TARBIYA: Journal of Education in Muslim Society, 11(2), 2024, 211-224
DOI: http://doi.org/10.15408/tjems.v11i2.44701
Website: http://journal.uinjkt.ac.id/index.php/tarbiya
p-ISSN: 2356-1416, e-ISSN: 2442-9848

EVALUATING CHATGPT'S ACCURACY ACROSS COGNITIVE LEVELS IN ACADEMIC ASSESMENTS

Astutiati Nurhasanah[1], Fadhilah Suralaga[2*], Ida Rosyidah[3], Zahrotun Nihayah[4], Riri Fitri Sari[5], Ade Solihat[6], Nabila Ernada[7]
[1]National Research and Innovation Agency (BRIN), Tangerang Selatan, Indonesia
[2,3,4] Syarif Hidayatullah State Islamic University of Jakarta, Indonesia
[5,6] Universitas Indonesia, Depok, Indonesia
[7] Design Academy Eindhoven, Netherlands
E-mail: fsuralaga@gmail.com

Received: 14th March 2024; Revised: 27th October 2024; Accepted: 28th December 2024

Abstract

This study evaluates the accuracy of ChatGPT's free version in answering academic questions based on Bloom's Taxonomy cognitive levels (C1–C6) and disciplines (physics, social sciences, and religious studies) at two universities in Jakarta. A mixed-method approach was used, combining statistical and content analyses. Thirty-five lecturers from UIN Jakarta and the University of Indonesia submitted exam questions in Bahasa Indonesia to ChatGPT, and the responses were scored on a 0–100 accuracy scale. Results show that ChatGPT performs well on multiple-choice questions (C1–C3) in physics but struggles with higher-order tasks (C5–C6) requiring synthesis, evaluation, and creativity. In social sciences, accuracy was consistent, particularly in theoretical questions, though ChatGPT faced challenges with data-driven analysis and practical application. Religious studies exhibited high accuracy across all cognitive levels due to the structured and doctrinal nature of the material.Statistical analysis revealed significant differences in accuracy between lower and higher cognitive levels in physics ($p = 0.005$) and religious studies ($p = 0.011$), but no significant difference in social sciences ($p = 0.137$). ANOVA ($p = 0.464$) showed no significant differences across disciplines. This study highlights ChatGPT's effectiveness in answering lower to intermediate-level questions (C1–C4) but identifies limitations with higher-level tasks (C5–C6). These findings encourage educators to design questions that assess deeper cognitive skills while utilizing AI's strengths in supporting learning and knowledge acquisition.

Keywords: ChatGPT; Bloom's Taxonomy; AI in education; cognitive skills; academic assessment

Abstrak

Studi ini mengevaluasi akurasi versi gratis ChatGPT dalam menjawab pertanyaan akademik berdasarkan tingkat kognitif Taksonomi Bloom (C1–C6) dan disiplin ilmu (fisika, ilmu sosial, dan studi keagamaan) di dua universitas di Jakarta. Pendekatan mixed-method digunakan, menggabungkan analisis statistik dan konten. Sebanyak 35 dosen dari UIN Jakarta dan Universitas Indonesia mengajukan soal ujian dalam Bahasa Indonesia ke ChatGPT, dan jawaban yang dihasilkan dinilai pada skala akurasi 0–100. Hasil penelitian menunjukkan bahwa ChatGPT unggul pada soal pilihan ganda (C1–C3) di bidang fisika, tetapi kesulitan pada tugas tingkat tinggi (C5–C6) yang membutuhkan sintesis, evaluasi, dan kreativitas. Pada ilmu sosial, akurasi cenderung konsisten, terutama pada soal teoretis, meskipun ChatGPT menghadapi tantangan dalam analisis berbasis data dan penerapan praktis. Pada studi agama, ChatGPT menunjukkan akurasi tinggi di semua tingkat kognitif karena struktur materi dan interpretasi doktrin yang jelas. Analisis statistik menunjukkan perbedaan signifikan pada akurasi antara tingkat kognitif rendah dan tinggi di fisika ($p = 0,005$) dan studi agama ($p = 0,011$), tetapi tidak pada ilmu sosial ($p = 0,137$). Hasil ANOVA ($p = 0,464$) menunjukkan tidak ada perbedaan signifikan antar disiplin ilmu secara keseluruhan. Studi ini menyoroti efektivitas ChatGPT dalam menjawab soal tingkat rendah hingga menengah (C1–C4) tetapi mengidentifikasi keterbatasan pada tugas tingkat tinggi (C5–C6). Temuan ini mendorong pendidik untuk merancang soal yang mengukur keterampilan kognitif mendalam sambil memanfaatkan kekuatan AI dalam mendukung pembelajaran dan akuisisi pengetahuan.

Kata kunci: *ChatGPT; Taksonomi Bloom; AI dalam pendidikan; keterampilan kognitif; penilaian akademik*

How to Cite: Nurhasanah, A., Suralaga, F., Rosyidah, I., Nihayah, Z., Sari, R. F., Solihat, A., & Ernada, N. (2024). Evaluating ChatGPT's Accuracy Across Cognitive Levels in Academic Assessments. *TARBIYA: Journal of Education in Muslim Society, 11*(2), 211-224. https://doi.org/10.15408/tjems.v11i2.44701

*Corresponding author

Introduction

ChatGPT is an artificial intelligence (AI) program developed by OpenAI (https://chat.openai.com/chat) that engages users in conversational interactions. Trained using deep learning techniques, ChatGPT processes text inputs and generates responses by leveraging large-scale datasets to recognize language patterns, context, and semantics, enabling it to produce human-like responses across a wide array of topics (OpenAI, 2022).

