Contemporary Educational Technology, 2024, **16**(2), ep509
ISSN: 1309-517X (Online)

OPEN ACCESS

Research Article

My AI students: Evaluating the proficiency of three AI chatbots in *completeness* and *accuracy*

Reginald Gerald Govender [1*]

0000-0002-3143-4050

[1] University of KwaZulu-Natal, Durban, SOUTH AFRICA
[*] Corresponding author: govenderr4@ukzn.ac.za

Citation: Govender, R. G. (2024). My AI students: Evaluating the proficiency of three AI chatbots in *completeness* and *accuracy*. *Contemporary Educational Technology, 16*(2), ep509. https://doi.org/10.30935/cedtech/14564

ARTICLE INFO

Received: 27 Dec 2023

Accepted: 4 Apr 2024

ABSTRACT

A new era of artificial intelligence (AI) has begun, which can radically alter how humans interact with and profit from technology. The confluence of chat interfaces with large language models lets humans write a natural language inquiry and receive a natural language response from a machine. This experimental design study tests the capabilities of three popular AI chatbot services referred to as my AI students: Microsoft Bing, Google Bard, and OpenAI ChatGPT on *completeness* and *accuracy*. A Likert scale was used to rate *completeness* and *accuracy,* respectively, a three-point and five-point. Descriptive statistics and non-parametric tests were used to compare marks and scale ratings. The results show that AI chatbots were awarded a score of 80.0% overall. However, they struggled with answering questions from the higher Bloom's taxonomic levels. The median *completeness* was 3.00 with a mean of 2.75 and the median *accuracy* was 5.00 with a mean of 4.48 across all Bloom's taxonomy questions (n=128). Overall, the *completeness* of the solution was rated mostly incomplete due to limited response (76.2%), while *accuracy* was rated mostly correct (83.3%). In some cases, generative text was found to be verbose and disembodied, lacking perspective and coherency. Microsoft Bing ranked first among the three AI text generative tools in providing correct answers (92.0%). The Kruskal-Wallis test revealed a significant difference in *completeness* (asymp. sig.=0.037, $p<0.05$) and *accuracy* (asymp. sig.=0.006, $p<0.05$) among the three AI chatbots. A series of Mann and Whitney tests were carried out showing no significance between AI chatbots for *completeness* (all p-values>0.015 and $0<r<0.2$), while a significant difference was found for *accuracy* between Google Bard and Microsoft Bing (asymp. sig.=0.002, $p<0.05$, $r=0.3$ medium effect). The findings suggest that while AI chatbots can generate comprehensive and correct responses, they may have limits when dealing with more complicated cognitive tasks.

Keywords: artificial intelligence, chatbots, generative text, *completeness*, *accuracy*

INTRODUCTION

OpenAI's ChatGPT's popularity reached one million users within five days of its release in 2022, which is astounding compared to other popular online services like social media and entertainment (**Figure 1**). ChatGPT powered by GPT-3 model[1] can generate writing that resembles human language. The advent of technology has brought about significant transformations in multiple domains of society, most notably education (Govender, 2021). The promise of enhancing teaching and learning experiences can be observed through the advancements of technologies and the emergence of several artificial intelligence (AI)-powered tools, like OpenAI chat generative pre-trained transformer (GPT), Google Bard, Microsoft Bing, etc. These AI tools possess sophisticated large language models (LLM), a type of machine learning model that handles natural language processing (NLP). Hence, these tools are equipped with text generation, text classification, conversational question answering, and language translation (Zamfirescu-Pereira et al., 2023) that confer distinctive educational advantages while resembling a chatbot (Hwang & Chang, 2023; Okonkwo & Ade-Ibijola,

[1] At the time of testing.

Figure 1. ChatGPT vs. other online services at one million user mark (adapted from Buchholz, 2023, where *one million backers, **one million nights books, & ***one million downloads)

2021). The use of AI chatbots in educational activities has gained significant attention for its potential to support student engagement and learning processes (Su & Yang, 2023).

These chatbots can engage students in dynamic conversations, providing them with personalized feedback and guidance. Research has demonstrated that AI chatbot technologies, such as ChatGPT, can enhance student interaction and learning processes, enrich learning experiences, and potentially improve student motivation, engagement, and learning outcomes (Adiguzel et al., 2023). A significant attribute of such technologies in education is their ability to provide personalized assistance, guidance, and support to learners whenever needed. However, it is essential to consider both the benefits and potential risks associated with such AI tools to ensure their effective integration into the educational setting. It had been found that emotions and reflections are absent when engaging with AI chatbots (Tlili et al., 2023).

AI chatbots like Open AI ChatGPT, Google Bard, and Microsoft Bing, while providing convenience to access information, give rise to ethical issues, particularly in the realm of education. A potential risk associated with using them is the potential for students to misuse these technologies by using them as a means of cheating. Moreover, instead of helping students work through problems, they can answer questions in full and provide in-depth explanations. The ability of AI chatbots to generate content instantaneously raises concerns about academic integrity. However, while AI chatbots like ChatGPT, Bard, and Bing can generate highly coherent and contextually relevant responses, they may not always produce accurate or reliable information (Khurana et al., 2023; Yin et al., 2022). The double meaning of words and context contributes to the inaccurate output. Besides security concerns such as impersonation and identity theft, NLPs cannot access external information (Yin et al., 2022). Hence, the inability to provide updated information and its algorithm may cause a struggle with complex or unconventional questions.

Furthermore, AI powered tools such as chatbots lack the ability to think critically or interpret complex concepts, which limits their effectiveness in specific educational contexts (Tlili et al., 2023). The convenience and efficacy of AI chatbots are enticing for students. As online education becomes increasingly popular due to the aftereffects of the COVID-19 pandemic (Jones & Sharma, 2020; Naidu, 2022), using AI chatbots in the teaching and learning setting makes it easier for students to receive immediate help with their studies. Simply typing in questions or prompts (input) and receiving instant generative text responses (output) can lead to reliance on these chatbot outputs without engaging in the necessary critical thinking and initiating problem-solving.

