ChatGPT vs Bard: Which is a Better Writer?

Ai Leng Ng
School of Humanities and Behavioural Sciences
Singapore University of Social Sciences
Singapore
Alng006@suss.edu.sg

Justina Ong
School of Humanities and Behavioural Sciences
Singapore University of Social Sciences
Singapore
JustinaOng001@suss.edu.sg

***Abstract*— The emergence and subsequent widespread adoption of Artificial Intelligence (AI) Language models such as ChatGPT and Google Bard has sparked intense interest in the capabilities and applications of AI language models in education and language learning. However, not much is known about how AI language models perform linguistically. This study examined the effects of AI language models, ChatGPT and Bard, on the writing quality and linguistic features (i.e., lexical sophistication, syntactic complexity, and cohesion) of AI written output. Data was collected from both AI language models, which were tasked to answer a total of 120 essay prompts on topics ranging from Arts and Humanities, Social Sciences, and Science. The responses generated from both AI language models were analyzed for their writing quality, lexical sophistication, syntactic complexity, and cohesion. Results revealed that ChatGPT is stronger in producing higher quality text which uses more sophisticated words and complex sentences, whereas Bard is stronger in creating more cohesive texts. Results of the study are discussed in the context of AI language models' training methods, and implications for using AI in education, as well as, language teaching and learning.**

Keywords— AI language models, writing quality, lexical sophistication, syntactic complexity, cohesion

I. Introduction

Since the release of ChatGPT in November 2022, Artificial Intelligence (AI) language models have continued to gain attention, particularly in the global education community. As educational institutions move towards integrating generative AI tools such as ChatGPT and Google Bard into classrooms, the need to understand the ability of these tools by studying the output they produce becomes essential for educators. The use of AI language models in the educational context appears to be a significant topic of inquiry likely due to their ability to answer almost any question and create almost any type of text. Not only does the output produced appear almost indistinguishable from human writing [1], it has been often considered coherent, convincing and of high quality.

Analyzing the linguistic properties of AI written output enables users to identify linguistic features that AI language models use to craft coherent and convincing responses. Additionally, understanding the strengths and weaknesses of AI language models, not only improve users' AI literacy but also provide insights on what can be harnessed from their potential applications in education. This may assist teachers in utilizing AI language models as text generation tools in providing students with targeted, relevant, and relatable discourse text, in support of their language learning.

This paper will examine how the different AI language models influence the linguistic features and writing quality of AI written output. Specifically, this paper aims to analyze the patterns of language use in AI written output for their (i) lexical sophistication, (ii) syntactic complexity, (iii) cohesion, and (iv) writing quality in order to answer the following research question: What are the effects of AI language models (ChatGPT vs. Bard) on the lexical sophistication, syntactic complexity, cohesion, and writing quality of AI written output?

II. Literature Review

A. Background of AI Language Models

AI language models are machine learning systems that use neural networks to generate texts in response to prompts [2]. These systems are trained on prediction tasks, which require them to predict the likelihood of a word given its surrounding context. Through this training, these models have demonstrated an ability to produce language output that closely resembles human writing. These models specialize in processing and generating text by leveraging on massive datasets to identify grammar, syntax, semantics, and context.

ChatGPT is an AI language model developed by OpenAI that has been trained on an extensive amount of data and optimized for producing human-like text. It is based on generative pre-trained transformer (GPT) 3.5 architecture, a left-to-right architecture where it predicts the most likely continuation of words through pattern recognition and statistical likelihood. Its unidirectional training means that ChatGPT predicts the current word in a sentence by considering the words to the left of the current word. As such, ChatGPT only considers words preceding the current word being generated and has no knowledge of any words to the right of the current position word.

Bard is an AI language model developed by Google AI that uses bidirectional encoder representations from Transformers (BERT) as part of its architecture. Bard was trained using masked language modelling (MLM) where it was given sentences with words masked out and tasked with predicting the missing words. Bard is a bidirectional model which predicts words by considering words from left-to-right and right-to-left. Additionally, Bard has access to the internet. Thus, Bard responds to prompts by breaking down the query into parts of speech, and then accessing its database, including searching the web for information, to find the most relevant information for its response.

ChatGPT and Bard are currently two of the most advanced and popular AI language models in use, each with their distinct strengths and weaknesses owing to differences in their underlying models' architecture, training data and training approaches.

