International Journal of Science Education and Teaching
Vol. 3, No. 3, pp. 126 - 138, Sep. – Dec. 2024
http://doi.org/10.14456/ijset.2024.11

Comparison of ChatGPT and Gemini AI in Answering Higher-Order Thinking Skill Biology Questions: Accuracy and Evaluation

Thoriqi Firdaus[1*], Siti Aminatus Sholeha[2], Miftahul Jannah[2], and Andre Ramdani Setiawan[2]
[1]*Natural Science Education, Universitas Negeri Yogyakarya, Indonesia*
[2]*Natural Science Education, Universitas Trunojoyo Madura, Indonesia*
**Email: thoriqifirdaus.2023@student.uny.ac.id*

Received: 25 November 2024; Revised: 30 December 2024; Accepted: 30 December 2024

Abstract. AI is becoming increasingly prevalent and advancing over time, yet the accuracy of this intelligent technology remains a subject of scrutiny. This study aims to provide an in-depth evaluation of the capabilities of two platforms, ChatGPT and Gemini AI, by analyzing and comparing their performance, assessing answer accuracy, and offering comprehensive recommendations. A quantitative comparative approach was employed to evaluate the performance of ChatGPT and Gemini AI in answering Higher-Order Thinking Skills (HOTS) questions. The questions utilized were HOTS-based items on the subject of biology. The analysis shows that ChatGPT's accuracy rate (55%) is slightly higher than Gemini AI's (50%). However, Gemini AI's average score (0.5) is higher than ChatGPT's (0.4), meaning Gemini AI gives overall more accurate answers, even though its percentage of correct responses is lower. This difference is likely due to the types of questions and specific cognitive aspects involved. ChatGPT demonstrated strengths in questions requiring analysis and evaluation, while Gemini performed better in creation-based questions. Both systems faced challenges with questions that integrated complex cognitive processes and procedural knowledge, highlighting opportunities for further improvement in their respective knowledge-processing algorithms. The standard deviations for ChatGPT and Gemini are nearly identical, at 0.5026 and 0.5130, respectively, indicating a comparable level of consistency in the responses of both models. The mean standard error for ChatGPT (0.1124) is slightly lower than that of Gemini (0.1147), suggesting that ChatGPT's mean estimates are marginally more stable. This study highlights that ChatGPT and Gemini AI exhibit distinct strengths and weaknesses in answering Higher-Order Thinking Skills (HOTS) questions. ChatGPT excelled in cognitive dimensions involving analysis (C4) and factual knowledge, providing detailed and comprehensive answers. In contrast, Gemini AI demonstrated an advantage in the creation dimension (C6) and tasks requiring concise, straightforward responses, such as producing or planning solutions.

Keywords: comparison, ChatGPT, Gemini AI, HOTs, Biology

INTRODUCTION

Technological advancements have transformed educational paradigms, driving a shift towards the digitalization of learning (Firdaus, 2023). Among these advancements, Artificial Intelligence (AI) has emerged as a key tool for streamlining contemporary learning processes (Nirwani & Priyanto, 2024). AI's ability to provide rapid responses and analyze large amounts of data has made it increasingly prevalent in education. However, concerns about the accuracy and reliability of AI-generated answers persist, especially in educational contexts where incorrect information could have significant implications (Johnson et al., 2023).

The integration of AI in education aligns with 21st-century learning goals, particularly in fostering Higher-Order Thinking Skills (HOTS). These skills, which include analysis, evaluation, and creation, are essential for student success in the Society 5.0 era, where cognitive abilities serve as a critical benchmark (Firdaus, 2022; Latifah & Maryani, 2021). HOTS goes beyond memorization and comprehension, requiring students to apply knowledge creatively and solve complex problems. As such, the effectiveness of AI platforms in supporting these cognitive processes has become a critical area of investigation.

Among the most widely used AI platforms in education are ChatGPT by OpenAI and Gemini AI by Google. Each platform has distinct strengths and applications. ChatGPT excels in speed, text processing, and efficiency, while Gemini AI is better suited for tasks that require deep analysis and comprehension (Rane et al., 2024). Research by Bahil et al. (2024) found that Gemini AI outperformed ChatGPT in terms of accuracy, achieving a rate of 66% compared to ChatGPT's 62%. However, the suitability of these platforms for addressing HOTS questions requires further investigation, particularly in understanding how their differences affect learning outcomes.

The integration of AI into educational contexts also presents challenges. A World Bank report (2024) identified ChatGPT as the most visited AI platform with 2.3 billion visits, far surpassing other platforms such as Gemini AI (123 million visits). Despite their widespread adoption, these platforms often produce answers that may seem accurate but can lead to misconceptions if not critically evaluated (Dalalah & Dalalah, 2023; Karatas et al., 2024). This highlights the need for teacher oversight and robust mechanisms to monitor the accuracy of AI-generated responses.