RESEARCH PROBLEM

A significant challenge during the COVID-19 pandemic was assessment quality control because of students' dishonest behavior and plagiarism during online assessments (Garg & Goel, 2022; Newton et al., 2022). While there has been a rise in the popularity of online AI power tools that are based on NLP models like ChatGPT, Bard, Perplexity, Shako, etc., there have been reports of educational institutions banning the use of such tools

(Tlili et al., 2023). In addition, plagiarism software *Turn it in* released an update not too long after the release of GPT-3 with AI writing detection capabilities (Dheda, 2023). However, NLP AI tools must be embraced through strategic and responsible implementation in teaching, learning and research (Allam et al., 2023; Hwang & Chang, 2023).

In 2019, OpenAI made a blog post regarding the future of language models and their implications (Radford et al., 2019). The company promoted and encouraged research around AI. Research on bias, misuse and detection of NLP models is vital for future development and social harmony (Borenstein & Howard, 2021; Selwyn, 2022). Hence, this research is essential for developers to feel confident releasing larger and more advanced language models. While NLP tools can answer a question based on input text, the generated human-like output's *completeness* and *accuracy* must be investigated.

THEORETICAL FRAMEWORK

A seminal paper by Turing (1950) begins with the statement, "I propose to consider the question, can machines think?" (p. 433). Early work by Alan Turing gave rise to the imitation game, also known as the Turing test (Gonçalves, 2023), which tests if machines can demonstrate behavior equivalent to humans to the extent that it is indistinguishable (Turing, 1950). It evaluates the machine's capability to engage in natural language conversations and simulate human-like responses (Hodges, 2009). This test has been widely used to assess the level of AI chatbots' cognitive abilities and their ability to pass as human-like entities over the years. Some experts propose alternative methods for evaluating AI chatbots' cognitive abilities and their resemblance to human-like entities (Berrar & Schuster, 2014; Jannai et al., 2023; Moor, 1976; Shieber, 1994), while others argue that the Turing test is simply outdated (Carter, 2023; Shin, 2023).

Bloom's taxonomy has been the standard for question creation and design, demanding different levels of cognitive engagement through a response to a question. Bloom's taxonomy was created by Benjamin Bloom in 1956, which classifies the cognitive demand required in learning into levels of complexity (Forehand, 2010). The classification ranges from simple recall to complex synthesis and evaluation. It is widely used in education and assessment to measure a student's understanding and critical thinking skills (Bibi et al., 2020). Bloom's taxonomy consists of six levels: remember, understand, apply, analyze, evaluate, and create (Forehand, 2010). Each level represents a different cognitive skill that an individual should demonstrate when answering questions or completing tasks. Teachers use Bloom's taxonomy to create more effective assessments that encourage higher order thinking and a deeper understanding of the subject matter. This classification system has proven to be a valuable tool for promoting meaningful learning experiences for students across various educational settings (Bibi et al., 2020).

Wang et al. (2023) points out that *accuracy* refers to the absence of inaccuracy between recorded and real-world quantities while *completeness* is defined as whether all important data is captured. In this study *completeness* was defined as the amount of information that overlapped with the memorandum guide and *accuracy* was defined as the quantity of information (Nguyen, 2021). While the Turing test is often used to evaluate the cognitive ability of AI chatbots, it may not fully capture their proficiency in different cognitive demands (Marcus et al., 2016). Therefore, incorporating Bloom classifications in *completeness* and *accuracy* assessment process can provide a more nuanced understanding of an AI chatbot's cognitive capabilities. Additionally, this approach enables developers to identify areas for improvement and enhance the overall performance of AI chatbots.

The classification of cognitive engagement levels is a valuable tool for educators and assessors to gauge a student's level of comprehension and ability to think critically. Hence, selecting questions categorized in the different levels of Bloom's taxonomy allows for assessing the cognitive ability of AI chatbot responses. By utilizing questions categorized by the various Bloom classification levels, the cognitive ability of AI chatbots in question can be assessed more comprehensively, focusing on *completeness* and *accuracy*. This approach provides valuable insights into AI chatbot's generative text ability and human-like responsiveness to understanding and critical thinking.

Table 1. A crosstabulation between number of questions per examination & Bloom's taxonomy

		Bloom's taxonomy levels						Total
		Remember	Understand	Apply	Analyze	Evaluate	Create	
Number of questions per examination	IT P2 Nov 2019	2	1	1	1	3	1	9
	IT P2 Nov 2020	2	1	0	4	1	2	10
	IT P2 Nov 2021	1	1	3	2	1	2	10
	IT P2 Nov 2022	2	4	3	0	2	2	13
Total		7	7	7	7	7	7	42

Note. Based on 42 questions fed into three AI chatbots (42×3=126)

Table 2. Summary of weighted marks for 42 questions

Variable	Value
n	42
Mean	1.76
Median	2.00
Mode	2.00
Standard deviation	0.95
Range	5.00
Minimum	1.00
Maximum	6.00
Sum	74.00

METHOD

The research design employed in this study was a confirmatory experimental design. Since the hypotheses is described as priori and then examined based on empirical evidence (Nilsen et al., 2020). Given the capabilities of NLP tools, this study examines the response provided by three popular AI chatbot services. In this study, my AI students are chatbot interfaces: Open AI's ChatGPT, Google's Bard, and Microsoft's Bing[2]. The study sets out to assess and determine if differences exist in *completeness* and *accuracy* measurements across the different AI tools (ChatGPT, Google Bard, and Microsoft Bing). The following hypothesis was proposed:

H_0. There is no difference in *completeness* and *accuracy* among the three selected AI-generated responses[3].

