ChartGPT-3.5's Performance on Introduction to Medical Laboratory Sciences Test: AI Triumphs in Theory but Stumbles in Diagrams

Parks Makhoahle[1], Xolani Khohliso[2]
[1]Centre for Quality Health and Living, Faculty of Health and Environmental Sciences, Central University of Technology, Free State, South Africa
[2]Curriculum and Academic Staff Development, Centre for Innovation in Learning and Teaching, Central University of Technology, Free State, South Africa
Corresponding Author: xkhohliso@cut.ac.za

Abstract

The inclusion of Artificial Intelligence (AI) in health education has seen a sizable uptake among universities and medical students, with a growing focus on enhancing AI tools to facilitate learning and decision-making. This study evaluates the performance of AI on a formal test introducing medical science concepts. A 90-mark test, written by 60 students, was sent to the ChatGPT3.5 window for answering. AI completed 5 questions, obtaining 72.2% (65/90) on the theory part, leaving 3 diagram questions (27.8% [25/90]) with reasons. ChatGPT3.5 achieved perfect scores on theory questions; however, it struggled to interpret questions with diagrams. Thus, this study explores the implications of AI in medical education and the areas that require further development. A more advanced ChatGPT could have resulted in a better performance if it were upgraded to analyse the diagram. In essence, while AI contributes valuable insights and efficiency in certain domains of teaching and learning, it should be viewed as a complementary tool rather than a replacement for the irreplaceable human element. The relationship between AI and human expertise holds the potential to enhance teaching and learning in healthcare settings, combining the strengths of both for more comprehensive and reliable patient care, which is necessary for the development of students' competencies. ChatGPT-4.5 generally has higher cost implications compared to ChatGPT-4 and ChatGPT-3.5 due to its advanced capabilities and improved performance. In some countries and institutions with limited financial resources, it's important to leverage free access to ChatGPT for educational purposes, while being mindful that advanced versions may incur costs. Integrating ChatGPT as a teaching tool may require financial investment to access higher-tier services and ensure effective implementation.

Keywords: ChatGPT3.5, Artificial intelligence, Questions, Medical Laboratory Sciences

INTRODUCTION

The prediction of the machine that will be trained and work to imitate humans happened a long time ago. Gardner et al. (2021) reported this dream by Page (1966) and elaborated on this phenomenal idea, asserting that the machine will have a significant influence on global education and students' empowerment if trained on that specific content. To some extent, ChatGPT emerged as one of the influential artificial intelligences (AI) making ancient predictions a reality. OpenAI changed the technology landscape by developing Chat Generative Pre-Trained Transformer (ChatGPT), which is a natural language processing model launched in November 2022 (OpenAI, 2023). ChatGPT reviewed studies identified five key uses in teaching and assessment, extending from creating course content to performing language translation in the different languages (Lo., 2023). Lo (2023) also proposes that students can use it while preparing for formal formative assignments and preparation for summative assessments.

Students can draft papers in preparation for assessments and have ChatGPT assist with responses so that they can evaluate their own errors, and that will empower the students to be ready for any form

assessment. Lo (2023) found that the chatbot can act as a useful scaffolding tool where lecturers as instructors can use it to create formal formative assessment tasks and provide individualise student performance evaluation. Another reviewed study by Crompton and Burke's (2023) indicated that assessment and evaluation emerged as the most common use of AI technologies in higher education, with more emphasis on automatic assessment, test generation, feedback, online activity review, and the evaluation of educational resources. The use of artificial intelligence (AI) among medical professionals and students has gained interest in recent years. ChatGPT has gained interest in the healthcare field, owing to its high performance by answering almost 60% in the United State Medical Licensing Exam (USMLE) (Kung et al.,2023; Choi et al., 2023, ChatGPT., 2023).