Thinking is a process that involves the brain's cognitive system and emotions in addressing and resolving problems (Firdaus et al., 2022). In Indonesia, the K13 curriculum prioritizes the development of HOTS as a key learning objective, emphasizing students' abilities to analyze, evaluate, and create (Asrafil et al., 2020). However, the increasing reliance on AI tools in education raises concerns about their potential to undermine critical thinking skills. Misuse of AI-generated content could lead to misconceptions and negatively impact students' understanding of complex concepts (Hidayatullah et al., 2024; Owan et al., 2023). Addressing these challenges requires a comprehensive evaluation of AI platforms to determine their efficacy in supporting HOTS development.

Research indicates that AI's adoption in education has grown significantly over the past decade, with studies on "AI" and "Education" accounting for 70% of related publications since 2010 (Chen et al., 2020). While many of these studies highlight the benefits of AI, its application in solving HOTS-related questions remains underexplored. The need for educators and developers to enhance AI tools to better support learning processes is therefore paramount (Jinhe et al., 2022).

Artificial Intelligence (AI) has become a transformative force across various fields, including education, leveraging advancements such as Natural Language Processing (NLP) to facilitate communication in human languages (Obaigbena et al., 2024). NLP enables seamless interactions between humans and machines, enhancing the

accessibility and functionality of AI systems. These capabilities, combined with advances in computer vision, allow AI to recognize and interpret gestures, emotions, and facial expressions, making it a ubiquitous tool in daily life (Obaigbena et al., 2024). Despite these advantages, over-reliance on AI for scientific responses can undermine cognitive processes, especially when the accuracy of such responses is not thoroughly verified (Danry et al., 2023; Johnson et al., 2023).

In the educational sector, AI promotes active student engagement and creates an interactive learning environment. However, its use is not without challenges. Misinterpretation of AI-generated content due to a lack of critical evaluation can lead to misconceptions, negatively affecting students' critical thinking skills (Dilekli & Boyraz, 2024). This issue underscores the need for careful evaluation of AI-generated feedback to ensure its alignment with educational goals. Advanced data analytics, a feature of AI, has also sparked debates about its role in monitoring performance and generating personalized recommendations with consistency and precision (Heaven, 2020).

Two widely recognized AI platforms, ChatGPT by OpenAI and Gemini AI by Google, exemplify the diverse applications of AI in education. Research indicates that ChatGPT excels in tasks requiring contextual intelligence and reasoning, while Gemini AI is preferred for tasks necessitating extensive analysis and deep comprehension (Rane et al., 2024). Studies comparing their performance in various domains provide mixed results. For example, Carla et al. (2024) found that ChatGPT demonstrated superior analytical performance in assisting medical professionals, whereas Gemini AI faced significant limitations, particularly in complex tasks. These findings highlight the platforms' respective strengths and weaknesses, emphasizing the importance of selecting an appropriate tool based on the specific needs of the task.

The integration of AI in education aligns with the demands of 21st-century learning, particularly in fostering Higher-Order Thinking Skills (HOTS). Defined in the revised Bloom's Taxonomy, HOTS encompasses cognitive processes such as analysis, evaluation, and creation, which are critical for addressing complex problems (Jaenuddin et al., 2020; Latifah & Maryani, 2021). In Indonesia, the K13 curriculum emphasizes HOTS as a core objective to enhance students' abilities in solving, evaluating, and creating solutions (Asrafil et al., 2020). As AI becomes more integrated into classrooms, its potential to support HOTS development is increasingly evident. However, concerns remain about its accuracy and the potential for misuse, which could lead to misconceptions and undermine students' understanding (Hidayatullah et al., 2024; Owan et al., 2023).

AI systems' ability to simplify complex material has been widely recognized, with studies showing their positive impact on learning outcomes (Joseph et al., 2013). For instance, research by Wang et al. (2023) indicates that inaccuracies in AI-generated answers often stem from the specificity of the posed questions, leading to varying performance across platforms. ChatGPT demonstrates strengths in data processing and contextualizing responses but struggles with deep understanding in certain domains (Yasmar & Amalia, 2024). Conversely, Gemini AI, with its latest model improvements, has enhanced its reasoning capabilities (Muchlis & Maulida, 2024).

Given the increasing adoption of AI, a comprehensive evaluation of its efficacy in addressing HOTS questions is essential. This study aims to compare the performance of ChatGPT and Gemini AI in answering HOTS questions. By analyzing their respective strengths, weaknesses, and accuracy, this research seeks to provide insights into the suitability of these platforms for fostering higher-order cognitive processes in education.

RESEARCH OBJECTIVES

This study aims to provide an in-depth evaluation of the capabilities of two prominent artificial intelligence platforms, ChatGPT and Gemini AI, in answering Higher-Order Thinking Skills (HOTS) questions in biology. The primary objective is to analyze and evaluate the performance and accuracy of ChatGPT and Gemini AI in addressing HOTS-based biology questions. This includes a comprehensive comparison of the platforms, focusing on parameters such as accuracy levels, relevance of responses, and alignment with scientific validity criteria and educational content. The analysis seeks to uncover the distinct characteristics of each platform, including ChatGPT's strengths in text processing efficiency and Gemini AI's advantages in deep reasoning and resource integration. Additionally, the objective encompasses identifying limitations, such as potential inaccuracies or misconceptions, that could affect the platforms' effectiveness in HOTS-based learning environments. By integrating performance analysis and accuracy evaluation into a single objective, the study aims to provide a cohesive understanding of these platforms' foundational capabilities.