Three NLP tools powered by AI, namely ChatGPT, Microsoft Bing and Google Bard, were asked seven random questions per the six Bloom's taxonomy levels (create, evaluate, analyze, apply, understand, and remember). These questions were from the South African national senior certificate information technology grade 12 examination paper two from 2019, 2020, 2021, and 2022. Information technology examination two is a theoretical-based assessment, usually handwritten and requires textual responses, making it ideal for this experiment. This resulted in forty-two questions per AI chatbot (7×6=42), respectively; nine questions were selected from the 2019 exam, 10 from the 2020 exam, 10 from the 2021 exam, and 13 from the 2022 exam (**Table 1**).

Table 2 depicts the summary of the 42 questions. These 42 questions carried a grand total of 74 marks, with a mean (M) of 1.76 and a standard deviation (SD) of 0.95. The minimum weighted mark of a question was one, and the maximum weighted mark was six.

Seminal work by Taylor (1986) defines five characteristics of judging information quality: *accuracy*, *completeness*, *currency*, *reliability*, and *validity*. A comparative assessment study by Fichman (2011) examined responses from a series of question-and-answer websites by adopting a binary metrics scale (yes/no) to rate *accuracy*, *completeness*, and *verifiability* of the responses. Nguyen (2021) conducted a study on medical Internet generated content through Google snippets and adopted a 5-point Likert scale focusing specifically on *accuracy* and *completeness*. Research conducted by John et al. (2010) used regression analysis to assess the answering of information of community driven question-answering services on the Internet by adopting scales based on *accuracy*, *completeness*, *presentation*, and *reasonableness*. A pilot study conducted by Li et al.

[2] As of the later part of 2023 Microsoft Bing chat has changed to Microsoft Copilot.
[3] Null hypothesis.

Table 3. Distribution of questions based on mark-weighting

Mark per exam question	n_1	n_2	Percentage (%)
1	18	54	42.86
2	20	60	47.62
3	2	6	4.76
4	1	3	2.38
5	0	0	0.00
6	1	3	2.38
Total	42	126	100

Note. Consider n_1 to be size within a cell & n_2 to be size of entire sample

(2016) found that *accuracy* and *completeness* were the most used criterion for measuring the quality of an answer on a social web platform. In a computer-based automated essay scoring system, metrics based on *accuracy* and *completeness* were adopted (Ramesh & Sanampudi, 2022). Similarly, Blooma et al. (2008) found that *accuracy* and *completeness* were the key predictors of the best answer in evaluating responses provided by a question-and-answer computer retrieval system. Guided by the plethora of previous studies, this study will adopt *accuracy* and *completeness* in assessing the responses from AI chatbots. As supported by Wang et al. (2023) researchers focus on *accuracy* and *completeness* because these two variables are found to be the most significant in decision-making.

A three-point rating scale in the form a Likert was used to classify *completeness* with 1-*incomplete (limited response)*, 2-*adequate (succinct response)*, and 3-*comprehensive (verbose response)*. An answer is regarded as *Incomplete* if it addresses some aspects of the question, but significant parts are missing or incomplete; *adequate* if it addresses all aspects of the question and provides the minimum amount of information required to be considered complete; and *comprehensive* if it addresses all aspects of the question and provides additional information or context beyond what was expected.

A five-point rating scale in the form Likert was adopted for *accuracy* with 1-*completely incorrect*, 2-*more incorrect than correct*, 3-*approximately equal correct and incorrect*, 4-*more correct than incorrect* and 5-*correct*. The results were listed descriptively and were compared among AI chatbots using the Kruskal Wallis and Mann-Whitney U testing.

The researcher created *completeness* and *accuracy* scales based on relevant literature (Fichman, 2011; Nguyen, 2021; John et al., 2010; Ramesh & Sanampudi, 2022). To ensure consistency, the researcher followed a process of entering questions one at a time into each AI chatbot and prompted the chatbot to be specific in its output. The generative text output formed the dataset, which was then marked against the examination memorandum. The process was repeated twice after a three-five-day interval to ensure internal validation; if the new response's evaluation (marks and scales) differed from the original, it was updated accordingly.

Table 3 shows the mark-weighting allocation based on the randomly chosen questions. Of the 42 questions, two mark-weighted[4] questions were the majority (n=20), followed by one mark-allocated questions (n=18), three mark-allocated questions (n=2), four mark-allocated questions (n=1) and six mark-allocated questions (n=1). There were no five mark-allocated questions.

The 42 questions were fed into the three NLP tools (3×42=126) driven by AI technology, providing human-like responses through its chatbot interface. Based on the 126 entries, there were 54 one-mark weighted questions, 60 two-mark weighted questions, six three-mark weighted questions, three four-mark weighted questions, and three six-mark weighted questions (**Table 3**).

RESULTS

There were four one-mark weighted questions categorized under the understanding level of Bloom's taxonomy that were marked totally incorrect, receiving a zero (**Table 4**). In addition, two six-mark weighted questions categorized under the create level of Bloom's taxonomy were marked incorrect (partially correct). The latter indicates that AI NLP tools may not have a preference in producing a correct answer based on the cognitive level of the question. However, 17 zero marks were awarded, while no remember level categorized questions were marked incorrect.

[4] Mark weighted is the mark allocated to the question as per the examination.

Table 4. Mark weight of question vs. mark awarded

Mark awarded to AI	Weighting of questions marks	Bloom's taxonomy levels							Mark awarded matching mark weight
		Remember	Understand	Apply	Analyze	Evaluate	Create	Total	
0	1		4	0	1	1	2	8	
	2		1	2	1	2	0	6	
	3		0	0	1	0	0	1	
	6		0	0	0	0	2	2	
	Total		5	2	3	3	4	17	
1	1	12	8	6	5	8	7	**46**	
	2	1	0	1	2	0	2	6	
	Total	13	8	7	7	8	9	52	
2	2	8	8	12	9	10	1	**48**	
	Total	8	8	12	9	10	1	48	
3	3				2		3	**5**	
	6				0		1	1	
	Total				2		4	6	
4	4						3	**3**	
	Total						3	3	
Total	1	12	12	6	6	9	9	54	85.0%
	2	9	9	15	12	12	3	60	80.0%
	3	0	0	0	3	0	3	6	83.0%
	4	0	0	0	0	0	3	3	100%
	6	0	0	0	0	0	3	3	0.0%
	Total	21	21	21	21	21	21	126	

Note. There were no five-mark weighted questions

When the mark weighted per question matches the mark awarded, this means that AI received full marks for that question; hence, AI response is similar/matches the criteria set in the memorandum. A total of 52 one marks were awarded. Therefore, an 85.0% matching (scored full marks) was obtained on one-mark weighted questions (**Table 4**). Most of these marks belonged to the remember level of Bloom's taxonomy, while six were partially awarded marks to other levels.