Cooling the growing concern of the anti-ChatGPT groups, Halaweh (2023) highlights that acceptance of the existing tools like search engines and spreadsheets that are used to aid with searching for information, calculations, and organizing data without concern, then ChatGPT should be accepted in good spirit as it will help people during this technology driven world. Halaweh (2023) suggests that as there are no concerns with using these tools, there should be no concern with using ChatGPT's abilities to produce and edit texts by considering it as a tool to save time and effort. It is not debatable that issues such as the ethicality of using ChatGPT due to its capabilities of producing texts quickly and can cause ethical issues when used wrongly by the authors. Another study reported that ChatGPT was found to have generated articles that received positive peer-review process because all three articles were commended for publication (Dowling and Lucey., 2023). The unethical escape happens when authors who used AI claimed this work to be theirs. Improper use may pose some issues for users as well. Claiming falsified biased data, not having up-to-date information, and generating incorrect or fake information may ruin the author's reputation. The consequences of not enhancing and embracing ChatGPT and its proper use will lead students to be involved in plagiarism and have them bypass plagiarism detectors (Lo, 2023, p. 8). Mhlanga (2023) reiterates the use of AI responsibly and ethically in education, highlighting empowering students about AI limitations, transparency in the use of ChatGPT, respect for privacy, and accuracy of information.

Building on the above context, the present study aimed to evaluate ChatGPT-3.5's performance on an introductory medical laboratory science test by directly comparing its answers to those of first-year students. The important objective was to discover how closely ChatGPT – 3.5's answers aligned with human responses in terms of accuracy, and to identify specific areas, specifically diagram-based questions, where the AI struggled. This investigation contributes to the debate on AI integration in higher education in South Africa, particularly relevant in the era of National Health Insurance (NHI) implementation, by examining whether such AI tools can enhance learning outcomes without compromising the essential role of human expertise.

LITERATURE REVIEW

Olney et al. (2017) and Yang et al. (2021) reported the reliability of AI in teaching and learning and in the generation of tests. As has been suggested by Lo (2023), subjects' facilitators can benefit from using ChatGPT as it can assist in drafting course syllabi, teaching resources, and assessment tasks as long as the final touch to ensure accuracy of the generated content is addressed. Al-Worafi et al. (2023) tested the feasibility of ChatGPT in the curriculum design and found improved expect rating from 50%-92%, evidence suggesting that AI can be a useful tool. ChatGPT-3, freely available was used to assess the quality of the students generated short answers, Moore et al. (2022), although the focus was on the extent to which students are able to generate quality test items, the ChatGPT3 results indicated that the free ChatGPT-3 was useful in evaluating and assessing the content of students' work**.** In support of the notion which emphasizes the need for human opinion as the final word on AI generated output, ChatGPT-3 in a study matched human evaluation only for 40% of the questions as compared to human expect classifying 68% of questions as low quality. It's evident that the AI model reported most questions as high quality, For GPT-3 this figure was only 9%. The researchers conclude that ChatGPT-3 overestimated the quality of the questions. As every assessment must be aligned to cognitive levels

of understanding, in assigning the items to the levels of Bloom's taxonomy**,** there was a disagreement between ChatGPT-3 and human experts in 68% of the questions. Moore et al. (2023) continued testing the new model; at that time, GPT-4 identified 79% of the flaws identified by human annotators and matched 62% of the human quality evaluations**.** This may indicate that as language models become more advanced, they can perform pedagogical tasks more effectively, approximating expert performance.

AI enhancement and its use in learning also for students' usage as a tool to prepare tasks online such as automatic assessment and question correction purposes get global popularity. The is growing global interest in the use features which makes it easier for lecturers and subject instructors (SI's) to prepare and conduct quizzes and tests practically and easily. Lectures and SI's no longer need to make questions and corrections manually. Kejarcita platform supported features to be used in the application of automatic assessment is one of the AI driven quiz creation and automatic correction (https://kejarcita.id/). This made a development in the AI period as it granted lectures to easily and practically create online assessments. Kejarcita platform is user friendly as it requires the subject details, academic level, number of questions to be used, and bloom taxonomy cognitive level of difficulty. Following which, the lecturer only needs to share the assessment link with the students to do it directly online.

Student online assessment results can be directly accessed and accepted automatically on module by the lecturer. The results score will appear with a list of wrong questions, correct questions, and discussion. This creates a smooth learning experience as lecturers no longer have to bother manually correcting and assessing the results of online assessment. The lot is mostly handled by the AI system. AI technology is easing the burden of administrative workload, such as preparing lesson plans, assessing exams, checking student blackboard activities, and others.