The second objective is to deliver comprehensive recommendations for using AI platforms in HOTS-based learning. These recommendations will address the needs of students, educators, and AI developers, guiding them in selecting and refining platforms that are more effective, accurate, and aligned with 21st-century educational requirements. Moreover, the findings will offer valuable insights to AI platform providers, enabling them to implement continuous improvements and evaluations of their technologies. By doing so, this study seeks to contribute to the optimal use of AI in education, ensuring these tools effectively enhance HOTS-oriented learning and meet the evolving demands of modern education.

METHODOLOGY

This study adopts a quantitative comparative approach to evaluate the performance of ChatGPT and Gemini AI in answering Higher-Order Thinking Skills (HOTS) questions. The quantitative comparative method is designed to numerically measure differences between two or more variables, enabling objective statistical analysis. Comparative quantitative research involves comparing two or more groups or variables to identify their differences or similarities. According to Sugiyono (2012), comparative research compares the presence of one or more variables within a single sample, in this case, using different AI platforms. The variables compared in this study are the performances of ChatGPT and Gemini AI in addressing HOTS questions.

The questions used in the study are HOTS-based biology questions adopted from Yuliani's (2017) research. These questions were selected due to their strong reliability, with an estimated coefficient of 0.93, which categorizes the instrument as highly reliable. This confirms that the measurements obtained using this instrument are dependable. The average item logit value of 0.0 indicates that the instrument can assess higher-order thinking abilities. As Bond and Fox (2013) stated, an average item logit of 0.0 represents a random value that reflects a 50:50 probability, indicating a balance between respondents' ability levels and the difficulty of the questions. If the average item logit does not reach 0.0, the instrument is generally considered less effective in accurately measuring the intended abilities.

Summary Of 197 Measured Person

	Total Score	Count	Measure	Model Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
MEAN	10.2	20.0	.03	.52	.99	.1	1.00	.1
S.D.	3.9	.3	1.01	.10	.11	.7	.27	.8
MAX.	17.0	20.0	1.84	1.07	1.33	2.4	2.82	3.3
MIN.	1.0	18.0	-3.22	.46	.55	-1.4	.14	-1.0

REAL RMSE .54 TRUE SD .86 SEPARATION 1.59 Person RELIABILITY .72
MODEL RMSE .53 TRUE SD .86 SEPARATION 1.64 Person RELIABILITY .73
S.E. OF Person MEAN = .07
Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .72

Summary Of 20 Measured Item

	Total Score	Count	Measure	Model Error	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
MEAN	100.6	196.5	.00	.16	1.01	-.1	1.00	.1
S.D.	21.0	.8	.64	.02	.08	1.1	.13	1.1
MAX.	176.0	197.0	.53	.25	1.20	2.4	1.35	2.2
MIN.	79.0	194.0	-2.50	.16	.90	-2.1	.85	-1.6

REAL RMSE .17 TRUE SD .62 SEPARATION 3.73 Item RELIABILITY .93
MODEL RMSE .16 TRUE SD .62 SEPARATION 3.81 Item RELIABILITY .94
S.E. OF Item MEAN = .15

UMEAN=.0000 USCALE=1.0000
Item RAW SCORE-TO-MEASURE CORRELATION = -.99
3931 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 4572.60 with 3715 d.f. p=.0000
Global Root-Mean-Square Residual (excluding extreme scores): .4472
Capped Binomial Deviance = .2526 for 3931.0 dichotomous observations

Figure 1. Realibility HOTs Question
Source: Yuliani's (2017) research

The accuracy data analysis was conducted using a t-test statistic, employing the following calculation formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Meanwhile, the evaluation analysis was performed using descriptive analysis and presented in a data observation table. The results of ChatGPT and Gemini AI responses were compared based on the accuracy of their answers.

RESULTS AND DISCUSSION

Analyzing and comparing the capabilities of ChatGPT and Gemini AI

The research outcomes were analyzed to evaluate the performance of two systems, ChatGPT and Gemini AI, in answering questions designed based on cognitive dimensions, cognitive processes, and types of knowledge. This analysis focused on identifying the accuracy levels of both systems' responses compared to the correct answers and exploring performance patterns relative to the characteristics of the questions. The results provide insights into the strengths and weaknesses of each system in handling various levels of complexity and types of knowledge assessed.