This platform has made a significant impact on academia, aiding academics in gathering and summarizing data, as well as organizing it into coherent writing. This allows for a single summarized version from multiple sources. Kung et al. (2023) suggested that since ChatGPT responses were very consistent, students could easily follow its explanation, making it a useful source of information. This leads to enhancing students' productivity (Ningrum, Saputra, Mahardika, Sari, 2024). Educators can create participatory learning media, compile interesting learning scenarios, and gain more informative teaching materials (Dwiyono, Wahyudi, and Tannarong, 2024). It even able to improve student' skills in writing English, develop an interactive, personalized and efficient educational environment (Lestari, 2024). Additionally, ChatGPT's responses often provide insightful, deductive reasoning, and introducing students to novel and non-obvious concepts (Kung et al., 2023). However, ChatGPT's effectiveness also influenced by the user's perception and interactions (Tu, 2024), showing how its performance also related to the questions and prompts it receive from the user.

While ChatGPT can be a helpful tool for drafting articles, reviewing materials, or rephrasing academic texts, many argue that it should not be relied upon to generate entirely original content. (Arif et al., 2023; Bom, 2023). Concerns have been raised about ChatGPT's potential issues, including ethical, copyright, transparency, and legal concerns, risks of bias, plagiarism, lack of originality, inaccuracies with the risk of hallucination, limited knowledge, incorrect citations, cybersecurity issues, and the potential to spread misinformation (Sallam, 2023). Therefore, it is recommended that the AI bot should be used primarily for brainstorming, generating content like outlines, making summaries, or employing it as an editor (Lingard, 2023).

Furthermore, many are concerned that ChatGPT's ability to automatically generate text by simply inputting questions—while avoiding plagiarism detection—could be misused by students during assessments of their learning and skill development (Arif et al., 2023; O'Connor & ChatGPT, 2023; Teixeira da Silva, 2023). Although ChatGPT draws information from various internet-accessible sources, previous studies have highlighted errors in literature synthesis, citations, problem statements, research gaps, and data analysis (Rahman et al., 2023). It has also been known to produce incorrect or fabricated information and can bypass plagiarism detectors (Lo, 2023). Additionally, due to sourcing from limited databases, ChatGPT may struggle to distinguish between facts and opinions (Suharmawan, 2023). These limitations raise concerns about the reliability of ChatGPT-generated answers in educational contexts, particularly for exam questions, potentially impacting teaching and learning strategies. Students relying on ChatGPT to study past exam papers may receive misleading information if the AI generates inaccurate content. Consequently, educators might need to adjust their assessment methods to ensure that student submissions reflect genuine understanding.

In Indonesia, ChatGPT has been widely adopted in educational settings, ranging from secondary schools to universities. Previous studies have highlighted its potential to enhance learning experiences, improve productivity, and support instructional design. However, most of this research has focused on

DOI: 10.15408/ tjems.v11i2.44701

ChatGPT's impact on learning dynamics, academic integrity, or its effectiveness as a teaching aid, leaving a critical gap in evaluating the accuracy of its responses, particularly within the framework of Bloom's Taxonomy. Accurate answers are crucial in guiding student learning and shaping effective educational strategies. Inaccurate or misleading responses could hinder students' understanding and affect educators' approaches to teaching and assessment.

Empirical studies on ChatGPT's effectiveness in learning strategies and instructional design are scarce, highlighting the need for further investigation (Chang, Chen, & Tang, 2024). Despite ChatGPT's growing presence in educational contexts, limited research has examined its performance in supporting Bloom's Taxonomy-based learning objectives, particularly in the Indonesian context. This study seeks to address these gaps by exploring three key issues: first, the accuracy of ChatGPT's responses to questions posed by educators; second, the variation in accuracy across different academic disciplines (physical sciences, social sciences, and religious studies) and cognitive levels as classified by Bloom's Taxonomy; and finally, the implications of these findings for educators' learning strategies and instructional design.

Bloom's taxonomy has been serving as a valuable framework for many educators in Indonesia in planning and structuring curricula as well as effective teaching strategies to achieve an enhancement in student competencies. It is a framework used to classify learning objectives based on the level of complexity of thinking and cognition. It categorizes learning objectives according to the complexity of cognitive processes, ranging from the simplest to the most complex: remembering, understanding, applying, analysing, evaluating, and creating (Armstrong, 2010), which, in this study, are assigned as Cognitive Levels 1 (C1) to 6 (C6), respectively.

Method

This study employed a mixed-method approach, utilizing statistical analysis alongside content analysis. Statistical analysis was used to describe the accuracy level of ChatGPT's responses and to examine differences in accuracy across Bloom's cognitive levels and academic disciplines. Content analysis was applied to identify the characteristics of the exam questions and the answers provided by ChatGPT (Krippendorff, K., 2004). The evaluation was conducted using ChatGPT online platform (https://chat.openai.com/chat) (accessed on 18 March 2023 - 23 February 2024).