Technological improvements to these processes will give lecturers enough time to monitor each student's progress, implement individualized improvement actions for each student, and focus on improving teaching techniques. The lecturer can do so because the AI system works independently according to programmed instructions, learns from each student's habits, and identifies those at risk. Then AI can provide individualized recommendations for material that each student and others need to study again, based on the results each has obtained. The AI system will work independently according to programmed instructions and can learn from the user's or student's habits. Another advantage of being technologically inclined as an institution is that AI can provide recommendations for material that needs more attention, based on its analysis of student performance. One good example is an AI-based strategy for a university's English-aided education system that is provided to assist. This benefited students as some functions of the English teaching system are enhanced and humanised when combined with English teaching (Bin & Mandal, 2019). The use of AI in English education is explored to increase the quality and efficiency of English pedagogy.

Artificial intelligence has made improvements in health, education and other disciplines regarding teaching and learning, its use for assessment (Memarian & Doleck, 2023). Nevertheless, much consideration was not given to the use of theories that directly inform technologies within education, specifically assessment for learning. ChatGPT3.5 offers interpret questions, questions and response, perform translation across several languages, text production in dissimilar demonstration forms and can analyze and summarize the data in different forms (Dwivedi et al., 2023). Several authors classified AI in education through different methods, yet none clearly showed the correlation between learning theories and assessment (Gao, Nagel, & Biedermann, 2019; Van Vaerenbergh & P′erez-Suay, 2022; Zhai et al., 2021). As the learning outcomes should be aligned to content to be taught, a literature review predicted a lack of research in addressing AI technologies embedded deeply with educational theories and assessment strategies (Chen et al. (2020).

Because of the slow response of higher institutions due to various and dynamic qualms surrounding assessment capacity, the research highlights more focus to be on the transfer of expectations,

assumptions, evidence, and outcomes (Hargreaves, 2005). Warm bodies can design assessments for learning that focus on the idea of informing learners about the assessment expectations, collecting evidence concerning the communicated expectations, and making sense of assessments done via qualitatively documented (e.g., through a rubric or memorandum) means. Adopting this approach will enable acceptance that assessment is eligible to error, whether good or bad what is important is the emphasis on what transpired during the assessment. Hence the study is interested in understanding the AI ChatGPT3.5 strength as a free tool used by less resourced countries. To do so, we let ChatGPT3.5 respond to an Assessment for learning activity so that we can identify gaps in the free AI tool.

In a low resource economic setting like in Sub Saharan Africa and South Africa including other developing countries, artificial intelligence models like ChatGPT can play a very important role in assisting teaching and learning more especially ChatGPT 3.5 because it's a free one. ChatGPT4 is available free for a certain time and have costs implications, hence on this study we focus more on ChatGPT 3.5 that is used by most people in South Africa and poor countries facing economic constrains. Despite success of ChatGPT in answering medical questions, one needs to assess its accuracy further for future adjustment or awareness to its users to identify its limitations as it being reported to have variable accuracy in certain specialty questions (Gupta et al., 2023; Suchman et al., 2023). It was reported that ChatGPT-4 can comprehend specific field questions and provide accurate answers in ophthalmology (Moshifar et al., 2023). Nonetheless low economic countries cannot afford to use ChatGPT-4 due to limited access and resort to continue with ChatGPT3.5, that is why more focus is on the free ChatGPT on this study.

ORIGINALITY OF THE WORK

To our knowledge, this study is the first to directly compare ChatGPT-3.5's performance with that of human students on a formal formative assessment in medical laboratory science. Previous research has evaluated ChatGPT on standardised exams or purely textual questions banks but our work introduces a novel element by including diagram-based questions that the AI must tackle alongside students. Through focusing on the free ChatGPT – 3.5 model in a resource - limited South African university context, this research addresses a gap in the literature regarding AI's capabilities in real classroom settings. This originality lies in highlighting how a freely accessible AI performs on domain -specific educational tasks including visual interpretation and where it falls short, providing new insights for the integration of AI tools in medical education.

METHODOLOGY

A formal formative assessment for the module named introduction to medical laboratory sciences was created in Microsoft Word. This module is presented at national qualification framework level 6 (NQFL 6). The students who are enrolled on this module are at first year level of higher institutions post passing grade 12 or exit level secondary education system in South Africa, all have passed first semester module which is a prerequisite to continue with this module second semester first year level. This group consist of 60 students enrolled for bachelor on medical laboratory Sciences in Biomedical Technology pitched at NQFL8 presented for 4 years according to South African higher education system. Table 1 below presents an overview of the participants demographics.