Table 1: Comparison of Responses: ChatGPT vs. Gemini

Code	Cognitive Dimension	Cognitive Process	Knowledge Dimension	ChatGPT Answer	Gemini Answer	Correct Answer
A1	C4 (Analyzing)	Attributing	Factual	E	E	E
A2	C4 (Analyzing)	Organizing	Procedural	D	D	D
A3	C4 (Analyzing)	Organizing	Procedural	D	D	A
A4	C5 (Evaluating)	Examining	Factual	A	C	C
A5	C4 (Analyzing)	Distinguishing	Procedural	E	A	A
A6	C4 (Analyzing)	Organizing	Conceptual	B	B	C
A7	C5 (Evaluating)	Critiquing	Conceptual	B	B	B
A8	C6 (Creating)	Producing	Metacognitive	E	C	A
A9	C5 (Evaluating)	Critiquing	Factual	D	A	D
A10	C6 (Creating)	Producing	Conceptual	C	C	C
A11	C4 (Analyzing)	Distinguishing	Factual	A	E	D
A12	C6 (Creating)	Formulating	Metacognitive	B	B	B
A13	C6 (Creating)	Planning	Factual	A	A	E
A14	C6 (Creating)	Planning	Conceptual	B	B	D
A15	C5 (Evaluating)	Critiquing	Metacognitive	B	B	B
A16	C6 (Creating)	Producing	Conceptual	B	E	E
A17	C6 (Creating)	Planning	Metacognitive	B	B	B
A18	C6 (Creating)	Formulating	Procedural	A	A	C
A19	C5 (Evaluating)	Critiquing	Conceptual	B	B	A
A20	C4 (Analyzing)	Attributing	Factual	C	C	E

The analysis results indicate that ChatGPT demonstrates a slightly higher accuracy rate than Gemini, with 55% correct answers versus 50% for Gemini. Although the margin is small, ChatGPT has an advantage in understanding and responding to certain questions. Regarding cognitive dimensions, ChatGPT outperformed Gemini in tasks requiring analysis skills (C4), answering 4 out of 8 questions correctly, compared to Gemini's three correct answers. Both systems performed equally well for the evaluation dimension (C5), each providing four correct answers out of 6 questions. However, in the creation dimension (C6), Gemini showed a slight edge, answering 4 out of 6 questions correctly, while ChatGPT managed three correct answers.

Regarding cognitive processes, ChatGPT excelled in questions involving organizing processes, with two correct answers compared to Gemini's 1. However, Gemini demonstrated greater consistency in tasks requiring producing and planning processes, highlighting its ability to handle more complex question types. ChatGPT, on the other hand, struggled more with questions requiring distinguishing or examining, though its overall performance remained satisfactory.

From the perspective of knowledge dimensions, ChatGPT showed superior performance in questions based on factual knowledge. Gemini, meanwhile, delivered nearly comparable results to ChatGPT on questions involving conceptual and metacognitive knowledge. However, both systems faced significant challenges with procedural knowledge-based questions, with higher error rates observed in this dimension. Certain questions, such as A3, A6, A13, and A18, presented a high difficulty level for both systems. On these questions, the responses from both ChatGPT and Gemini diverged from the correct answers, indicating that the combination of complex cognitive processes and procedural knowledge posed significant challenges. Questions A3, A6, A13, and A18 incorporate visual elements such as images and diagrams, which significantly increase the complexity for AI systems to generate accurate responses. Specifically, image-based questions (A3, A6, and A18) and diagram-based questions (A13) require the AI to interpret

visual data, a capability that remains a notable limitation for most natural language processing (NLP) models.

In addition to their visual components, these questions engage cognitive processes classified under Bloom's Taxonomy as C4 (analyzing) and C6 (creating). AI systems face challenges with C4 tasks because they require synthesizing information from multiple modalities, including textual explanations, visual data, and contextual knowledge. For example, A3 and A6 necessitate the integration of image interpretation with biological concepts, a task that exceeds the current capabilities of text-based AI models like ChatGPT and Gemini AI. Similarly, questions involving C6 processes, such as A13 and A18, pose difficulties because they demand the generation of novel and creative outputs. These questions often require the AI to not only interpret diagrams but also propose original solutions or construct new concepts based on limited or incomplete information. This highlights a critical gap in the ability of current AI systems to perform tasks that simulate higher-order cognitive skills, especially those requiring creativity and deep reasoning.

Understanding why these questions are challenging provides valuable insights for improving AI systems. Enhancements such as multimodal training, which integrates textual and visual data processing, or more robust algorithms for handling abstract and creative reasoning, could address these limitations. Future AI training models should focus on bridging these gaps to improve performance in tasks that combine visual interpretation and higher-order cognitive skills. Conversely, questions like A1, A2, A7, A10, A12, A15, and A17 demonstrated that both systems could consistently provide correct answers. These questions generally involved factual or conceptual knowledge paired with relatively straightforward cognitive processes.

Evaluating the accuracy of responses

A comparison between ChatGPT and Gemini AI in answering HOTS biology questions requires further analysis. This analysis aims to evaluate and compare the performance of the two models based on group statistical results. The data provided includes sample size (N), mean score (Mean), standard deviation (Std. Deviation), and standard error of the mean (Std. Error Mean), as presented in Table 2. Using this data, we can assess the consistency of each model's responses and determine whether there are significant differences between the two models.