Participants were instructed to input examination questions they had previously posed to their students during mid-term or end-of-term assessments into ChatGPT. These questions were classified by the participants according to Bloom's Taxonomy cognitive levels (C1–C6) and academic disciplines (physical sciences, social sciences, religious studies). The questions were submitted in Bahasa Indonesia, and ChatGPT generated responses in the same language. Each response was then reviewed by the same participant for accuracy and scored on a scale from 0 to 100, with 0 indicating complete inaccuracy and 100 indicating complete accuracy. Participants also provided a rationale for the scores they assigned.

The study involved 35 lecturers from higher education institutions in Jakarta: State Islamic University Syarif Hidayatullah Jakarta (representing religious-based university) and University of Indonesia (representing general university), who collectively taught 51 different subjects. Of these lecturers, 22 taught physical science-related subjects, 22 taught social sciences, and 7 taught subjects related to religious studies. The number 35 does not mean to be representative, yet it is still illustrative samples from those who are willing to be research participants. The participants were asked to make exam questions in which there are similar or varying levels of question, put them into ChatGPT, then gave

DOI: 10.15408/ tjems.v11i2.44701

scores to the ChatGPT's responses. Due to the limited number of participants, religious studies do not have educators that offer C6 level question.

The collected data were categorized into two cognitive groups based on Bloom's Taxonomy: lower cognitive levels (C1–C3) and higher cognitive levels (C4–C6) (Anderson & Krathwohl, 2001). These classifications were then used to conduct the subsequent analysis. Statistical analyses, including T-Tests and ANOVA (Field, A. 2018; Hair, J.F, W.C., Babin, B.J, & Anderson, R.E, 2019), were conducted using Microsoft Excel. Two-tailed T-Tests assuming unequal variance (heteroscedastic) were performed to assess the significance of differences in accuracy scores between Bloom's lower cognitive levels (C1 – C3) and higher levels (C4 – C6), within each academic discipline. Additionally, One-Way ANOVA was utilized to evaluate the significance of differences in accuracy scores across the different academic disciplines.

Findings and Discussion

Findings

The accuracy scores of ChatGPT's responses to questions posed by lecturers are summarized in Tables 1, 2, and 3, categorized by academic disciplines: physical sciences, social sciences, and religious studies, respectively. The cognitive levels of the questions, according to Bloom's Taxonomy, are also included in these tables. A total of 83 questions were analyzed, with 54 of these falling within the lower cognitive levels (C1–C3), while the remaining questions corresponded to higher cognitive levels (C4–C6) (Figure 1).

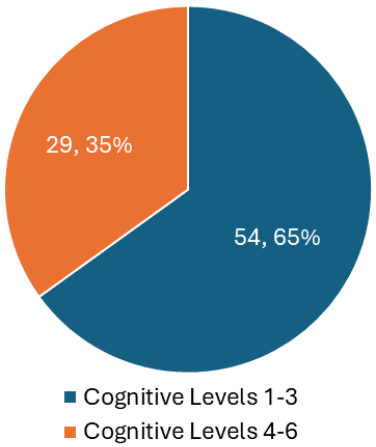

Figure 1. Distribution of the 83 questions addressed by lecturers to ChatGPT. The numbers within the pie chart indicate the count of questions at each cognitive level and their respective percentages of the total

As could be observed in Table 1, the accuracy scores of ChatGPT response to C1 – C3 questions addressed in physical science subjects ranged from 0 to 100. The 0 and 100 accuracy scores were given to ChatGPT's response to multiple-choice questions that required precise, definitive answers, so the scoring options were only either accurate (100) or inaccurate (0). Out of the forty physical science-related questions posed to ChatGPT, 30 were at cognitive levels C1–C3, while the remaining 10 were at levels C4–C6. Additionally, 10 of the questions were multiple-choice, with eight at the C1–C3 level and two at the C4–C6 level.

DOI: 10.15408/ tjems.v11i2.44701

Table 1. The Accuracy Scores for Chatgpt's Responses to Physical Science-Related Questions, Categorized By Course and Bloom's Cognitive Level

	Course	Cognitive Levels 1-3	Score	Cognitive Levels 4-6	Score
1	Gastrointestinal Module	C2	70	C4	70
2	Community Medicine	C2	0		
		C2	0		
3	Medical Ethics	C2	100		
4	Molecular Biology	C2/C3	100		
			100		
			100		
			0		
5	Research Methodology	C2	100		
		C2	100		
6	Principles of Evolution	C2	90		
7	Evolution	C2	85		
8	Conservation Biology	C2	95		
9	Primatology	C1	100		
10	Mathematics	C3	100	C4	0
		C1	100	C6	0
11	Cell Biology	C2	100	C6	10
				C6	10
12	Pharmacy Laboratory Practical Class			C4	75
				C4	75
				C4	75
13	Pharmacy Laboratory Theory			C4	75
14	Technology of Tablet Preparation	C1	100		
		C1	100		
		C1	100		
15	Gerontological Nursing	C1	100		
		C2	100		
16	Family Nursing	C2	100		
17	Environmental Chemistry	C2	80		
		C3	80		
18	Eye Anatomy	C1	100		
19	Mathematics			C6	50
20	Biology	C 2	80		
21	Pharmacy	C 3	100		
		C 3	100		
22	IT	C2	80		