Table 1: Demographic characteristics of the student participants

Characteristics	Value
Age	Between 17 – 19 years
Female	52%
Male	48%
Africans	98%
Whites	2%
Academic Level	1st Years

Prerequisite module Passed	60 (100%)

These questions covered several introductory aspects to medical lab science as listed in Table 2. These questions range between 6 to 15 marks long and were written by first year's medical lab sciences students. A free account was created on the website https://beta.openai. com/overview and used to log in https://openai.com/chatgpt. This study was done using the free version of ChatGPT available at that time (ChatGPT 3.5). Once logged in the following steps were followed to answer the questions: 1.) The text and graphs of the questions were copied and pasted into the textbox, 2.) The following sentence was pasted into the textbox, exactly as written: "answer the following test written by the First-year students in Medical laboratory Sciences students." 3.) The 90 marks questions were then copied, pasted into ChatGPT3.5 inbox. A total of 8 questions were passed with three questions pictures not being supported to paste. Having said this, it must be noted that the student evidence, is deliberately not included in the paper because the focus is on Chat GPT not students' work.

As for the instrument validity, the questions where moderated by qualified medical technologist for a first-year entry level compliance in line with the Health Profession Council of South Africa (HPCSA), Department of Higher Education (DHET), and Council of Higher Education (CHE). These questions were rated according to the Blooms taxonomy cognitive levels and then graded accordingly with each given mark based on the memorandum. ChatGPT 3.5 responses were reviewed by the assessor using the following criteria: 1= answers were accurate requiring only minor arranging, 2= answers would require substantial amendments in order to be appropriate, 3= answers were incorrect or significantly misleading. The results are shown in Table 3. For a response to receive a score of 1 all points or more should be in the assessor's memo needed. This approach ensure rigorous evaluation of the responses, the worst score from the reviewer was used for the final analysis. For answers requiring substantial modifications to be useful (score = 2) were further characterized in Table 4. ChatGPT wrote the questions and answers in accordance with what was in its vocabulary at that time text.

Regarding the instrument reliability, the AI's answers were evaluated by 3 reviewers using the predefined criteria. In order to ensure reliability, any discrepancies in scoring were resolved through discussion, and the most rigorous score was recorded for each answer. This approach, using the lowest reviewer score, provided a conservative and reliable assessment of ChatGPT's performance. No formal statistical reliability metric was calculated due to the small number of tests items, but the consistent scoring process by independent reviewers helped reinforce the reliability of the results.

RESULTS AND DISCUSSION

RESULTS

Overall ChatGPT 3.5 fails to produce appropriate answers to questions containing the diagram for introduction to medical laboratory sciences courses taught by the Faculty in the Department of Health Sciences. An appropriate question used in the study is shown below. ChatGPT correctly wrote and identified the correct answer from its library. For 5/7 questions ChatGPT gave explicit answers and 25% were not attempted.

Table 2: Introduction to medical laboratory sciences topics used in the question paper

- Glass use and properties
- Cylinder
- Pipette
- Centrifugation
- Particles properties
- Autoclave
- Safety

- Accuracy and precision
- Analyse the drawing

Table 3: Reviewers' Scores: Questions and Answers Generated by ChatGPT

Score	Define criteria	#	%
1	Questions and answers were correct	5/8	62.5%
2	Questions and answers would require addition or modification to be appropriate	1/8	12.5%
3	Answers were incorrect or significantly misleading or never attempted	2/8	25%

In the above table 3, clearly ChatGPT -3.5 attempted to all eight questions on the test, providing correct answers for five of them (62.5%), a partially correct answer that required additional modifications for one question (12.5%), and incorrect or no answer for the remaining two questions (25%). This performance corresponded to an overall score of 65 out of 90 marks (72.2%). With 65 of those marks coming from the theory-based questions. In contrast, ChatGPT scored 0 out of the 25 allocated marks to the two diagram-based questions. The results highlight that the model's strong performance on textual theory items and its failure on visual diagram items, as exemplified by the fully accurate AI generated answer shown in figure 1.