Table 2: Statistical Comparison Results

Group Statistics	N	Mean	Std. Deviation	Std. Error Mean
ChatGPT	20	0.4	0.5026246899500346	0.11239029738980327
Gemini	20	0.5	0.512989176042577	0.11470786693528086

The statistical analysis indicates that ChatGPT and Gemini's performance in answering questions exhibits nearly equivalent characteristics. The comparison is conducted on a balanced dataset based on an identical sample size of 20 for each model. The average score (mean) for Gemini was slightly higher than that of ChatGPT, at 0.5 compared to 0.4. This suggests that Gemini, on average, provides somewhat more accurate responses than ChatGPT. However, the small mean difference of 0.1 necessitates further statistical testing to determine its significance. Regarding performance variation, the standard deviations for ChatGPT and Gemini were nearly identical, at 0.5026 and 0.5130, respectively. This reflects that both models exhibit similar levels of consistency in answering questions. The comparable variation indicates that ChatGPT and Gemini demonstrate similar fluctuations in performance on the tested data. Furthermore, the standard error of the mean for ChatGPT was slightly smaller than that for Gemini, at 0.1124 versus 0.1147. ChatGPT's mean score estimation is marginally more stable than

Gemini's. Although Gemini showed a higher average score, the levels of variation and consistency between the two models are nearly identical. The small difference in average scores suggests that the performance of the two models in answering questions is not significantly different. Advanced statistical analysis, such as a significance test, must confirm whether this difference is statistically meaningful or simply due to random variability in the data.

t-Test Analysis

A t-test is necessary to determine the statistical significance of the differences between the two groups, ChatGPT and Gemini, regarding their average scores. The table includes key metrics from the t-test analysis, such as:

- t-Statistic
- Degrees of Freedom (df)
- Two-tailed Significance (Sig. 2-tailed)
- Mean Difference
- Standard Error of Difference (Std. Error Difference)

By evaluating these values, the analysis will establish whether the observed mean difference between ChatGPT and Gemini is statistically significant or merely attributable to chance variations within the dataset.

Table 3: Uji T ChatGPT dan Gemini dalam Menjawab Soal

t-statistic	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
-1.0000002	19	0.3298768009211	-0.0999999999998	0.09999999999998

The t-test results reveal a t-statistic value of -1.0 with degrees of freedom (df) equal to 19. This indicates a minor difference between the mean scores of ChatGPT and Gemini. The significance value (Sig. 2-tailed) is 0.330, which exceeds the standard significance threshold of 0.05. This means the observed difference in mean scores between the two groups is not statistically significant. The mean difference between the two groups is -0.1, suggesting that ChatGPT's average score is slightly lower than Gemini's. However, with a standard error of difference of 0.1, this discrepancy is too small to be considered meaningful. The negative value in the mean difference simply indicates the direction of the difference ChatGPT scoring lower than Gemini but does not imply any substantive or significant disparity. This analysis concludes that although there is a slight difference in mean scores between ChatGPT and Gemini, this difference is not statistically significant. In other words, the performance of both groups can be considered equivalent in the context of this test. Further research with a larger sample size or an alternative test design may be necessary to explore performance differences between these models further.

Evaluation of ChatGPT and Gemini AI Responses

Based on the analysis of AI usage (ChatGPT and Gemini AI) in solving higher-order thinking Skill (HOTS) questions, here is a detailed narrative based on cognitive dimensions, answer accuracy, explanation quality, consistency, as well as the strengths and weaknesses of each AI. The analyzed questions involve cognitive dimensions such as analyzing (C4), evaluating (C5), and creating (C6) and include knowledge dimensions spanning factual, conceptual, procedural, and metacognitive domains. Both AIs could provide responses aligned with the required cognitive levels. However, variations were observed in the depth of explanation, particularly for questions requiring extensive exploration. Regarding answer accuracy, both AIs generally provided answers that matched the correct key, although differences in delivery were noted. For example, in Question 1, ChatGPT and Gemini AI accurately explained moss's role in the ecosystem.

In Question 2, both provided correct answers regarding the functions of antipodal and synergid cells, but ChatGPT's explanation was more detailed than Gemini AI's.

Regarding explanation quality, ChatGPT delivered structured and in-depth explanations. For example, besides answering the question, ChatGPT elaborated on related processes or mechanisms, such as the ecosystem impact of moss on the carbon cycle. In contrast, Gemini AI provided concise and direct answers, focusing on the main points. While these answers were relevant to the questions, Gemini AI often lacked the deep elaboration necessary for a more comprehensive understanding. Regarding response consistency, ChatGPT provided more consistent, detailed answers to high-difficulty questions. At the same time, Gemini AI tended to perform optimally on questions based on factual or simple conceptual knowledge. For metacognitive questions, Gemini AI showed weaknesses in delivering exploratory responses.

As for strengths and weaknesses, ChatGPT's strength lies in its detailed and comprehensive explanations, making it suitable for fostering deep understanding. However, its lengthy responses can make it difficult for users to pinpoint the main answers quickly. Gemini AI's strength is providing concise and focused answers, ideal for tasks requiring straightforward responses. However, it lacks the depth for questions requiring extensive exploration and elaboration.

ChatGPT and Gemini AI were assessed in solving higher-order thinking Skill (HOTS) questions across various categories, focusing on accuracy, explanation quality, and analytical capabilities.