Table 2. The Accuracy Scores For Chatgpt's Responses to Social Science Questions, Categorized by Course And Bloom's Cognitive Level

	Course	Cognitive Levels 1-3	Score	Cognitive Levels 4-6	Score
1	Behaviour modification	C1	100		
		C2	100		
2	Forensic Psychology	C2	80		
3	Theory of Social Change			C5/C6	35
				C4	50
4	Mental Health	C2	80	C4	70
5	Pre-Marriage Counselling	C3	80	C4	80
6	Innovation Management			C4	70
7	Pancasila and Citizenship			C4	95
8	Intercultural and Interfaith Communications	C2	85		
		C3	85		
9	Politics and Communication			C4	80
10	Management of Public Relation and Service	C1	90		
		C2	90		
11	Gender in Islam	C2	95		
12	Counselling Psychology	C2	95		
13	Counselling Guidance			C4	85
14	Education Psychology 2			C4	90

DOI: 10.15408/ tjems.v11i2.44701

	Course	Cognitive Levels 1-3	Score	Cognitive Levels 4-6	Score
15	Creativity Theory and Development			C6	70
16	Development Theory			C4	90
17	Islamic Statehood			C4	70
18	Sociology	C3	95		
19	Cultural Anthropology			C4	95
20	Islamic Economics			C4	98
21	Sharia Economics	C2	80		
22	Publication and Publishing	C2	50		

Table 3. The Accuracy Scores For Chatgpt's Responses to Questions From Religious Studies, Again Classified by Courses and Bloom's Cognitive Levels

	Course	Cognitive Levels 1-3	Score	Cognitive Levels 4-6	Score
1	Modern Islamic Philosophy	C2	100		
		C2	100		
2	Local Religions	C2	85	C4	78
3	World Religions	C2	88		
4	Minor Religions	C2	85		
5	Tafseer (Interpretation of the Quran)	C2	80	C4	75
		C3	90		
6	Islamic studies	C1	80	C4	80
		C2	90		
		C3	70		
7	History of Islamic Civilisation			C4	80
				C4	72

Of the 28 social science questions (Table 2), 14 were at cognitive levels C1–C3, and the remaining 14 were at levels C4–C6. Table 2 describes that the accuracy scores for C1–C3 questions in social science discipline ranged from 50 to 100. A score of 50 was assigned to a question related to publication and publishing procedures because ChatGPT failed to provide a complete explanation of the stages involved, which could critically impact the outcome of the publishing process. This omission might be attributed to the unavailability of comprehensive online information following the implementation of new publishing regulations (Nabawi et al., 2022). In contrast, questions from courses such as Behavior Modification and Forensic Psychology, which involved well- documented theories and concepts, received full accuracy scores of 100.

The accuracy scores of ChatGPT's responses to C4-C6 questions within the social science courses ranged from 35 to 95. Lower scores (35 and 50) were given for responses to questions in the Theory of Social Change course, where ChatGPT provided a general explanation of concepts but failed to delve into specific data or elaborate on phenomena based on empirical observations. On the other hand, ChatGPT achieved higher accuracy scores (85–95) in courses like Educational Psychology, where it successfully integrated analysis and evaluation according to the relevant theories.

A total of 15 questions were posed from the religious study discipline (Table 3), with 10 at cognitive levels C1–C3 and 5 at levels C4–C6. As can be seen in Table 3, across all questions in religious studies, ChatGPT's responses consistently received high accuracy scores (>70), regardless of the cognitive level. In physical science subjects (Figure 2A), the majority (86.67%) of responses to C1–C3 questions were highly accurate (scores >70). However, for C4–C6 questions, equal number (40%) of responses fell within the lower (scores ≤35) and higher (scores >70) accuracy ranges. The figure also describes that for both C1 – C3 and C4 – C6 level questions within the physical science courses, only smaller percentage received medium scores (35<score≤70). This trend is likely influenced by the fact that 25% of the physical science questions were multiple-choice, leading to a binary scoring outcome (either fully accurate or inaccurate).

DOI: 10.15408/ tjems.v11i2.44701

In contrast, as shown in Figure 2A and Figure 2B, a higher percentage of ChatGPT's responses to exam questions in social sciences and religious studies achieved high accuracy scores, irrespective of the cognitive level of the questions.

The results of the statistical analyses are presented in Table 4 and Table 5. The analyses revealed that within the physical science disciplines, there was a highly significant difference in accuracy scores between Bloom's lower (C1-C3) and higher (C4-C6) cognitive levels ($p = 0.005$). Similarly, in religious studies courses, the accuracy scores also showed a significant difference between these cognitive levels ($p = 0.011$, $p < 0.05$). In contrast, within the social sciences, the T-test indicated no significant difference in accuracy scores between the lower and higher Bloom's levels ($p = 0.137$, $p > 0.05$). Furthermore, ANOVA results showed no significant difference in ChatGPT's performance across the different academic disciplines ($p = 0.464$).