Despite the wide range of applications of ChatGPT and Bard, they are not without their limitations. AI language models may have the ability to generate human-like text through extensive training on vast corpora of text, enabling

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00066

them to capture patterns of syntax, semantics, and stylistic nuances. However, this linguistic ability comes with factual inaccuracy as users have been known to encounter challenges in discerning the veracity of AI-generated information. Even when the presented facts are incorrect, the use of language have lead readers to believe that the information is credible and well-supported. This tendency for misrepresentation has potentially far-reaching consequences in domains including education, journalism, and public discourse.

B. *Linguistic Features of Writing Quality*

Research on writing quality has utilized linguistic features to explore the characteristics of successful writing across various genres. Lexical sophistication, syntactic complexity, and cohesion have been shown to correlate with writing quality in several studies [3], [4], [5].

Lexical sophistication is a multi-dimensional construct which characterizes how uncommon or advanced words are [6]. It is commonly measured using distributional properties such as word frequency and word range. Syntactic complexity is another multi-dimensional construct that measures the degree of variation and elaboration of grammatical structures [7]. Syntactic complexity represents the range and sophistication of a written text [4]. Thus, by analyzing the number and variety of clauses, subordinating conjunctions, and other grammatical features used in a text, the complexity of the structures used to construct sentences can be revealed [8]. Mean length of sentence, mean length of clause and mean length of T-unit have been used to capture the degree of syntactic complexity in writing in several studies [4], [9].

Text cohesion refers to linguistic devices that create links within a text [10]. It can occur locally at sentence level, globally across paragraphs, or even across texts [5]. Cohesion is a vital element in aiding readers in distinguishing unified whole texts from sentences which are unrelated but adjacent to one another [11]. If cohesion in a text is absent, readers will face difficulties relating the lexical items within the text, thereby affecting their comprehension of the text.

In the context of AI language model output, the use of lexical sophistication, syntactic complexity, and cohesive devices influences readers' perceptions of how persuasive the texts are, and correspondingly, affects the acceptance of the information as accurate and true, even when information presented may be factually inaccurate or biased. This underscores the importance of analyzing the linguistic features used by AI language models. Apart from studying the features to aid in teaching writing, understanding how AI language models use linguistic devices to create persuasive texts has implications for the responsible use of AI-generated content in education and academia.

C. *Previous Research on AI Language Models in Education, Assesment and Linguistics*

The development of ChatGPT and Bard has inspired a surge of research into their applications for language teaching and assessment. Studies have been conducted to test AI language models' ability to perform academic tasks such as sitting for an exam at undergraduate [12] and postgraduate level [13]. They did so by including exam questions as prompts, and grading the output based on existing marking schemes. ChatGPT was found to be able to provide answers with well elaborated explanations for basic questions from a Wharton postgraduate course, however, advanced process analysis questions were not answered well [13]. The findings highlight ChatGPT's ability to provide elaborated and well-written answers for only some types of questions.

ChatGPT's highly fluent and semantically coherent outputs shared such a close resemblance to human-authored texts, that even experts faced challenges distinguishing them apart. Casal and Kessler [14] investigated the extent to which linguists and reviewers from top Applied linguistics journals could distinguish AI-authored research abstracts from human-authored ones. Reviewers used continuity, coherence, and writing quality as criteria to distinguish AI-authored texts from human-authored ones. Results showed that linguistics experts had difficulty distinguishing AI-authored texts from human-authored texts, with a success rate of identifying human-authored texts at 44% and AI-authored texts at 34%.

Linguistic features of writing have also been studied alongside the use of AI, in the areas of Automatic essay scoring (AES). Mizumoto and Eguchi [15] investigated ChatGPT's reliability and accuracy in AES and found that the use of linguistic features enhanced the accuracy of ChatGPT's scoring. Using ChatGPT to score 12,100 essays, they compared ChatGPT scores against benchmark levels and explored the extent to which linguistic features influence AES with ChatGPT. By comparing the scores against 45 linguistic features which capture lexical sophistication, syntactic complexity, and cohesion, the authors found that including linguistic features in their regression model significantly improved the prediction of benchmark levels. This suggests that a consideration of the linguistic features in the text may produce more reliable and accurate prediction of essay ratings.