- **Question Code A1:** Analyzing the role of moss in life. ChatGPT exhibited strong analytical skills, detailing the benefits of moss, such as preventing erosion, retaining water, and aiding soil formation. It also provided relevant ecological context. In contrast, Gemini AI only gave a core response, stating the importance of moss in ecosystems without elaboration.
- **Question Code A2:** Functions of antipodal cells and synergids. ChatGPT provided an in-depth explanation, describing their location and role in double fertilization. It mentioned antipodal cells as nutrient providers and synergids as chemical signalers guiding pollen tubes to the ovum. Gemini AI’s response was correct but concise, lacking a detailed mechanism.
- **Question Code A3:** Identifying ovule location in a Gymnosperm reproduction diagram. ChatGPT identified the correct location and explained the ovule as the site of sperm and ovum fusion, enhancing biological understanding. Gemini AI gave the correct location but without context or explanation.
- **Question Code A4:** Evaluating errors in moss metagenesis. ChatGPT elaborated on the stages, explaining that the zygote develops into sporogonium before producing spores and detailing the functions of archegonia and antheridia. Gemini AI answered correctly but without a comprehensive explanation.
- **Question Code A5:** Optimal temperature for moss growth based on a graph. ChatGPT provided a comprehensive response, explaining how 22–28°C supports moss metabolism and how extreme temperatures reduce physiological activity. Gemini AI only stated the optimal temperature without linking it to biological mechanisms.
- **Question Code A6:** Identifying errors in statements about Gymnosperms and Angiosperms. ChatGPT analyzed differences in seed protection, reproductive tools, and seed structure, aiding conceptual understanding. Gemini AI only mentioned core differences without further discussion.
- **Question Code A7:** Determining corn plant characteristics based on an image. ChatGPT detailed monocot features, such as fibrous roots, parallel leaves, and separate flowers, providing rich context. Gemini AI answered correctly but omitted specifics.

- **Question Code A8:** Identifying a fruit not classified as a Dicot. ChatGPT correctly identified coconut as a Monocot and melinjo as a Gymnosperm, explaining scientific classifications. Gemini AI provided the correct answer but lacked elaboration.
- **Question Code A9:** Identifying a fruit not classified as a Dicot. ChatGPT accurately described the coconut as a Monocot and melinjo as a Gymnosperm, including Monocot characteristics and Gymnosperm seed properties. Gemini AI's answer was correct but lacked depth.
- **Question Code A10:** Designing an experiment to distinguish Monocots from Dicots using fruits. ChatGPT outlined a structured procedure, including seed observation, to identify cotyledons. The explanation connected methods to outcomes. Gemini AI's response was correct but lacked logical reasoning and procedural details.
- **Question Code A11:** Determining if a flower is complete and perfect. ChatGPT explained the criteria for completeness (presence of peduncle, receptacle, petals, stamens, and pistils) and perfection (both reproductive organs). Gemini AI stated the result without elaborating on the supporting parts.
- **Question Code A12:** Identifying ferns based on characteristics like segmented stems and small spiral leaves. ChatGPT identified the plant as *Equisetum sp.*, adding habitat details, such as moist mountainous areas. Gemini AI provided the species name without additional context.
- **Question Code A13:** Explaining differences between tropophyll (sterile) and sporophyll (fertile) leaves. ChatGPT linked their functions to photosynthesis and spore production, integrating their roles in the fern life cycle. Gemini AI provided the correct answer but lacked depth.
- **Question Code A14:** Classifying and identifying benefits of Pteridophytes. ChatGPT included classifications such as *Adiantum* (ornamental), *Lycopodium* (herbal medicine), and *Azolla* (green fertilizer), connecting benefits to plant classes. Gemini AI listed classifications without discussing uses.
- **Question Code A15:** Explaining haploid and diploid stages in Pteridophyte metagenesis. ChatGPT explained that the gametophyte arises from spores via meiosis (haploid), while the sporophyte arises from a zygote (diploid), linking meiosis to the plant life cycle. Gemini AI's response, though accurate, lacked detailed connections.
- **Question Code A16:** Outlining a practical experiment to observe *Azolla pinnata* and *Sphagnum sp.* using eosin solution. ChatGPT gave a detailed procedure, including preparation, observation duration, and analysis. Gemini AI provided a brief response without technical specifics.
- **Question Code A17:** Identifying tools for anatomical observation of *Sphagnum sp.* ChatGPT listed tools like microscopes, slides, pipettes, tweezers, and cutters, explaining each tool's purpose. Gemini AI mentioned only basic tools, omitting functional details.
- **Question Code A18:** Formulating a hypothesis for eosin absorption experiments. ChatGPT proposed a theory based on physiological differences between *Azolla* and *Sphagnum*, such as vascular tissue efficiency, supported by biological reasoning. Gemini AI offered a simpler hypothesis without elaboration.
- **Question Code A19:** Evaluating incorrect statements about moss and fern morphology. ChatGPT explained the gametophyte dominance in mosses and sporophyte dominance in ferns, highlighting additional distinctions like true roots, stems, and leaves. Gemini AI only identified generational dominance differences.
- **Question Code A20:** Determining germination types in Monocots and Dicots. ChatGPT provided examples of epigeal germination (cotyledons above ground) and hypogeal germination (cotyledons underground). Gemini AI answered correctly but omitted comparative details.