Table 4: Explanations on inappropriate questions or answers

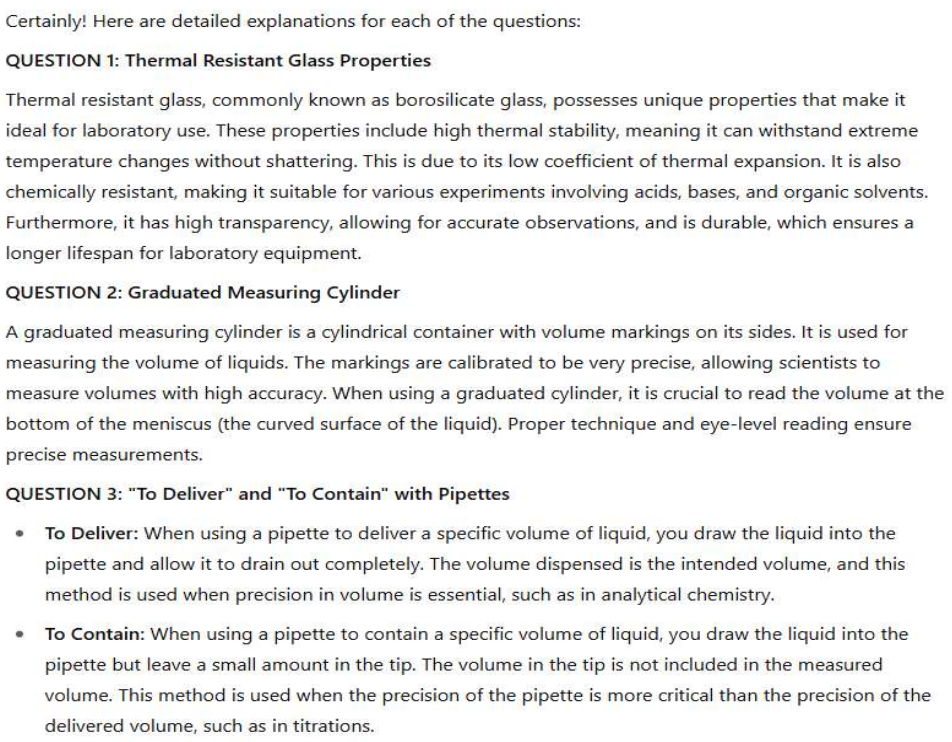
Score	Define criteria	#	%
2.1	There is no explanation for incorrect answers	0/8	0%
2.2	Not enough detail in the explanation for the incorrect/no answers	2/8	25%
2.3	Question written as a negative question, i.e. modification of the question context	0/8	0%

The exceptional performance of AI and the 12 students highlights its potential as an effective tool for assessing theoretical knowledge. However, AI ChatGPT-3.5's incapability to answer questions involving diagram interpretation exposed a critical limitation. While twelve human students navigated the visual aspects of the questions, AI struggled to provide accurate responses. Thus, these results correlate with the study done by Gupta et al., 2023, and Suchman et al., 2023, which showed that AI has identified limitations and variable correctness in certain disciplines.

DISCUSSION

The 60 students participated in the formal test, which included several theory questions in the Introduction to Medical Laboratory Science formative continuous assessment. Following the test, AI analyzed and responded to questions. Remarkably, twelve students and AI demonstrated exceptional proficiency, answering all theory questions correctly and achieving a perfect score of 100% (Brown, 2022). However, 45 of 60 students faced challenges answering two additional questions that required diagram interpretation.

The high accuracy of AI in theory questions underscores its potential as a valuable resource for medical education, teaching and learning, and assessment. Nevertheless, the limitations in diagram interpretation reveal areas where ChatGPT 3.5 currently faces challenges compared to human counterparts. Further, the news development in AI can comprehend and interpret visual data, but in countries with poor resources, relying on ChatGPT3.5 is crucial for governments and institutions of higher learning to invest in AI technology and its comprehensive integration into medical education. For students who find difficulty in some or all questions attempted correctly by ChatGPT-3.5, AI would play a critical role as teaching and learning tool as found in a study by Memariana and Doleck, 2023.



Certainly! Here are detailed explanations for each of the questions:

QUESTION 1: Thermal Resistant Glass Properties

Thermal resistant glass, commonly known as borosilicate glass, possesses unique properties that make it ideal for laboratory use. These properties include high thermal stability, meaning it can withstand extreme temperature changes without shattering. This is due to its low coefficient of thermal expansion. It is also chemically resistant, making it suitable for various experiments involving acids, bases, and organic solvents. Furthermore, it has high transparency, allowing for accurate observations, and is durable, which ensures a longer lifespan for laboratory equipment.