ChatGPT was also used in writing classrooms to assist students in developing their argumentative essays. ChatGPT has been used to support students in the various stages of essay writing [16]. In the study, ChatGPT was used as a writing evaluator that provided feedback to students in order to scaffold the structural and linguistic aspects of essay development. ChatGPT was instructed to adopt the role of writing tutor and was tasked to provide feedback on the outline, content or language used based on an evaluation rubric provided by the authors. The feedback provided by ChatGPT was subsequently assessed. Overall, the authors found that ChatGPT could provide feedback on different aspects of student writing, but not all feedback was useful for learning. ChatGPT was able to provide feedback on the essay outline, evaluate the quality of the supporting evidence and provide suggestions for changes in language use. However, the feedback and suggestions were found to be too general, overlapping, and superficial. Additionally, the suggested edits were found to be mostly evaluative, lacked elaboration and limited the learning potential for students.

While the applications of AI language models have been explored and assessed in various aspects of teaching, few studies have considered the linguistic features of the output produced by these language models. One such study compared the differences between ChatGPT output and human writing by conducting a qualitative analysis of human writing available online and ChatGPT's responses to the same questions [17]. Results showed that ChatGPT wrote in an organized manner with clear logic. Subsequent linguistic analysis showed that, compared to humans, ChatGPT used more conjunctions, creating text that appeared more logical. Additionally, there was a frequent co-occurrence of conjunctions with nouns, verbs, and prepositions, which created more informative and objective text, and reflected a clear structure within the writing. Another study compared the novelty of syntactic structure of GPT-2, a predecessor to ChatGPT, to human writing in order to determine whether language models copy from their training data [18]. AI-generated text was compared with human-generated text and evaluated on how novel the syntactic structure was. The objective was to determine whether AI could apply syntactic rules to its output, or if it was merely copying from its training data. The authors found that GPT-2 could indeed create novel sentence structure. Furthermore, most of the generated sentences had overall syntactic structures that were even more novel than human-generated text. However, humans rely on meaning and intent to guide their writing whereas AI language models have no clear goal when generating text. The authors thus noted that novelty is not necessarily a desired trait, as a model can increase novelty by generating random words, but its output may become incoherent.

While several studies [16], [17], [18], have been focused on the successes of ChatGPT's applications, this is premature without a careful assessment of the texts it generates. The current open-ended experimental approach to generative AI research highlights a lack of a robust understanding of generative AI's capabilities [19], thus more focus needs to be placed on understanding the linguistic features of AI written output. With a paucity of studies that examine how AI language models perform linguistically, this study contributes to filling this gap by focusing on the linguistic constructs in AI written output.

III. Methodology

A. Context of the Study

The main aim of this study is to examine the effects of AI language models on the writing quality, lexical sophistication, syntactic complexity, and cohesion of AI written output. To examine linguistic features produced by AI language models, a few Natural Language Processing (NLP) tools such as (i) the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) [20], (ii) the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) [21], (iii) the Tool for the Automatic Analysis for Cohesion (TAACO) [22], and (iv) Grammarly have been employed in this study. Such tools have been used in several studies to measure the lexical sophistication [5], [23], [24], syntactic complexity [3], [15], [25], and cohesion [15], [23], [26] in texts.

B. Procedures of the Study

First, a total of 120 task prompts were designed and created using Bloom's taxonomy framework [27]. The framework was employed to design essay questions based on a range of topics from the domains of Arts and Humanities, Science, and Social Sciences. Second, to collect the data, each prompt was input into the web interfaces of ChatGPT (https://chat.openai.com) and Google Bard (https://bard.google.com/chat), where each model generated a unique response to the prompt. Altogether, the dataset comprised of a total of 240 responses, with 120 texts generated by ChatGPT and 120 texts generated by Bard.

Next, to measure lexical sophistication, syntactic complexity, and cohesion, NLP tools, namely, TAALES, TAASSC, TAACO were used. *Grammarly* (https://app.grammarly.com/), an AES system, was used to score the texts for writing quality. The results collected from *Grammarly* and the NLP tools formed the data used for statistical analysis. Analysis of variance (ANOVA) was performed to examine the effects of AI language models on lexical sophistication, syntactic complexity, cohesion, and writing quality of AI written output.

C. Data Collection Tools

This study used Grammarly, TAALES, TAASSC, and TAACO to measure writing quality, lexical sophistication, syntactic complexity, and cohesion respectively. Table 1 lists the selected linguistic features indices and the respective NLP tools used for measurement.