There were 48 two marks awarded; most were from the applied level of Bloom's taxonomy, with one two-mark awarded for a create level question of Bloom's taxonomy. An 80.0% matching was found for two-mark weighted questions (**Table 4**).

Analyze level categorized questions received two three-marks and create level categorized questions received four three-marks, including one six-mark question receiving three out six marks (**Table 4**). A matching of 83.0% was found for three-mark weighted questions (**Table 4**). There were three four-marks awarded for Bloom's Create level questions (**Table 4**). A 100% matching was found for the four mark-weighted questions.

A 0.0% matching was found for six mark-weighted questions, with one six mark-weighted question receiving three out of six attaining 50.0% (**Table 4**), indicating AI's inability to generate an answer that fulfils the total solution. Based on AI NLP tools (my AI students) tested, they achieved greater than 80.0% for one, two, three, four mark-weighted questions in obtaining the full marks but received 0.0% for six mark-weighted questions.

Upon a visual inspection (**Figure 2**), no zero marks were awarded for remember level of Bloom's taxonomy. Suggesting that my AI students tested produced all correct responses for questions structured around recalling facts and basic concepts.

Table 5 shows the scored responses of my three AI students. The total mark of the 42 questions was 74 (**Table 2**); hence, the grand total of all three AI students was 222 (3×74=222).

The awarded mark of the 126 questions fed into three AI tools was 178 out of 222 (80.0%), with M=1.41 and SD=0.87. The minimum scored mark was zero and the maximum scored mark was four.

Microsoft Bing had the highest score, awarded 19 one mark, 18 two marks, and three three marks, while only being awarded a zero once (**Table 6**). ChatGPT was awarded 18 one mark, 16 two marks, and two three marks, while Google Bard had the most incorrect responses (11 zeros). All three AI tools scored a maximum

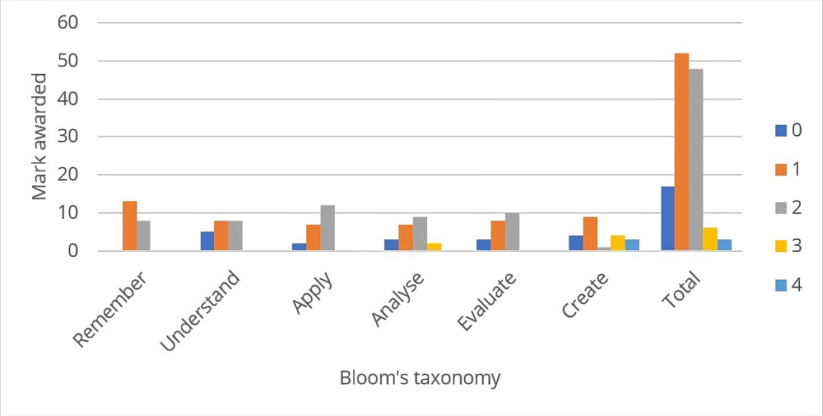

Figure 2. A visual representation showing marks awarded vs. Bloom's taxonomy (0, 1, 2, 3, & 4 are representative of marks awarded) (Source: Author)

Table 5. Awarded mark to AI

Variable	Value
n	126
Mean	1.41
Median	1.00
Mode	1.00
Standard deviation	0.87
Range	4.00
Minimum	0.00
Maximum	4.00
Sum	178.00

Table 6. A crosstabulation between AI tool & marks awarded

AI tools	Bloom's taxonomy levels					Total
	0	1	2	3	4	
ChatGPT	5	18	16	2	1	42
Google Bard	11	15	14	1	1	42
Microsoft Bing	1	19	18	3	1	42
Total	17	52	48	6	3	126

Table 7. A crosstabulation between AI tool & marks awarded

AI tools	Remember	Understand	Apply	Analyze	Evaluate	Create	Mark awarded	Percentage (%)
ChatGPT	9	9	10	12	10	10	60	81.0
Google Bard	10	6	10	7	7	10	50	68.0
Microsoft Bing	10	9	11	12	11	15	68	92.0
Total	29	24	31	31	28	35	178	80.0

Note. Total mark of 126 questions was 74×3=222

of four marks, while the maximum mark weighed question was six. A closer look into the awarded marks reveals Microsoft Bing (92.0%) taking first place, followed by ChatGPT (81.0%) in second place and Google Bard (68.0%) in third place (**Table 7**).

To further assess the performance of my three AI students, **Table 8** depicts that among the 126 AI-generated responses, the median *accuracy* value was five (M=4.48, SD=1.23) and the median *completeness* value was three (M=2.75, SD=0.47). Most responses were limited, obtaining 76.2% for *completeness;* however, containing correct elements achieved 83.3% *accuracy*.

To identify if there exist differences in *completeness* and *accuracy* measurements across my AI students (ChatGPT, Google Bard, and Microsoft Bing), the following hypothesis was tested.

H_0. There is no difference in *completeness* and *accuracy* among the three AI-generated responses.