Policy Recommendations

In advancing higher-order thinking skills (HOTS) based learning, the indispensable role of teachers remains at the forefront, even amidst the growing integration of AI tools such as ChatGPT and Gemini AI. While these tools hold considerable potential to enrich the educational process, their efficacy hinges on deliberate and well-structured implementation strategies. Teachers must be active facilitators and validators with actionable methodologies to incorporate AI into lesson planning, assessment, and classroom management.

Teachers can use ChatGPT to craft analytical scenarios (C4) or in-depth evaluative materials during lesson planning. At the same time, Gemini AI is well-suited for designing creative tasks that involve planning and production (C6). In assessment, teachers play a pivotal role in validating AI-generated responses to ensure accuracy and alignment with learning objectives. For instance, ChatGPT's comprehensive and detailed answers can serve as a foundation for facilitating in-depth classroom discussions encouraging critical and analytical thinking. Conversely, Gemini AI's concise responses can be utilized as comparative tools for exploring alternative solutions or initiating talks.

AI can also enhance classroom management when strategically integrated into learning activities. ChatGPT excels in providing elaborate explanations to help students grasp complex concepts, whereas Gemini AI is ideal for quick, formative assessments like quizzes. However, direct teacher intervention is crucial for tasks that necessitate intricate cognitive processes or procedural knowledge. Teachers can guide students through complex problem-solving steps, ensuring they understand the logical connections between each phase and the expected outcomes.

To maximize the utility of AI, teachers require a structured framework for validating AI-generated responses. This framework should include aligning AI outputs with curricular standards, posing follow-up questions to deepen students' comprehension, and engaging students in critically evaluating AI-provided answers as an exercise in analytical reasoning. Such measures enhance learning accuracy and reinforce students' critical, creative, and problem-solving skills.

Consequently, policies governing the use of AI in education must unequivocally establish teachers as the primary arbiters of these tools. Teachers must ensure that AI serves as a complement to, rather than a replacement for, their pedagogical expertise. By maintaining this balance, AI technologies such as ChatGPT and Gemini AI can become powerful instruments in creating an effective, meaningful, student-centered learning environment that fosters enduring understanding.

CONCLUSION

This study demonstrates that ChatGPT and Gemini AI exhibit distinct strengths and weaknesses in addressing questions about higher-order thinking skills (HOTS). ChatGPT excels in cognitive dimensions involving analysis (C4) and factual knowledge, offering detailed and comprehensive responses. In contrast, Gemini AI performs better in creation (C6) tasks and processes requiring concise, direct answers, such as production or planning. However, both AIs encounter challenges with procedural knowledge-based questions and complex cognitive processes. The average accuracy difference between the two is not statistically significant, indicating that their overall performance is relatively comparable, with each AI excelling in different areas. The consistency and quality of their responses vary depending on the complexity of the questions and the type of knowledge being assessed. ChatGPT and Gemini AI can significantly contribute to learning, particularly in HOTS-based questions. Nevertheless, it is essential to emphasize that the successful implementation of AI in education depends heavily on how teachers utilize these technologies to enhance learning rather than as substitutes for their roles in guiding and supporting students.

IMPLICATIONS

ChatGPT, with its ability to provide detailed and comprehensive responses, is well-suited for supporting analytical and evaluative tasks that require in-depth exploration. Its more elaborate answers can serve as a foundation for class discussions, encouraging students to think critically and develop a deeper understanding of concepts. On the other hand, Gemini AI, with its concise and direct answers, is ideal for quick learning activities such as quizzes or tasks requiring time efficiency. These differences highlight the potential for both AIs to be used complementarily to address diverse learning needs.

However, the study also identifies limitations in both AIs, particularly with questions involving procedural knowledge and more complex cognitive processes. These challenges underscore the critical role of teachers in validating AI-generated answers and providing direct guidance to students. Teachers are responsible for ensuring that students receive correct answers and grasp the underlying reasoning. In this context, AI responses can serve as starting points for discussion or as tools to explain concepts, but teachers remain the primary agents in fostering students' understanding.

Furthermore, integrating AI into learning must be strategically designed to maximize its effectiveness. ChatGPT can support tasks requiring in-depth elaboration, while Gemini AI can be utilized in simpler, more straightforward contexts. The use of AI should complement, not replace, the essential human interactions that are central to the educational process. With the right approach, these technologies can enhance learning quality, help students develop higher-order thinking skills, and support teachers in creating richer and more diverse educational experiences.

ACKNOWLEDGEMENTS

We extend our gratitude to KOPI ALINEA (Komunitas Peneliti Akademi Literasi Sains dan Budaya) for facilitating this research process. The support provided, including access to discussions, references, and feedback, has been instrumental in developing ideas and completing this study. The presence of KOPI ALINEA as a space for collaboration and learning has been a vital part of this research journey. We hope this community continues to grow and provide valuable benefits to other researchers in the future.