Grammarly, a free AES system, was used to rate all 240 responses. AES was adopted because it has been known to mitigate risks associated with human raters, such as inconsistent or inaccurate scoring [15]. Grammarly has been found to be an effective tool for assisting in marking and providing feedback [28]. AES, such as Grammarly, employs statistical and rule-based methods to analyze text, and have been found to have high reliability and validity [29]. Using Grammarly also removes the risks of individual marker differences, or marker bias towards any AI language model.

The NLP tools, TAALES, TAASSC, and TAACO were selected as they have been shown to be reliable measures of lexical sophistication, syntactic complexity, and cohesion, and have been used in several studies for the analysis of linguistic features in written text [3], [23], [26]. Studies have established their reliability by ensuring score consistency across administration of these tools [20], [21], [22]. The computational indices used in this study are as follows:

D. TAALES Indices

In this study, lexical sophistication was operationalized using word frequency and word range. Word frequency refers to the number of times a particular word appears in a reference corpus. Less frequently occurring words are considered more sophisticated than words that appear more frequently [20]. Measuring how often particular words appear in a corpus provides an indication of the relative commonness or rarity of words. Word range refers to the number of texts in a corpus where a particular word appears. Words with low range values (i.e., words that occur in fewer contexts) are considered more

TABLE I. Selected Indices of Linguistic Features and Corresponding NLP Tools

Linguistic Features	**Selected Indices and Measurement Tools**	
	Selected Indices	***NLP Tools***
Lexical Sophistication	Word Frequency	TAALES
	Word Range	
Syntactic Complexity	Mean Length T-unit	TAASSC
	Mean Length Clause	
	Mean Length Sentence	
	Dependent Clause per T-unit	
	Complex Nominal per Clause	
Cohesion	Positive Connectives	TAACO
	Negative Connectives	
	Word Overlap Between Adjacent Sentences	
	Word Overlap Between Adjacent Paragraphs	

sophisticated. Studies have shown that essays which use words that appear less frequently or words that occur in fewer contexts tend to be considered higher in writing quality [17], [24], [30].

E. *TAASSC Indices*

Syntactic complexity was measured using mean length of T-unit, mean length of sentence, mean length of clause, dependent clause per T-unit, and complex nominals per clause. These indices capture the global, causal, and phrasal complexity in a text and can therefore provide a comprehensive measure of syntactic complexity in written text [15], [25], [31].

F. *TAACO Indices*

Cohesion was measured using positive connectives and negative connectives, which help readers form connections between sentences in a text [32]. Local and global cohesion were also captured using word overlap between two adjacent sentences and word overlap between two adjacent paragraphs respectively [26].

IV. Results and Findings

A. *Effects of AI Language Models on Writing Quality, Lexical Sophistication, Syntactic Complexity and Cohesion*

Table 2 shows descriptive statistics and ANOVA results of AI generated responses by AI language models. Table 2 shows that first, ChatGPT (M = 91.83, SD = 4.05) performed better than Bard (M = 81.34, SD = 3.47) for writing quality (F(202.66), p <.001, ω^2 = .45). Second, ChatGPT (M = -0.47, SD = 0.59) performed better than Bard (M = 0.47, SD = 0.60 for lexical sophistication (F(71.78), p < .001, ω^2 = .23). As lower lexical sophistication scores indicate a higher level of sophistication, this means that ChatGPT had greater lexical sophistication than Bard. Third, ChatGPT (M = 0.49, SD = 0.52) performed better than Bard (M = -0.49; SD = 0.45) for syntactic complexity (F(111.87), p < .001, ω^2 = .32). On the other hand, Bard (M = 0.42, SD = 0.44) performed better than ChatGPT (M = -0.42, SD = 0.36) in cohesion measures (F(130.14), p < .001, ω^2 = .35).

Fig. 1 illustrates the comparative differences between ChatGPT and Bard for writing quality, lexical sophistication, syntactic complexity, and cohesion of AI written output.