Table 8. Summary of *completeness* & *accuracy*

Likert scale	Statement	n	Percentage (%)	M	Median	SD	Min.	Max.
Completeness	1-Incomplete (limited response)	96	76.2	2.75	3.00	.472	1	3
	2-Adequate (succinct response)	28	22.2					
	3-Comprehensive (verbose response)	2	1.6					
Accuracy	1-Completely incorrect	9	7.1	4.48	5.00	1.225	1	5
	2-More incorrect than correct	7	5.6					
	3-Approximately equal correct and incorrect	3	2.4					
	4-More correct than incorrect	2	1.6					
	5-Correct	105	83.3					

Note. Based on 126 AI-generated responses

Table 9. Mean rank of independent group (AI) & Kruskal-Wallis results

	AI tools	n	Mean rank	Kruskal-Wallis H	df	Asymp. sig.
Completeness of AI	ChatGPT	42	58.95	6.601	2	.037
	Google Bard	42	72.24			
	Microsoft Bing	42	59.31			
	Total	126				
Accuracy of AI	ChatGPT	42	65.01	10.227	2	.006
	Google Bard	42	54.58			
	Microsoft Bing	42	70.90			
	Total	126				

Note. Test statistic: Kruskal-Wallis test & grouping variable: AI

Table 10. Mann-Whitney U results based on *accuracy* & *completeness*

Comparison	AI tools		Mann-Whitney U	Wilcoxon W	Z-values	Asymp. sig. (2-tailed)
AI comparison 1	ChatGPT vs. Google Bard	*Accuracy*	735.5	1,638.5	-1.791	.073
		Completeness	697.5	1,600.5	-2.365	0.018
AI comparison 2	ChatGPT vs. Microsoft Bing	*Accuracy*	799.0	1,702.0	-1.458	0.145
		Completeness	875.5	1,778.5	-0.072	0.942
AI comparison 3	Google Bard vs. Microsoft Bing	*Accuracy*	654.0	1,557.0	-3.057	0.002
		Completeness	699.5	1,602.5	-2.343	0.190

Note. Alpha=0.015

The Kruskal-Walllis test is a non-parametric test that is an alternative to the one-way analysis of variance ANOVA (McKight & Najab, 2010). The test is performed when one wants to investigate differences among more than two groups and the data is non-normal and qualitative, an ordinal scale/continuous.

This study's independent variable was the group of three AI-powered students: ChatGPT, Google Bard, and Microsoft Bing. The dependent variables were *completeness* and *accuracy*. The responses were tested using a Kruskal-Wallis test to evaluate the differences across three AIs for *completeness* and *accuracy* (**Table 9**). The test revealed significant results for *completeness* (asymp. sig.=0.037, $p<0.05$) and *accuracy* (asymp. sig.=0.006, $p<0.05$) among the three AI chatbots.

Significant differences were found for the independent variables for *completeness* and *accuracy*. Hence, the null hypothesis ($\mathbf{H_0}$) was rejected, and the alternative was accepted ($\mathbf{H_1}$) There is a difference in *completeness* and *accuracy* among the three AI-generated responses. The mean rank of *completeness* (**Table 9**) for ChatGPT, Google Bard, and Microsoft Bing, respectively, were 58.95, 72.24, and 59.31, while the mean rank for *accuracy* (**Table 9**) for ChatGPT, Google Bard, and Microsoft Bing, respectively, were 65.01, 54.58, and 70.90.

Since significant differences exist in *completeness* and *accuracy*, a series of Mann-Whitney U tests were carried out to determine where the difference is (**Table 10**). Mann-Whitney U test is a non-parametric alternative to the independent sample t-test (McKnight & Najab, 2010). A Bonferroni adjustment was implemented to prevent the likely cause of type one error rate inflation (Emerson, 2020). The alpha is 0.05 (5.0%) and the number of comparisons is three (ChatGPT vs. Google Bard; ChatGPT vs. Microsoft Bing, and Google Bard vs. Microsoft Bing), hence 0.05/3=0.015.

A significant difference exists in AI comparison three, *accuracy* (**Table 10**) between Google Bard and Microsoft Bing (p<0.015).

To determine the effect size, the *Z*-value can be used to approximate the value of *r*, as follows: $r = \frac{Z}{\sqrt{n}}$, where *Z* is the *Z* statistics and *n* is the number of cases. *Z*-value is -3.057 (the effect size is of concern and not the direction of effect; hence, the absolute value is considered) and *n* is 126.

$$r = \frac{3.057}{\sqrt{126}} = 0.272.$$

Cohen (1988) highlights that if *r* is .1=small effect, .3=medium effect, and .5=large effect. Hence, 0.272 indicates a medium effect between Google Bard and Microsoft Bing. Search engine browser companies have developed these AI tools; however, only Microsoft Bing extends its generative text from internet sources (Microsoft, 2023).

Non-significant differences were found between the other AI comparisons with effect sizes, further supporting a small to no effect (**Table 10**). ChatGPT vs. Google Bard (asymp. sig.=0.018, p>0.015, r=0.211); ChatGPT vs. Microsoft Bing (asymp. sig.=0.942, p>0.015, r=0.006) and Google Bard vs. Microsoft Bing (asymp. sig.=0.19, p>0.015, r=0.209).

In summary the Kruskal-Wallls test revealed a significant difference in *completeness* (p<0.05) and *accuracy* (p<0.05) (**Table 9**) of generative text response exits among all three AI chatbots.

By carrying out a series of Mann-Whitney tests, further analysis was conducted on *accuracy* by comparing each AI chatbot against the other. Results showed a significant difference between Google Bard and Microsoft Bing (asymp. sig.=0.002, p<0.015, r=0.272 indicative of a medium effect, **Table 10**). Similarly, a series of Mann-Whitney tests were conducted on *completeness*, resulting in non-significant results. Non-significant results are likely because the impact is too small to be captured or the study was underpowered (Edelsbrunner & Thurn, 2023). However, reporting the effect size with the p-values avoids confusion since p-values without effect sizes lack meaning (Visentin et al., 2020). While a significance in *completeness* exists among all three AI chatbots (**Table 10**), when compared to each other, there is little to no significance with all p-values>0.015 and 0<r<0.2 (**Table 10**).