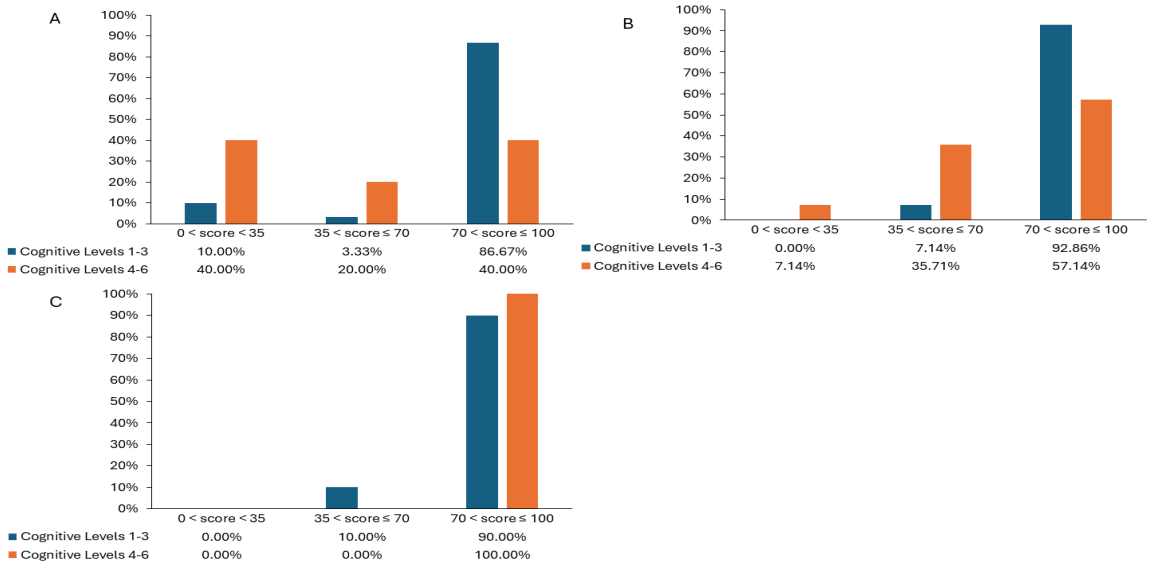

Figure 2. The Distribution of Accuracy Scores For Chatgpt's Responses Within the Science (A), Social Sciences (B), and Religious Studies (C) Disciplines

Table 4. P - Values From T - Tests Performed to Measure the Significance of Accuracy Score Differences Between Bloom's C1 - C3 And C4 - C6, Within Each Academic Discipline.

	Science	Social Science	Religious Study
P value	0.005	0.137	0.011

Table 5. One-Way ANOVA Results to Analyse the Significance of Score Differences Between Academic Disciplines

Groups	Count	Sum	Average	Variance		
Science	40	3000	75	1280.769		
Social	28	2283	81.53571	254.9246		
Religious Studies	15	1253	83.53333	79.98095		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1122.989	2	561.4946	0.775107	0.464082	3.110766
Within Groups	57952.7	80	724.4087			
Total	59075.69	82				

DOI: 10.15408/ tjems.v11i2.44701

Discussion

The findings of this study provide valuable insights into the accuracy of ChatGPT's responses to examination questions across different academic disciplines and cognitive levels. Overall, the results demonstrate that while ChatGPT can provide highly accurate answers to certain types of questions, its performance varies significantly depending on the subject matter and the complexity of the questions, as classified by Bloom's Taxonomy.

Accuracy in Physical Science Disciplines

In the science disciplines, ChatGPT's accuracy was notably high for questions classified at the lower cognitive levels (C1–C3), particularly for those requiring precise, factual answers. This outcome is consistent with the nature of C1–C3 questions, which often involve recall and understanding of well-established facts and concepts (Armstrong, 2010). One example got a score 100 by the educator:

Two reasons that cause genetic mutation are … and …

ChatGPT: Genetic mutations can be caused by various factors, but two main causes of genetic mutations are chemical mutagens and radiation.

- **Chemical Mutagens:** Many chemical compounds can cause genetic mutations by damaging or altering the structure of DNA. Examples of chemical mutagens include substances such as cigarette smoke, industrial pollutants, chemicals in certain foods and beverages, and other chemicals that can lead to DNA damage.
- **Radiation:** Exposure to radiation, whether in the form of X-rays, gamma rays, or exposure to alpha, beta, or neutron radiation, can damage the structure of DNA and cause mutations. Radiation can also come from natural sources like cosmic radiation or from human activities such as medical or nuclear radiation exposure.

Both of these factors can damage DNA molecules, disrupt the DNA base sequence, or cause structural changes in chromosomes. As a result, genetic mutations can occur, which can affect gene expression, protein function, or lead to the development of genetic diseases."

On the other hand, the accuracy of responses decreased markedly for higher cognitive level questions (C4–C6), where a deeper understanding, analysis, and synthesis of information were required (Armstrong, 2010). However, most of the questions in C4 physical sciences, although came in the form of study cases, were ended with multiple choices, helping the AI to answer more correctly.

Furthermore, one question example categorized as a C6 question, which asking the students to create a picture of a scheme, shows the limitation of the Free ChatGPT, which the AI was unable to generate, resulting a low score of just 10.

The prevalence of low scores in the case of C6 questions suggests that while ChatGPT excels at retrieving and reproducing factual information, it struggles with more complex tasks that involve critical thinking and the application of knowledge in novel contexts (Cong-Lem et al., 2024).