As mentioned earlier, this study investigated the effects of AI language models on writing quality and linguistic features of written output from ChatGPT and Bard. The key finding from the study is that the type of AI language model had an effect on writing quality, lexical sophistication, syntactic complexity, and cohesion of AI written output. Specifically, (1) both ChatGPT and Bard achieved high overall writing quality scores, with ChatGPT performing significantly better than Bard; (2) The results of lexical sophistication show that ChatGPT was significantly stronger than Bard in generating uncommon words that occur in fewer contexts; (3) Similarly, the results of syntactic complexity show that ChatGPT was significantly stronger than Bard in generating more complex sentences; (4) In contrast, the results of cohesion show that Bard was significantly stronger than ChatGPT in generating more cohesive texts. These findings suggest that ChatGPT is currently the stronger AI language model in producing higher quality text that uses more sophisticated words and complex sentences (Findings 1–3), whereas Bard is stronger in creating cohesive texts (Finding 4).

TABLE II. Descriptive Statistics and ANOVA Results of AI Generated Responses by AI Language Model

Dependent Variables	**AI Language Models**		**ANOVA Summary**		
	ChatGPT (n=120)	***BARD (n=120)***	***F***	***p***	***ω²***
Writing Quality	91.83 (*M*) 4.05 (*SD*)	81.34 (*M*) 3.47 (*SD*)	202.66	<0.001	0.454
Lexical Sophistica-tion	-0.472 (*M*) 0.586 (*SD*)	0.471 (*M*) 0.597 (*SD*)	71.78	<0.001	0.227
Syntactic Complexity	0.494 (*M*) 0.521 (*SD*)	-0.494 (*M*) 0.453 (*SD*)	111.87	<0.001	0.317
Cohesion	-0.418 (*M*) 0.364 (*SD*)	0.418 (*M*) 0.442 (*SD*)	130.14	<0.001	0.352

a. Note: Standardized means reported for Lexical Sophistication, Syntactic Complexity and Cohesion

To the best of available knowledge, the current study has not been previously researched. Furthermore, there is also a paucity of studies that examine how AI language models perform linguistically. The findings of this study are in line with the qualitative findings of [12] and [13], who found that ChatGPT could produce well-written and elaborated high-quality texts. ChatGPT's use of relevant content-specific words and its ability to form sentences which were appropriately complex may have contributed to the authors' qualitative assessments that ChatGPT's responses to task prompts were well-elaborated and well-written, leading to its high performance in undergraduate and postgraduate tasks.

B. *Explanation of Effects of AI Language Models on Linguistic Features*

The most feasible reason which could explain ChatGPT's performance compared to Bard is the differences in their training data and methods. ChatGPT likely performed better than Bard in lexical sophistication due to its autoregressive training approach, where it can predict its next word through pattern recognition and statistical likelihood which is based on words preceding its current word. This may have allowed ChatGPT to draw from a wider range of word tokens. As it predicts its output from left-to-right, its word prediction is less constrained compared to Bard. ChatGPT's training approach appeared to give it a lexical advantage. In a similar vein, ChatGPT's training approach could have explained how it outperformed Bard in constructing more complex sentences. Its autoregressive approach facilitated more flexible and creative text generation. In contrast, Bard's masked language modelling approach limited Bard to a set of pre-determined phrases or patterns. This difference may have allowed ChatGPT to create output which had higher syntactic complexity compared to Bard.

A major area where Bard outperformed ChatGPT was in how it produced texts which were more cohesive. One possible explanation for how Bard produced texts which were significantly higher in cohesion compared to ChatGPT is the differences in the models' underlying architecture. It appears that the bidirectional architecture of Bard optimized the model for creating cohesive texts, as it created the condition where Bard can consider the text as a whole. In contrast, ChatGPT's GPT-3.5 framework generated its words unidirectionally, which left it at a disadvantage in creating more cohesive texts, as it only used preceding words for reference when predicting