DISCUSSION

An analysis of 126 questions from the differing levels of Bloom's taxonomy revealed an 80.0% matching (scored full marks), providing correct AI text generative answers among three popular NLP chatbots as my students. The score percentage is like an experiment by Zhu et al. (2023) that found an 84.0% pass rate when examining the chatbot on the basic life support exam. Mostly full marks were scored for the lower levels of questions within Bloom's taxonomy (remember and understand), which were usually one and some two mark allocated questions (**Table 4**). Exam questions allocated more than one mark generally demanded more cognitive power since the answers entailed complex thought, reasoning, and specific details. Thus, these types of questions fell into the categories within the upper echelons of Bloom's taxonomy and AI tools investigated in this study did not necessarily perform well. A 0.0% matching was found for six mark-weighted questions, with one six mark-weighted question receiving three out of six attaining 50.0% of the solution. Therefore, my AI students generally performed well on questions that required factual, simple, straightforward answers but struggled to generate responses for questions that required complex cognitive power. The responses to such questions were presented in a complex and confident manner; however, the generated text was disembodied and lacked perspective, displaying symptoms of AI hallucination. This study confirms the findings of Nguyen (2021) regarding low average scores, which were found for *accuracy* and *completeness* for questions that required in-depth specific responses.

Overall, the findings indicate that my AI students have the potential to generate complete and accurate answers. The means and medians of *completeness* (M=2.75; median=3.00) and *accuracy* (M=4.48; median=5.00) were found to be very close, indicating that the data set has a symmetrical distribution. AI chatbot responses rated *comprehensive* on the *completeness* scale[5] (mean *completeness*=2.75; median

[5] Three-point Likert scale.

completeness=3.00) and rated *more correct than incorrect* on *accuracy* scale[6] (mean *accuracy*=4.48; median *accuracy*=5.00) when compared against the memorandum for the Information Technology examinations. While my AI students provided comprehensive answers, their responses were sometimes too broad, offering additional information or context, hindering its *accuracy* (7.1% completely incorrect and 5.6% more incorrect than correct, **Table 8**). This is indicative that text-generative AIs favor more open-ended questions, welcoming varied degrees of responses. The findings of Zhu et al. (2023) affirm that AI responded well to open-ended questions, which revealed an increase from >80.0% correct answers to closed-ended questions to achieving >90.0% to the same question in an open-ended format.

Completeness of the solutions were found to be limited across (76.2% incomplete/ limited response, **Table 8**) all Bloom's taxonomic levels (all question types). While questions weighed more, AI underperformed, as mentioned, likely due to the higher cognitive demand required. Supported by Wang (2023) findings showed that while AI was able to generate impressive answers to Physical Science questions, it struggled with generating answers to conceptual questions requiring abstract thinking. There were significant differences found for the independent variables for *completeness* and *accuracy* among my AI students. As a result, the null hypothesis (**H_0**) was rejected, and the alternative was accepted **H_1.** There is a difference in *completeness* and *accuracy* among the three AI-generated responses. A closer look into the awarded marks reveals Microsoft Bing ranked first, followed by ChatGPT in second place and Google Bard in third. At the time of testing, Microsoft Bing was the only AI tool that harnessed the power of the Internet when generating its text output (Jabotinsky & Sarel, 2022). Thus, responses are a concoction of reinforced learning through a series of NLP algorithms while leveraging the resourcefulness of the world wide web.

The versions of the generative AI text tools utilized in this study should be considered, as well as their ability to generate potentially misleading responses that are presented authoritatively and persuasively. Improvement in the responses can be attributed to version updates, which include refined algorithms and parameters. In addition, repetitive feedback through similar questions asked by users allows for refining responses (Lee & Yeo, 2022). LLMs have the potential to improve and develop rapidly if trained by experts in the subject disciplines, thus churning reliable data and resulting in greater *completeness* and *accuracy* in the dissemination of knowledge. This study demonstrates the potential of generative AI text tools in teaching and learning spaces. Education governing bodies must include training on the potential benefits, limitations, and risks of using AI tools, raising awareness, and encouraging responsible use among students and teachers (Gulyamov & Rustambekov, 2023; Theophilou et al., 2023). This includes user privacy and data security as safeguards, while policies must be implemented to secure personal information and data when using these technologies.

The personalized conversational approach offered by generative AIs through chatbots has the effect of individual and specific feedback, resulting in enhanced student engagement with the subject topic (Adiguzel et al., 2023; Strzelecki, 2023). Moreover, these AI developers offer their platform via the Internet, providing 24 hours/seven days access at the user's convenience. This ensures students have continuous access to educational resources and support, leading to more independent and self-paced learning (Chen et al., 2023). AI's adaptability and customization can significantly benefit students with different learning styles and abilities, allowing for a more tailored and effective learning experience. However, the advanced capabilities to generate personalized and authentic responses may tempt students to rely on these technologies to provide answers during exams or assessments (Tlili et al., 2023). It's worth noting that the text created by all three AI chatbots was very predictable when compared and marked against the memorandum (178 out 222 marks, 80.0%). The sequentially produced words have a general blandness, which is evident; at the heart, these LLMs are algorithms that anticipate what word will follow next. This raises concerns about the validity and reliability of unsupervised assessments as it becomes more difficult to detect cheating. Efforts must be made to ensure the integrity of assessments by implementing stringent proctoring measures or utilizing alternative assessment methods that AI technologies cannot easily manipulate responses. Furthermore, AI tools can also assist teachers in designing curricula and instructional materials. They can provide valuable insights into student learning patterns, identify areas of improvement, and suggest appropriate teaching strategies.

[6] Five-point Likert scale.

Further research with a larger dataset offering more questions from the different Bloom's taxonomy levels is needed to validate the reliability of AI responses. This study used questions based on the Information Technology exam paper; thus, common conditions may impact the data training, leading to improved *completeness* and *accuracy*. There is always the possibility of going beyond information technology content so that the questions represent a wider body of knowledge. Other text-generative AI platforms can be included, and pre-post testing can allow the establishment of a benchmark. Other than text-generative output such as images, audio, etc., this may be of interest since this paper focuses specifically on text.

