content depth and originality. They evaluated the uniqueness of ideas, the strength of arguments, and the overall depth of the content, offering suggestions to enhance the essays' intellectual value.

Lastly, the AI models provided general feedback on writing style, tone, and voice. This evaluation focused on how the author expressed their ideas and whether the writing style was appropriate for the academic context..

3.4 Experimentation

To determine which AI tool had better performance in the evaluation of student essays, a multi-step process was followed:

Firstly, preprocessing of the selected essays was conducted to remove any personal identifiers, thus ensuring anonymity and compliance with privacy regulations. This step was crucial for protecting students' privacy whose essays were being evaluated.

Secondly, ERNIE Bot 4.0 Turbo and ChatGPT 4o were used to evaluate the preprocessed essays. These AI models provided scores and feedback on various aspects of the essays, including grammar, syntax, vocabulary usage, and overall quality. The AI evaluations served as the primary data for comparing the performance of the two models.

Thirdly, to ensure the validity of the AI evaluations, five experienced English teachers were assembled to independently evaluate the same set of essays. The teachers were blinded to the AI evaluations and assessed the essays based on the same criteria used by the AI models. The human evaluations provided a benchmark for comparing the performance of the AI tools.

Lastly, a comparison and analysis of the AI evaluations and human evaluations was conducted. The similarities and differences in the scores and feedback provided by the AI models and human evaluators were analyzed to assess the performance of each AI tool. This analysis revealed how closely the AI evaluations aligned with human judgments and identified areas for improvement in the AI models.

4. Result and Discussion

This study evaluated the performance of two AI tools, ERNIE Bot 4.0 Turbo and ChatGPT 4o, in assessing student essays. The evaluation process comprised several steps, including preprocessing the essays to ensure anonymity and privacy compliance, followed by AI evaluations, hu-

man evaluations, and a final comparison and analysis.

4.1 Result

The results of our analysis indicate that ChatGPT 4o and ERNIE Bot 4.0 Turbo employ distinct approaches in evaluating the writing samples. ChatGPT 4o, adopting a more critical perspective, identified multiple areas for improvement in the sample. For instance, in Sample 1, ChatGPT 4o emphasized the necessity of further developing the argument with specific examples and detailed explanations. This finding aligns with the lower overall score (5) assigned by ChatGPT 4o, reflecting its evaluation criteria. Additionally, issues related to coherence and cohesion were highlighted by ChatGPT 4o, including awkward phrasing, repetitive sentences, and disrupted flow due to inadequate transitions. The ChatGPT 4olso observed errors in lexical resource usage such as word choice, collocation errors, and the use of awkward or unclear phrases which impacted readability. Finally, grammatical errors like punctuation and sentence fragments that hindered clarity and accuracy were pointed out by ChatGPT 4o.

In contrast, ERNIE Bot 4.0 Turbo provided a more comprehensive evaluation, encompassing the sample's strengths and weaknesses. With an overall estimated score of 6.5, ERNIE Bot 4.0 Turbo commended the sample's strong coherence and cohesion, which is evident in its lucid introduction, well-developed body paragraphs, and succinct conclusion. The ChatGPT 4o also acknowledged a satisfactory range of vocabulary usage, incorporating some less common words and phrases, and demonstrating a diverse array of grammatical structures through appropriately applied complex sentence types. While ERNIE Bot 4.0 Turbo identified opportunities for greater lexical resource variation and precision, it did not critique the lexical resource as harshly as ChatGPT 4o. Similarly, ERNIE Bot 4.0 Turbo observed only minor significant grammatical errors that contributed to its higher overall score.

A notable disparity in the evaluations lies in the emphasis on the Task Response criterion. While both models acknowledged the sample's response to the task by addressing multiple perspectives, ERNIE Bot 4.0 Turbo exhibited an uneven emphasis, focusing more the opposing viewpoint. In contrast, ChatGPT 4o did not explicitly comment on the equilibrium of viewpoints in its evaluation.

The teachers' evaluations revealed a nuanced perspective, with certain essays garnering commendation for their coherent structure, extensive vocabulary, and precise grammar usage. Interestingly, the teachers placed significant emphasis on writers' ability to effectively engage with the task by providing well-supported arguments and addressing multiple perspectives. This focus closely aligned with ERNIE Bot 4.0 Turbo's evaluation which also acknowledged the sample's satisfactory response to the task. However, unlike AI models, teachers displayed greater leniency towards minor grammatical errors and lexical inaccuracies while prioritizing essay flow and coherence as key factors for evaluation. Furthermore, they were able to identify instances where writer creativity and originality enhanced the impact of the essay; dimensions that were not explicitly addressed in AI evaluations.

A comprehensive comparison and analysis of the AI evaluations and human evaluations revealed both similarities and differences in their approaches and findings. In terms of similarities, both ChatGPT 4o and ERNIE Bot 4.0 Turbo identified coherence and lexical resource as critical aspects of writing quality, aligning with the human evaluators' emphasis on these areas. Furthermore, all evaluators concurred that the samples exhibited a range of strengths and weaknesses, necessitating specific attention for improvement. However, several notable distinctions emerged in the evaluations. The AI models, particularly ChatGPT 4o, displayed a more discerning approach by highlighting grammatical errors and lexical inaccuracies that were often overlooked by human evaluators. This disparity can be attributed to the AI's adherence to stringent evaluation criteria while lacking inherent contextual understanding possessed by humans. Additionally, human evaluators placed greater significance on creativity, originality, and the writer's ability to meaningfully engage with the task at hand. Although acknowledging task response in the samples evaluated, AI models did not scrutinize this aspect as comprehensively as human evaluators did. Finally, scores assigned by human evaluators tended to be more lenient compared to those given by AI models due to their capacity to consider broader context and purpose of essays.