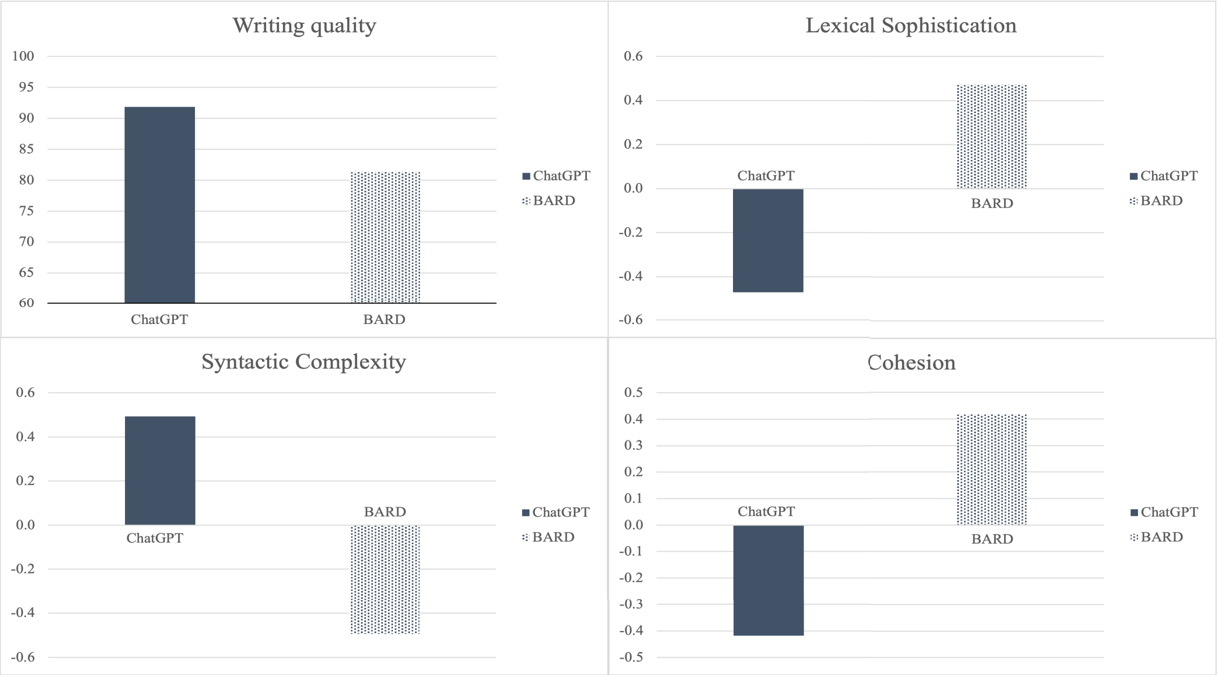

Fig. 1. Linguistic Features of AI written output by AI language models

its next word and cannot consider its generated text in its entirety. The findings suggest that in AI language model output, there appears to be a tradeoff between lexical sophistication and coherence. This was evidenced by ChatGPT's ability to generate text with more sophisticated words, whilst Bard was able to create more cohesive texts. Although uncommon or advanced words may enhance the quality of text, if overused or used inappropriately, it may result in incoherent or low readability text. In generating their text, AI language models may need to strike a balance between the use of advanced or uncommon words whilst ensuring clarity and logical flow. For example, using several infrequent and uncommon words within one text may lead to greater lexical sophistication, but it lowers word overlap between sentences and paragraphs. This may result in ChatGPT having an advantage in generating texts not only with greater lexical sophistication, but also with less cohesion compared to Bard. Although not directly comparable, these observations mirror previous findings on novel syntactic structures and cohesion [18], where AI can generate random words which would create highly novel syntactic structures but are incoherent. However, to substantiate these explanations, further investigations will need to be carried out to examine the relationship between lexical sophistication and coherence in AI generated texts.

C. *Conclusion and Implications*

The effects of AI language models on the various linguistic features of AI written output highlight the strengths and weaknesses of each model's training methods. To reiterate, the findings from this study point to how the different training methods and underlying model architectures manifest as linguistic differences between both models, where ChatGPT creates output with greater lexical sophistication and syntactic complexity, while Bard outperforms ChatGPT in its ability to write more cohesive texts. Most research from previous studies focus on the application of ChatGPT in teaching, learning, and assessment. However, there is a paucity of studies that examine how AI language models perform linguistically. The lack of information on how AI language models perform linguistically creates uncertainty for language teachers when selecting which AI language model to use for what purpose. Yet the direction that the global education community is moving towards is in incorporating AI language models into various stages of their workflow.