CONCLUSIONS

The study's analysis of my AI students' performance in generating responses to exam questions based on Bloom's taxonomy revealed that while they excelled at providing correct answers for questions at lower cognitive levels, they struggled with higher-weighted and complex questions requiring higher cognitive abilities. Microsoft Bing demonstrated the highest *accuracy*, followed by ChatGPT and Google Bard, indicating significant differences in the chatbots' performance. The findings suggest that AI NLP tools have the potential to generate complete and accurate answers but may face limitations in handling more complex cognitive tasks. The study also emphasized the need for the responsible use of AI NLP tools in education, highlighting their benefits for students with different learning styles and the potential concerns related to cheating in exams. Further research is recommended to validate AI responses' reliability and address the limitations in their cognitive abilities at higher levels of cognitive engagement.

Funding: This article was supported by National Research Foundation of South Africa (grant number: TTK2204092902).

Ethics declaration: The author declared that ethics approval was not needed since the study did not involve human subjects.

Declaration of interest: The author declares no competing interest.

Data availability: Data generated or analyzed during this study are available from the author on request.

REFERENCES

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology, 15*(3), ep429. https://doi.org/10.30935/cedtech/13152

Allam, H., Dempere, J., Akre, V., Parakash, D., Mazher, N., & Ahamed, J. (2023). Artificial intelligence in education: An argument of Chat-GPT use in education. In *Proceedings of the 9th International Conference on Information Technology Trends* (pp. 151-156). https://doi.org/10.1109/ITT59889.2023.10184267

Berrar, D. P., & Schuster, A. (2014). Computing machinery and creativity: Lessons learned from the Turing test. *Kybernetes, 43*(1), 82-91. https://doi.org/10.1108/K-08-2013-0175

Bibi, W., Butt, M. N., & Reba, A. (2020). Relating teachers' questioning techniques with students' learning within the context of Bloom's taxonomy. *FWU Journal of Social Sciences, 14*(1), 111-119.

Blooma, M. J., Chua, A. Y., & Goh, D. H. L. (2008). A predictive framework for retrieving the best answer. In *Proceedings of the 2008 ACM symposium on Applied Computing* (pp. 1107-1111). ACM. https://doi.org/10.1145/1363686.1363944

Borenstein, J., & Howard, A. (2021). Emerging challenges in AI and the need for AI ethics education. *AI Ethics, 1*, 61-65. https://doi.org/10.1007/s43681-020-00002-7

Buchholz, K. (2023). ChatGPT sprints to one million users. *Statista*. https://www.statista.com/chart/29174/time-to-one-million-users/

Carter, C. (2023). Machines hacking machines–Turing's legacy. In R. K. Nichols, C. M. Carter, C. Diebold, J. Drew, M. Farcot, J. P. Hood, M. J. Jackson, P. Johnson, S. Joseph, S. Khan, W. D. Lonstein, R. McCreight, T. Muehlfelder, H. C. Mumm, J. C. H. Ryan, S. M. Sincavage, W. Slofer, & J. Toebes (Eds.), *Cyber-human systems, space technologies, and threats*. https://kstatelibraries.pressbooks.pub/cyberhumansystems/chapter/6-machines-hacking-machines-turings-legacy-carter

Chen, Y., Jensen, S., Albert, L. J., Gupta, S., & Lee, T. (2023). Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers, 25*(1), 161-182. https://doi.org/10.1007/s10796-022-10291-4

Dheda, G. (2023). Can Turnitin detect ChatGPT? *Open AI Master*. https://openaimaster.com/can-turnitin-detect-chat-gpt/

Edelsbrunner, P., & Thurn, C. (2023). Improving the utility of non-significant results for educational research: A review and recommendations. *PsyArXiv*. https://doi.org/10.31234/osf.io/uxzwg

Emerson, R. W. (2020). Bonferroni correction and type I error. *Journal of Visual Impairment & Blindness, 114*(1), 77-78. https://doi.org/10.1177/0145482X20901378

Fichman, P. (2011). A comparative assessment of answer quality on four question answering sites. *Journal of Information Science*, *37*(5), 476-486. https://doi.org/10.1177/0165551511415584

Forehand, M. (2010). Bloom's taxonomy. *Emerging Perspectives on Learning, Teaching, and Technology, 41*(4), 47-56.

Garg, M., & Goel, A. (2022). A systematic literature review on online assessment security: Current challenges and integrity strategies. *Computers & Security, 113*(6), 102544. https://doi.org/10.1016/j.cose.2021.102544

Gonçalves, B. (2023). The Turing test is a thought experiment. *Minds and Machines, 33*(1), 1-31. https://doi.org/10.1007/s11023-022-09616-8

Govender, R. G. (2021). Embracing the fourth industrial revolution by developing a more relevant educational spectrum. In J. Naidoo (Ed.), *Teaching and learning in the 21st century* (pp. 30-49). Brill. https://doi.org/10.1163/9789004460386_003

Gulyamov, S., & Rustambekovich, R. S. (2023). Code of ethics for the responsible use of AI (chatbots) in science, education and professional activities. *Uzbek Journal of Law and Digital Policy, 1*(3).