The T-test result of the physical science disciplines ($p = 0.005$) further emphasizes that ChatGPT is more adept at answering questions that require lower-order cognitive skills, such as remembering and understanding. This aligns with previous observations where factual and well-defined answers, typically

DOI: 10.15408/ tjems.v11i2.44701

required at these cognitive levels, were accurately provided by ChatGPT. Interestingly, although the AI's performance has been observed as deteriorating with more complex, higher-order tasks, such as analysis, evaluation, and creation, which likely require deeper understanding and synthesis of information (Cong-Lem et al., 2024; Ray, 2023), the AI can still perform well enough (with scores around 70) on analysis and evaluation level due to the type of question being multiple choices, reflecting ChatGPT's ability to identify correct answers from a fixed set of options.

However, these findings contrast with those of Suárez et al. (2023), who examined ChatGPT's accuracy in answering endodontic (dental medicine) questions and found that the AI performed worse on easy questions compared to medium and difficult ones. In their study, ChatGPT was presented with questions categorized into three difficulty levels (Easy, Medium, Difficult), each requiring a simple Yes or No answer. The highest accuracy was observed at the medium level, followed by the difficult level, with the lowest accuracy on easy questions. This discrepancy may be due to differences in question complexity and format, as this paper involved more comprehensive questions rather than binary Yes/No responses. Further investigation is needed to understand the underlying factors contributing to these contrasting results.

Performance in Social Sciences

In contrast to the physical sciences, ChatGPT's performance in the social sciences has more variable, but generally higher for both lower (C1–C3) and higher (C4–C6) cognitive level questions. The AI demonstrated strong capabilities in addressing questions involving popular theories and well-documented concepts, as evidenced by the full accuracy scores received in subjects like Behavior Modification and Forensic Psychology.

However, the lower scores observed for some C4–C6 level questions, particularly in courses such as Theory of Social Change, highlight ChatGPT's limitations in areas requiring deep contextual understanding and the integration of empirical data. This finding aligns with previous research indicating that while ChatGPT can effectively summarize and explain theoretical content, it may falter when asked to apply these theories to specific, real-world scenarios that require detailed, context-specific knowledge (Cong-Lem et al., 2024). The moderate to high scores in other C4–C6 level questions within the social sciences suggest that ChatGPT can perform well when the questions align closely with its training data, particularly in cases where comprehensive analysis and evaluation are involved, but its performance is less reliable in more complex, data-driven inquiries (Cong-Lem et al., 2024).

The statistical analysis revealed the lack of significant differences in accuracy scores between lower and higher cognitive levels in the social sciences, suggesting that ChatGPT's performance is more consistent across these levels within this discipline. Social science questions may often involve concepts and theories that are widely discussed and documented, allowing the AI to provide more accurate responses even at higher cognitive levels. However, this consistency might also indicate that ChatGPT struggles equally with more complex cognitive tasks within this domain.

Performance in Religious Studies

The consistently high accuracy scores across all cognitive levels in religious studies suggest that ChatGPT is particularly effective in this discipline. This may be due to the structured and well-documented nature of religious texts and interpretations, which provide a clear basis for the AI to generate accurate and reliable responses. The high scores across C1–C3 and C4–C6 levels indicate that ChatGPT

DOI: 10.15408/ tjems.v11i2.44701

can handle both simple and complex inquiries within this field, likely because the questions involve well-established doctrines and interpretations that are well- represented in its training data. The significant difference between C1–C3 and C4–C6 levels in religious studies ($p = 0.011$) reflects ChatGPT's varying performance across different cognitive levels within this domain.

What might have caused the observed variation in ChatGPT's accuracy across different academic disciplines?

ANOVA results, indicating no significant difference in ChatGPT's performance across the different academic disciplines ($p = 0.464$), suggest that while there are variations in accuracy within disciplines, the overall performance of the AI is relatively uniform when comparing physical sciences, social sciences, and religious studies as a whole. This finding reinforces the idea that the observed variations in accuracy are more related to the cognitive demands of the questions rather than the subject matter itself.

These results underscore the importance of considering both the cognitive level and the discipline when evaluating the effectiveness of AI tools like ChatGPT in educational settings. While the AI demonstrates strong capabilities in handling lower-order cognitive tasks, its limitations become evident with more complex, higher-order questions, particularly in disciplines that require deep contextual understanding and critical thinking. Educators should be mindful of these strengths and limitations when integrating AI into their teaching and assessment practices.

The observed variation in ChatGPT's accuracy across different academic disciplines likely stems from several factors, including the nature of the content and the exam question within each discipline and the structure and availability of training data and the ChatGPT's features:

Nature of the Question and Content Across Disciplines

Disciplines like the physical sciences often deal with well-established, objective knowledge that is factual and structured. Interestingly, the educators within this field often use multiple choices rather than open-ended essays, providing ChatGPT with the opportunity to just choose rather than elaborate, even up to C4 level. The AI's training likely includes vast amounts of structured data from scientific literature, textbooks, and educational content, which enables it to retrieve and present factual information with high accuracy (Koteluk et al., 2021). This shows, from the questions at the C1 to C4 that tends to involve recall or simple application of these facts, compounded by the multiple choice questions, achieve high scores across different physical science subjects. However, when it gets to the highest cognitive level (C6), it still doesn't have the capacity to achieve high score (> 50) due to higher demands of creative thinking and the limitation of ChatGPT's free feature that unable to generate picture-based answers.