The results from this study pointed to one overarching implication, that is, ChatGPT and Bard open opportunities for language teachers to customize language and content for learners. Language teachers can use Bard to create appropriate and relatable model texts to target improving cohesion in writing, as Bard produces more cohesive texts than ChatGPT. Furthermore, they can customize topics specific to learner needs or proficiency level. Personalizing language input for learners can make targeted linguistic features more salient, which helps students relate to the texts used for language learning. As results from this study show that ChatGPT is stronger in lexical sophistication and syntactic complexity, teachers can use ChatGPT, instead of Bard, to aid in their preparation of material, through incorporating customized academic discourse texts that focus on these linguistic features. Without appropriate and relatable models of target academic genre, some L2 students may rely on a limited set of fixed expressions in writing [33], leading to repetitive writing styles. Using ChatGPT to create appropriate discourse texts for language teaching may increase students' awareness of the words, phrases, and grammatical structures typically used for specific genres that they can model in their own writing.

Furthermore, with ChatGPT and Bard, teachers can quickly and easily generate texts on new topics. Instead of spending time searching for authentic texts on the web, in newspapers or books, teachers can instead use the time to assess the generated text for appropriateness and tailor the output according to learners' language needs. Bard, for instance, will be particularly useful for this aspect of materials development, as its access to the internet allows teachers to create materials related to current events and affairs. Ensuring social relevance [34] refers to the need for teachers to take the societal, political, economic, and educational environment into account. It enables teachers to consider the social relevance of learning, what learners can relate to, and what learners need an awareness of. Using ChatGPT or Bard for this purpose can help to improve the efficacy and contextual appropriateness of teaching academic writing.

D. Limitations

This study does not consider the factual accuracy of AI written output in the measurement of writing quality, lexical sophistication, syntactic complexity, and cohesion. While the quantitative measurements used in this study provide a basis for comparison between both AI language models, the analyses in this study do not make an assessment of the veracity of arguments presented by each AI language model. One notable limitation of AI language models is their tendency to "hallucinate" or provide nonsensical, inaccurate, or contradictory statements. Factual inaccuracies or contradictory statements would negatively impact writing quality and cohesion which future studies can investigate.

References

[1] L. Floridi., and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," Minds and Machines, vol. 30, pp. 681-694, 2020.

[2] D. O. Eke, "ChatGPT and the rise of generative AI: threat to academic integrity?" Journal of Responsible Technology, vol. 13, 100060, 2020. https://doi.org/10.1016/j.jrt.2023.100060

[3] X. Zhang, X. Lu, and W. Li, "Beyond differences: Assessing effects of shared linguistic features on L2 writing quality of two genres," Applied linguistics, vol. 43(1), pp. 168-195, 2022.

[4] X. Lu, "A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development," TESOL quarterly, vol. 45(1), pp. 36-62, 2011.

[5] S. A. Crossley, "Linguistic features in writing quality and development: An overview," Journal of Writing Research, vol. 11(3), pp. 415-443, 2020. https://doi.org/10.17239/jowr-2020.11.03.01

[6] J. A. Read, Assessing vocabulary. Cambridge university press, 2000.

[7] J. M. Norris, and L. Ortega, "Towards an organic approach to investigating CAF in instructed SLA: The case of complexity," Applied linguistics, vol. 30(4), pp. 555-578, 2009.

[8] L. Ortega, "Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing," Applied linguistics, vol. 24(4), pp. 492-518, 2003.

[9] K. Wolfe-Quintero, S. Inagaki, and H. Kim, "Second language development in writing: Measures of fluency, accuracy, and complexity," Honolulu: University of Hawaii at Manoa, Second Language Teaching and Curriculum Center, 1998.

[10] M.A.K. Halliday and R. Hasan, "Cohesion in English," Longman, London, England, 1976.

[11] S. M. Lwin, "Discourse analysis," in Applied Linguistics for Teachers of Culturally and Linguistically Diverse Learners, N. Erdogan and M. Wei, Eds. IGI Global, Hershey. PA, 2019, pp. 239-261.

[12] W. Yeadon, O. O. Inyang, A. Mizouri, A. Peach, and C. P. Testrow, "The death of the short-form physics essay in the coming AI revolution," Physics Education, vol. 58, no. 3, p. 035027, 2023. https://doi.org/10.48550/arXiv.2212.11661

[13] C. Terwiesch, "Would Chat GPT3 get a Wharton MBA? A prediction based on its performance in the operations management course," Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania, 2023.