Hodges, A. (2009). Alan Turing and the Turing Test. In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the Turing Test*. Springer. https://doi.org/10.1007/978-1-4020-6710-5_2

Hwang, G. J., & Chang, C. Y. (2023). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments, 31*(7), 4099-4112. https://doi.org/10.1080/10494820.2021.1952615

Jabotinsky, H. Y., & Sarel, R. (2022). Co-authoring with an AI? Ethical dilemmas and artificial intelligence. *SSRN*. https://doi.org/10.2139/ssrn.4303959

Jannai, D., Meron, A., Lenz, B., Levine, Y., & Shoham, Y. (2023). Human or not? A gamified approach to the Turing test. *arXiv*. https://doi.org/10.48550/arXiv.2305.20010

John, B. M., Chua, A. Y. K., & Goh, D. H. L. (2010). What makes a high-quality user-generated answer? *IEEE Internet Computing, 15*(1), 66-71. https://doi.org/10.1109/MIC.2011.23

Jones, K., & Sharma, R. S. (2020). On reimagining a future for online learning in the post-COVID-19 era. *SSRN*. https://doi.org/10.2139/ssrn.3578310

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications, 82*(3), 3713-3744. https://doi.org/10.1007/s11042-022-13428-4

Lee, D., & Yeo, S. (2022). Developing an AI-based chatbot for practicing responsive teaching in mathematics. *Computers & Education, 191*(2022), 104646. https://doi.org/10.1016/j.compedu.2022.104646

Li, L., He, D., & Zhang, C. (2016). Evaluating academic answer quality: A pilot study on ResearchGate Q&A. In F. H. Nah, & C. H. Tan (Eds.), *HCI in business, government, and organizations: eCommerce and innovation* (pp. 1-14). Springer. https://doi.org/10.1007/978-3-319-39396-4_6

Marcus, G., Rossi, F., & Veloso, M. (2016). Beyond the Turing test. *AI Magazine, 37*(1), 3-4. https://doi.org/10.1609/aimag.v37i1.2650

McKight, P.E. & Najab, J. (2010) Kruskal-Wallis test. In *The Corsini encyclopedia of psychology* (pp. 1-10). https://doi.org/10.1002/9780470479216.corpsy0491

Microsoft. (2023). What is Bing Chat, and how can you use it? *Microsoft*. https://www.microsoft.com/en-us/bing/do-more-with-ai/what-is-bing-chat-and-how-can-you-use-it?form=MA13KP

Moor, J. H. (1976). An analysis of the Turing test. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 30*(4), 249-257. https://doi.org/10.1007/bf00372497

Naidu, S. (2022). Reimagining and reengineering education systems for the post-COVID-19 era. *Distance Education, 43*(1), 1-5. https://doi.org/10.1080/01587919.2022.2029652

Newton, P. M., & Keioni, E. (2022). How common is cheating in online exams and did it increase during the COVID-19 pandemic? A systematic review. *Journal of Academic Ethics*. https://doi.org/10.1007/s10805-023-09485-5

Nguyen, C. (2021). The accuracy and completeness of drug information in Google snippet blocks. *Journal of the Medical Library Association: JMLA, 109*(4), 613. https://doi.org/10.5195/jmla.2021.1229

Nilsen, E. B., Bowler, D. E., & Linnell, J. D. (2020). Exploratory and confirmatory research in the open science era. *Journal of Applied Ecology, 57*(4), 842-847. https://doi.org/10.1111/1365-2664.13571

Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence, 2*, 100033. https://doi.org/10.1016/j.caeai.2021.100033

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019). *OpenAI*. https://openai.com/research/better-language-models

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*(3), 2495-2527. https://doi.org/10.1007/s10462-021-10068-2

Selwyn, N. (2022). The future of AI and education: Some cautionary notes. *European Journal of Education, 57*(4), 620-631. https://doi.org/10.1111/ejed.12532

Shieber, S. M. (1994). Lessons from a restricted Turing test. *ArXiv*. https://doi.org/10.1145/175208.175217

Shin, B. (2023). The Turing test for measuring AI intelligence is outdated because of ChatGPT's wizardry, and a new test would be better. *Fortune*. https://fortune.com/2023/06/20/turing-test-proposed-update-ai-chatgpt-deepmind-cofounder/

Strzelecki, A. (2023). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2023.2209881

Su, J., & Yang, W. (2023). Unlocking the power of ChatGPT: A framework for applying generative AI in education. *ECNU Review of Education, 6*(3), 355-366. https://doi.org/10.1177/20965311231168423

Taylor, R. S. (1986). *Value-added processes in information systems.* Greenwood Publishing Group.

Theophilou, E., Koyuturk, C., Yavari, M., Bursic, S., Donabauer, G., Telari, A., Testa, A., Boiano, R., Hernandez-Leo, D., Ruskov, M., Taibi, D., Gabbiadini, A., & Ognibene, D. (2023). Learning to prompt in the classroom to understand AI limits: A pilot study. In R. Basili, D. Lembo, C. Limongelli, & A. Orlandini (Eds.), *Proceedings of the 22nd International Conference of the Italian Association for Artificial Intelligence* (pp. 481-496). Springer. https://doi.org/10.1007/978-3-031-47546-7_33

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments, 10*, 15. https://doi.org/10.1186/s40561-023-00237-x

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*, 433-60. https://doi.org/10.1093/mind/LIX.236.433

Visentin, D. C., Cleary, M., & Hunt, G. E. (2020). The earnestness of being important: Reporting non-significant statistical results. *Journal of Advanced Nursing, 76*(4), 917-919. https://doi.org/10.1111/jan.14283

Wang, J. (2023). ChatGPT: A test drive. *American Journal of Physics, 91*(4), 255-256. https://doi.org/10.1119/5.0145897

Wang, J., Liu, Y., Li, P., Lin, Z., Sindakis, S., & Aggarwal, S. (2023). Overview of data quality: Examining the dimensions, antecedents, and impacts of data quality. *Journal of the Knowledge Economy*. https://doi.org/10.1007/s13132-022-01096-6

Yin, D., Dong, L., Cheng, H., Liu, X., Chang, K. W., Wei, F., & Gao, J. (2022). A survey of knowledge-intensive NLP with pre-trained language models. *arXiv*. https://doi.org/10.48550/arXiv.2202.08772

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In A. Schmidt., K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, & M. L. Wilson (Eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-21). https://doi.org/10.1145/3544548.3581388

Zhu, L., Mou, W., Yang, T., & Chen, R. (2023). ChatGPT can pass the AHA exams: Open-ended questions outperform multiple-choice format. *Resuscitation, 188*, 109783. https://doi.org/10.1016/j.resuscitation.2023.109783

◆❖◆