In conclusion, while both AI evaluations and human evaluations offer valuable insights into the quality of writing, they differ in their methodologies and emphases. A combination of these evaluation methods could yield a more comprehensive and nuanced understanding of writing proficiency by harnessing the strengths inherent in both human and machine evaluation.

4.2 Discussion

The contrasting evaluations provided by ChatGPT 4o and ERNIE Bot 4.0 Turbo underscore the importance of considering diverse perspectives in AI-based educational assessment. ChatGPT 4o's critical approach, focusing on identifying specific areas for improvement, benefit learners seeking detailed feedback to enhance their writing skills. By pinpointing issues such as a lack of argument development, coherence, lexical precision, and grammatical errors, ChatGPT 4o offers a targeted roadmap for improvement.

However, ERNIE Bot 4.0 Turbo's more balanced assessment, which highlights strengths and weaknesses, may be more encouraging for learners, particularly those who have demonstrated some proficiency in the writing task. By praising the sample's coherence, lexical range, and grammatical structures, ERNIE Bot 4.0 Turbo reinforces positive aspects of the writing and motivates learners to continue building on their strengths.

The unequal emphasis on viewpoints ERNIE Bot 4.0 Turbo observes raises questions about the potential biases inherent in AI-based assessment systems. While the focus on the opposing viewpoint may not necessarily indicate a flaw in the sample's response to the task, it highlights the need for careful calibration of AI models to ensure that they evaluate writing samples in an unbiased and comprehensive manner.

Furthermore, our results suggest that ChatGPT 4o's critical analysis and ERNIE Bot 4.0 Turbo's balanced assessment could provide a more comprehensive and effective feedback loop for learners. By leveraging the strengths of both models, this study aimes to support educators and learners by providing valuable insights into their writing abilities, while identifying particular aspects that could benefit from improvement.

In conclusion, this study highlights the diverse approaches ChatGPT 4o and ERNIE Bot 4.0 Turbo adopted in evaluating writing samples according to IELTS Writing criteria. The contrasting evaluations offer valuable insights into AI-based educational assessment systems' potential benefits and limitations. Future research should explore ways to integrate the strengths of different AI models to create more comprehensive and effective assessment tools for learners.

5. Conclusion

This study offers a detailed comprehension of the performance of two leading AI language models, ERNIE Bot 4.0 Turbo and ChatGPT 4o, in evaluating first-year undergraduate persuasive essays within the social sciences domain. The findings demonstrate that while both models exhibit valuable assessment capabilities, they differ significantly in their approaches and focus. ChatGPT 4o adopts a more critical stance, effectively identifying flaws in essays, whereas ERNIE Bot 4.0 Turbo offers a more balanced assessment, recognizing both strengths and weaknesses.

Crucially, the study highlights the potential biases inherent in AI-based assessment systems, particularly the unequal emphasis on viewpoints observed in ERNIE Bot's evaluations. This underscores the need for continued research and development to mitigate such biases and ensure AI models provide fair and unbiased feedback.

The analysis contributes to the expanding body of literature on AI-assisted essay evaluation, highlighting the significance of integrating diverse AI capabilities to establish comprehensive and efficacious feedback mechanisms. By acknowledging the strengths and limitations of various AI models, researchers and educators effectively utilize the potential of AI in order to improve the learning experiences of students. As AI continues to permeate educational settings, this research provides a point of reference for educators seeking to incorporate AI-assisted essay evaluation systems into their teaching practices.

References

[1] Kaledio, P., Robert, A., & Frank, L. (2024). The impact of artificial intelligence on students' learning experience. Social Science Research Network. https://doi.org/10.2139/ssrn.4716747

[2] Zhu, L., Mou, W., Lai, Y., Lin, J., & Luo, P. (2024). Language and cultural bias in AI: comparing the performance of large language models developed in different countries on Traditional Chinese Medicine highlights the need for localized models. Journal of Translational Medicine, 22(1). https://doi.org/10.1186/s12967-024-05128-4

[3] Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing Assistant's impact on English language learners. Computers and Education. Artificial Intelligence, 3, 100055. https://doi.org/10.1016/j.caeai.2022.100055

[4] Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. Journal of Business Research, 120, 262–273. https://doi.org/10.1016/j.jbusres.2020.07.045

[5] Su, J., Zhong, Y., & Ng, D. T. K. (2022). A meta-review of literature on educational approaches for teaching AI at the K-12 levels in the Asia-Pacific region. Computers and Education. Artificial Intelligence, 3, 100065. https://doi.org/10.1016/j.caeai.2022.100065

[6] Chaudhry, I. S., Sarwary, S. A. M., Refae, G. A. E., & Chabchoub, H. (2023). Time to revisit existing student's performance evaluation approach in higher education sector in a new era of CHATGPT — a case study. Cogent Education, 10(1). https://doi.org/10.1080/2331186x.2023.2210461

[7] Theodosiou, A. A., & Read, R. C. (2023). Artificial intelligence, machine learning and deep learning: Potential resources for the infection clinician. Journal of Infection, 87(4), 287–294. https://doi.org/10.1016/j.jinf.2023.07.006

[8] Mazzone, M., & Elgammal, A. (2019). Art, creativity, and the potential of artificial intelligence. Arts, 8(1), 26. https://doi.org/10.3390/arts8010026

[9] Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj,

A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

[10] BRITISH COUNCIL. (2023). IELTS Writing Task 1 Guidelines. Retrieved from https://www.chinaielts.org/pdf/UOBDs_WritingT1.pdf

[11] BRITISH COUNCIL. (2023). IELTS Writing Task 2 Guidelines. Retrieved from https://www.chinaielts.org/pdf/UOBDs_WritingT2.pdf