[14] J. E. Casal and M. Kessler, "Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing," Research Methods in Applied Linguistics, vol. 2, no. 3, p. 100068, 2023. https://doi.org/10.1016/j.rmal.2023.100068

[15] A. Mizumoto and M. Eguchi, "Exploring the potential of using an AI language model for automated essay scoring," Research Methods in Applied Linguistics, vol. 2, no. 2, p. 100050, 2023. https://doi.org/10.1016/j.rmal.2023.100050

[16] Y. Su, Y. Lin, and C. Lai, "Collaborating with ChatGPT in argumentative writing classrooms," Assessing Writing, vol. 57, p. 100752, Jul. 2023. https://doi.org/10.1016/j.asw.2023.100752

[17] B. Guo et al., "How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection," arXiv:2301.07597, 2023.

[18] R. T. McCoy, P. Smolensky, T. Linzen, J. Gao, and A. Celikyilmaz, "How much do language models copy from their training data? Evaluating linguistic novelty in text generation using raven," Transactions of the Association for Computational Linguistics, vol. 11, pp. 652-670, 2023. https://doi.org/10.1162/tacl. a.00567

[19] V. Dentella, E. Murphy, G. Marcus, and E. Leivada, "Testing AI performance on less frequent aspects of language reveals insensitivity to underlying meaning," 2023, arXiv:2302.12313. https://doi.org/10.48550/arXiv.2302.12313

[20] K. Kyle, S. Crossley, and C. Berger, "The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0," Behavior Research Methods, vol. 50, pp. 1030-1046, 2018, doi: 10.3758/s13428-017-0924-4.

[21] K. Kyle, "Measuring syntactic development in L2 writing: Fine-grained indices of syntactic complexity and usage-based indices of syntactic sophistication," [Unpublished doctoral dissertation], Georgia State University, 2016. https://doi.org/10.57709/8501051

[22] S. A. Crossley, K. Kyle, and M. Dascalu, "The Tool for the Automatic Analysis of Cohesion 2.0: Integrating Semantic Similarity and Text Overlap," Behavioral Research Methods, vol. 51, no. 1, pp. 14-27, 2019. https://doi.org/10.3758/s13428-018-1142-4

[23] M. Kim and S. A. Crossley, "Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing," Assessing Writing, vol. 37, pp. 39-56, 2018

[24] D. S. McNamara, S. A. Crossley, and P. M. McCarthy, "Linguistic features of writing quality," Written Communication, vol. 27, no. 1, pp. 57-86, 2010.

[25] J. Dong, H. Wang, and L. Buckingham, "Mapping out the disciplinary variation of syntactic complexity in student academic writing," System, vol. 113, p. 102974, 2023, doi: 10.1016/j.system.2022.102974.

[26] S. A. Crossley, K. Kyle, and D. S. McNamara, "The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality," Journal of Second Language Writing, vol. 32, pp. 1-16, 2016.

[27] B. S. Bloom et al., "Taxonomy of Educational Objectives: The classification of educational goals. Handbook I: The Cognitive Domain," New York: David McKay Co., 1956.

[28] S. Koltovskaia, "Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study," Assessing Writing, vol. 44, p. 100450, 2020. https://doi.org/10.1016/j.asw.2020.100450

[29] Y. Attali and J. Burstein, "Automated essay scoring with e-rater® V. 2," The Journal of Technology, Learning and Assessment, vol. 4, no. 3, pp. 1-30, 2006.

[30] B. Laufer and P. Nation, "Vocabulary size and use: Lexical richness in L2 written production," Applied Linguistics, vol. 16, no. 3, pp. 307-322, 1995.

[31] E. Ziaeian and S. E. Golparvar, "Fine-grained measures of syntactic complexity in the discussion section of research articles: The effect of discipline and language background," Journal of English for Academic Purposes, vol. 57, p. 101116, 2022, doi: 10.1016/j.jeap.2022.101116.

[32] S. A. Crossley and D. S. McNamara, "Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication," Journal of Research in Reading, vol. 35, no. 2, pp. 115-135, 2012.

[33] J. Milton, "Lexical thickets and electronic gateways: Making text accessible by novice writers," in C. N. Candlin and K. Hyland (Eds.), "Writing: Texts, Processes and Practices," pp. 221-243, Routledge, 2014.

[34] B. Kumaravadivelu, "The postmethod condition: (E)merging strategies for second/foreign language teaching," TESOL Quarterly, vol. 28, no. 1, pp. 27-48, 1994