REFERENCES

Asrafil., Retnawati, h., & Retnowati, E. (2020). The Difficult of Students when Solving HOTS Problem and the Description of Students Cognitive Load After Given Worked Example as a Feedback. *International Conference on Science Education and Technology*. 1511(1). 1-11.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Carla, M. M., et al. (2024). Large Language Models as Assistance for Glaucoma Surgical Cases: a ChatGpt VS Google Gemini Comparison. *Graefe's Archive for Clinical and Experimental Ophthalmology.* 2945-2959.

Chen, L., Chen, P., & Lin, Z. (2020). Artificial Intelegence in Education: A Rivew. *Ieee Access.* (8). 75264-75278.

Dalalah, D., & Dalalah, O. M. (2023). The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. *The International Journal of Management Education*, *21*(2), 100822.

Danry, V., et al. (2023). Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Qestions Improve Human Logical Discernment Accuracy Over Causal AI Explanations. *Proceedings of the 2023 CHI Converence on Human Factors in Computing Systems.* 23(352). 1-13.

Dilekli, Y., & Boyraz, S. (2024). From “Can AI think” to “Can AI help thinking deeper?”: Is use of Chat GPT in higher education a tool of transformation or fraud?. *International Journal of Modern Education Studies.* 8(1). 49-71.

Firdaus, T. (2022). PENERAPAN MODEL DIRECT INSTRUCTION BERBASIS SETS PADA PEMBELAJARAN IPA UNTUK MENINGKATKAN KETERAMPILAN BERPIKIR KRITIS SISWA. *Natural Science Education Research (NSER)*, *5*(1), 119-134. https://doi.org/10.21107/nser.v5i1.15759

Firdaus, T., Ahied, M., Qomaria, N., Putera, D. B. R. A., & Sutarja, M. C. (2022). Jurnal Pembelajaran Sains. *Jurnal Pembelajaran Sains*, *6*(1).

Firdaus, T. (2023). Representative platform cyber metaverse terkoneksi BYOD sebagai upaya preventive urgensi digital pada sistem pendidikan Indonesia. *Jurnal Integrasi dan Harmoni Inovatif Ilmu-Ilmu Sosial*, *3*(2), 123-131. https://doi.org/10.17977/um063v3i2p123-131

Joseph, G. V., et al. (2024). Impact of Digital Literacy, Use of AI tools and Peer Collaboration on AI Assisted Learning: Perceptions of the University Students. *Digital Education Review.* 45. 43-49.

Karataş, F., Abedi, F. Y., Ozek Gunyel, F., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT’s effect on foreign language learners. *Education and Information Technologies*, 1-24.

Latifah, A., & Maryani, I. (2021). Developing HOTS Questions for the Materials of Human and Animals Respiratory Organs for Grade V of Elementary School. *Jurnal Prima Edukasia.* 9(2). 179-192.

Muchlis & Maulida, R. (2024). Komparasi Respons ChatGPT dan Gemini terhadap Command Pattern Identik dengan Metode Black Box. *Jurnal Teknik Informatika STMIK Antar Bangsa, 10*(2). 68-71. *https://ejournal.antarbangsa.ac.id*

Nirwani, N., Priyanto. (2024). Integrasi Artifical Intelegence dalam Pembelajaran Bahasa di SMP. *Jurnal Pendidikan Bahasa dan Sastra.* 7(1). 31-38.

Obaigbena, A., et al. (2024). AI and Human-Robot Interaction: A Riview of Recent Advances and Challenges. *GSC Advanced Research and Reviews.* 18(02). 321-330.

Owan, V. J., et al. (2023). Exploring the Potential of Artificial Intellegence Tools in Educational Measurement and Assessment. *Journal of Mathematics, Science and Technology Education.* 19(8). 1-15.

Rane, N. L., Choudhary, S. P., & Rane, J. (2024). Gemini or ChatGPT? Efficiency, Performance, and Adaptability of Cutting-Edge Generative Artificial Intelligence (AI) in Finance and Accounting. 1-12.

Rane, N. R., Choudhary, S.P., Rane, J. (2024). Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. *Journal of Applied Artificial Intelligence, 5*(1). 69-93. *https://doi.org/10.48185/jaai.v5i1.1052*

Sugiyono. (2012). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Bandung: Alfabeta.

Wang, Y., Shen, S., & Lim, B. (2023). RePrompt: Automatic Prompt Editing to Refine AI Generative Art Towards Precise Expressions. ArXiv, abs/2302.09466. *https://doi.org/10.1145/3544548.3581402*

World Bank. (2024). *Who on Earth Is Using Generative AI?*

Yasmar, R., & Amalia, D. R. (Analisis SWOT Penggunaan Chat GPT dalam Dunia Pendidikan Islam. *Fitrah Jurnal Studi Pendidikan*, 15(1). https://doi.org/10.47625/fitrah.v15i1.668

Yuliani, E. (2017). Pengembangan Manual Test berbasis Higher Order Thinking Skill (HOTs) Serta Implementasinya di SMA Unggul Negeri 8 Palembang. (Disestation, Universitas Islam Negeri (UIN) Raden Fatah Palembang).