On the other and, social sciences often involve more nuanced, interpretative, and context-dependent content, but ChatGPT still achieve great accuracy until C4 level due to its vast recorded data on the social science subjects. However, on the C6 level, which is the ability to create, ChatGPT achieve varying score, but not higher than 70. ChatGPT's training data, while extensive in social science, may not fully capture the depth and context required for the highest task in Bloom's Taxonomy.

Religious studies, on the other hand, often involve content that is more canonical and well-documented, with a clear structure in doctrinal texts. The AI's ability to accurately reproduce well-established interpretations of religious texts likely contributes to the higher accuracy observed in this

DOI: 10.15408/ tjems.v11i2.44701

area across level C1 until C4. However, since there was no C5-C6 level question, this research is unable to produce insights on how ChatGPT will answer a C5-C6 level question in religious studies.

Structure and Availability of Training Data and ChatGPT's Feature

The availability and quality of training data significantly impact ChatGPT's performance across disciplines. All three disciplines studied in this research have a wealth of well-organized, structured information that the AI can learn from, and the free ChatGPT performed well in all disciplines on the lower levels of Bloom's Taxonomy (C1-C3), even until C4. However, the type of question influence how ChatGPT can answer, such as how it shows with multiple choice in physical science enable the AI to achieve higher score and how the question that demands the AI to answer with a picture leads to lower score because the free feature cannot generate picture.

Implications for Educational Use

The results of this study have important implications for the use of ChatGPT in educational settings. While the AI demonstrates considerable potential as a tool for assisting with certain types of learning, particularly in subjects and cognitive tasks that align with its strengths, educators must be aware of its limitations (Yu, 2024).

The study indicates that C1-C4 level questions dominate educators' exam questions, and ChatGPT can provide highly accurate responses to these questions, especially when they come in the form of multiple choices. To prevent students from relying entirely on ChatGPT-generated answers, educators can create more varied and open-ended essay questions, developing questions at higher cognitive levels. Indeed, not all C1-C4 questions can be elevated to higher levels, but educators must still have strategies for assessing student responses to distinguish between answers taken entirely from ChatGPT and those elaborated through students' critical thinking. Certainly, identifying this is not easy, as many experts believe that detecting AI usage can be very challenging (De Carvalho, 2023; Fowler, 2023; Mujezinvoic, 2023 in Marais, E., Marais-Botha, R. & Coertzen, F., 2024). Therefore, to evaluate students' learning outcomes more objectively and comprehensively, educators can combine exam questions with project-based assignments, case studies, practical exercises, or problem-solving tasks that inherently require students to think at the C5 and C6 levels.

Conclusion

This study highlights the strengths and limitations of ChatGPT as an educational tool, particularly in its ability to answer examination questions across different academic disciplines and cognitive levels, making it an interesting novel learning medium for students. While the free AI feature performs well in areas requiring recall and understanding of factual information, its ability to handle more complex, creative, higher-order tasks like generating image is less reliable. Educators should consider these findings when incorporating AI into their teaching strategies, ensuring that students are encouraged to critically evaluate AI-generated content and supplement it with their own analysis and understanding, incorporating other references/learning sources, to develop more accurate and comprehensive understanding of their subjects.

The variation in ChatGPT's accuracy across different academic disciplines is multifaceted, involving the nature of the question and the content and the availability and structure of training data

DOI: 10.15408/ tjems.v11i2.44701

and ChatGPT's features. Understanding these factors is crucial for educators and researchers in designing examination and to prevent students from using ChatGPT without critical thinking, as it highlights the need for careful consideration of the AI's strengths and limitations in different disciplinary contexts.

We realize that this study is preliminary. However, it would be valuable to conduct further analysis to determine what precautionary measures lecturers should take to prevent the misuse of ChatGPT and other AI applications in exam settings and academic assignments, the use of ChatGPT from the perspective of students from various universities, and the use of ChatGPT by educators and students in relation to academic integrity and the factors that influence it.

Acknowledgements

We extend our sincere gratitude to all the participants in this study, whose contributions were essential to the realisation of our research. We are especially thankful to our fellow members of the MAI for their invaluable support throughout this study.

References

Anderson, L.W. & Krathwohl, D.R. (Ed.) (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Addison Wesley Longman, Inc.

Arif, T. Bin, Munaf, U., & Ul-Haque, I. (2023). The future of medical education and research: Is ChatGPT a blessing or blight in disguise?. *Medical Education Online*, 28(1). https://doi.org/10.1080/10872981.2023.2181052 .

Armstrong, P. (2010). Bloom's Taxonomy. *Vanderbilt University Center for Teaching*. https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/ .

Bom, H.-S. H. (2023). Exploring the Opportunities and Challenges of ChatGPT in Academic Writing: a Roundtable Discussion. *Nuclear Medicine and Molecular Imaging*, 57(4), 165–167. https://doi.org/10.1007/s13139-023-00809-2 .

Chang, C.-Y., Chen, I.-H., & Tang, K.-Y. (2024). Roles and research trends of ChatGPT-based learning: A bibliometric analysis and systematic review. *Educational Technology & Society*, *27*(4), 471-486. https://doi.org/10.30191/ETS.202410_27(4).TP03 .

Cong-Lem, N., Soyoof, A., & Tsering, D. (2024). A Systematic Review of the Limitations and Associated Opportunities of ChatGPT. *International Journal of Human–Computer Interaction*, 1–16. https://doi.org/10.1080/10447318.2024.2344142.

Dwiyono, Y., Wahyudi N., dan Tannarong Y. (2024). Pemanfaatan Chat GPT, Canva, dan Media Pembelajaran Interaktif untuk Peningkatan Kompetensi Pendidik. *COMMUNIO: Jurnal Pengadian kepada Masyarakat*, 2(1), 14-18. https://jurnal.litnuspublisher.com/index.php/jpkm/article/view/226.

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics*. London: Sage Publications.

Hair, J.F., Black, W.C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis*. Boston: Cengage Learning.

DOI: 10.15408/ tjems.v11i2.44701

Koteluk, O., Wartecki, A., Mazurek, S., Kołodziejczak, I., & Mackiewicz, A. (2021). How Do Machines Learn? Artificial Intelligence as a New Era in Medicine. *Journal of Personalized Medicine*, 11(1), 32. https://doi.org/10.3390/jpm11010032.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage Publications.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2). https://doi.org/10.1371/journal.pdig.0000198 .

Lestari, S. (2024). The Implementation of ChatGPT-based Learning for Higher Education in Indonesia: Systematic Literature Review, *ELS Journal on Interdisciplinary Studies in Humanities*, 7(2): June, 339-347. https://doi.org/10.34050/elsjish.v7i2.35265 .

Lingard, L. (2023). Writing with ChatGPT: An Illustration of its Capacity, Limitations & Implications for Academic Writers. *Perspectives on Medical Education*, 12(1), 261–270. https://doi.org/10.5334/pme.1072.

Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences*, 13(4), 410. https://doi.org/10.3390/educsci13040410.

Marais, E., Marais-Botha, R. & Coertzen, F. (2024). Constructing an artificial- intelligence higher education environment: Guidelines for the future. In L. Wood & O. Zuber-Skerritt (Eds.), *Shaping the future of higher educa- tion: Positive and sustainable frameworks for navigating constant change*, 173–192. Helsinki University Press. https://doi.org/10.33134- HUP-25-9 .

Nabawi, I. H., Febrina, Y., Pramono, H. E., Sutiarsih, Purwaningsih, H., Andini, R., Surawan, E., Budi, R. S., Afidhan, S., Rohanah, A., & Alfian, A. W. (2022). Petunjuk Teknis Layanan ISBN Perpustakaan Nasional RI (B. N. I. (BNI) dan K. I. N. (KIN), Koordinator Pengembangan dan Pengawasan Bibliografi, Ed.). *Perpustakaan Nasional RI.* https://isbn.perpusnas.go.id/docsurat/Petunjuk%20Teknis%20Layanan%20ISBN%20-%2020230127.pdf

Ningrum, A.R., Saputra, B.A., Mahardika, Y., Sari, N.P. (2024). Analisis Penerapan ChatGPT sebagai Alat Bantu Akademik dalam Meningkatkan Efisiensi dan Kreativitas Mahasiswa. *Seminar Nasional Amikom Surakarta (Semnasa) 2024*.

O'Connor, S., & ChatGPT. (2023). Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? *Nurse Education in Practice*, 66, 103537. https://doi.org/10.1016/j.nepr.2022.103537.

OpenAI. (2022, November 30). Introducing ChatGPT.

Rahman, M., Terano, H. J. R., Rahman, N., Salamzadeh, A., & Rahaman, S. (2023). ChatGPT and Academic Research: A Review and Recommendations Based on Practical Examples. *Journal of Education, Management and Development Studies*, 3(1), 1–12. https://doi.org/10.52631/jemds.v3i1.175.

DOI: 10.15408/ tjems.v11i2.44701

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003.

Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare*, 11(6), 887. https://doi.org/10.3390/healthcare11060887.

Suárez, A., García, V.D-F., Algar J., Sánchez, M.G., de Pedro M.L., Freire, Y. (2023). Unveiling theChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. *International Endodontic Journal*, 57, 108–113. https://doi.org/10.1111/iej.13985.

Suharmawan, W. Pemanfatan Chat GPT dalam Dunia Pendidikan. *Education Journal: Journal Education Research and Development*, 7(2), Agustus 2023. https://doi.org/10.31537/ej.v7i2.1248.

Teixeira da Silva, J. A. (2023). Is ChatGPT a valid author?. *Nurse Education in Practice*, 68. https://doi.org/10.1016/j.nepr.2023.103600 .

Tu, Y.-F. (2024). Roles and functionalities of ChatGPT for students with different growth mindsets: Findings of drawing analysis. *Educational Technology & Society*, *27*(1), 198-214. https://doi.org/10.30191/ETS.202401_27(1).TP01.

Yu, H. (2024). The application and challenges of ChatGPT in educational transformation: New demands for teachers' roles. *Heliyon*, 10(2), e24289. https://doi.org/10.1016/j.heliyon.2024.e24289.

DOI: 10.15408/ tjems.v11i2.44701