QUESTION 2: Graduated Measuring Cylinder

A graduated measuring cylinder is a cylindrical container with volume markings on its sides. It is used for measuring the volume of liquids. The markings are calibrated to be very precise, allowing scientists to measure volumes with high accuracy. When using a graduated cylinder, it is crucial to read the volume at the bottom of the meniscus (the curved surface of the liquid). Proper technique and eye-level reading ensure precise measurements.

QUESTION 3: "To Deliver" and "To Contain" with Pipettes

- **To Deliver:** When using a pipette to deliver a specific volume of liquid, you draw the liquid into the pipette and allow it to drain out completely. The volume dispensed is the intended volume, and this method is used when precision in volume is essential, such as in analytical chemistry.
- **To Contain:** When using a pipette to contain a specific volume of liquid, you draw the liquid into the pipette but leave a small amount in the tip. The volume in the tip is not included in the measured volume. This method is used when the precision of the pipette is more critical than the precision of the delivered volume, such as in titrations.

Fig 1. An example of the correct answer and explanations generated by ChatGPT

Figure 1 above shows an example of an answer that corresponds to human response to those question and no modification was required. In line with the need for a human voice, students are encouraged to use such output as guide and rewrite the answer according to individual understand without losing the context of understanding. ChatGPT-3.5 attempt provided sufficient response and scored a 1 as per Table 3. Majority of responses (62.5%) fell in this category.

Figure 2 below shows ChatGPT-3.5 response that requires the facilitator carefully read when awarding the marks because the examples given by the students because examples provided are limited to. Such an example of response score 2 in accordance with Table 3. This type of responses in this category accounted for (12.5%), and this correlated with the report done by Mhlanga (2023), who iterated the need to very AI responses and plagiarism in ensuring that examples are not only those from the AI generated outputs. Therefore, bringing comparison of AI and human responses, it became clear that beyond the scores, the nature of ChatGPT answers differed from those of the students in notable ways. For theory questions ChatGPT – 3.5 often produced highly detailed, well-structured answers, of similar nature to textbook explanations. In contrast, human students' answers varied in depth, and some were concise bullet point responses, and some included minor errors or omitted details that ChatGPT provided. Therefore, this proposes that the AI deliver thorough responses constantly, whereas students might summarise or occasionally misunderstand selected points. Nevertheless, for the diagram-based questions, the human students outperformed the AI not by virtue of greater factual comprehension, but by applying visual interpretation skills. Utmost students endeavored to describe or analyse the provided image, for example, identifying components of a laboratory instrument or interpreting a graphical result, whereas ChatGPT could not generate any meaningful answer due to a lack of visual input processing. When it came to partially correct or nuanced answers, students sometimes received partial credit for reasoning that was on the right track, even when their knowledge was incomplete. ChatGPT responses, on the other hand, were usually all or nothing, either a nearly perfect answer drawn from its training data or an inability to answer with no middle ground. Moreover, students occasionally used class – specific terminology or examples in their answers that the AI did not, indicating that students drew on lecture notes or practical lab experiences that are

outside the AI's text-only training. Consequently, this means that while ChatGPT – 3.5's correct answers corresponded or even exceeded student answers in precision and completeness, students' responses showed depths in areas requiring contextual insight and visual reasoning.

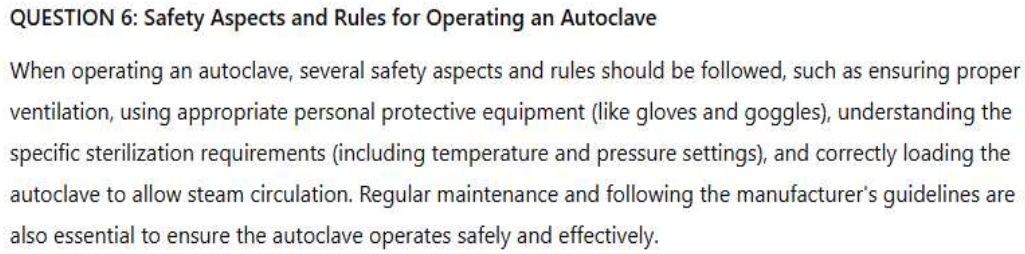


Fig 2. Example of the correct answer requiring a minor modification or explanations generated by ChatGPT

ChatGPT indicated challenges with the question seeking assistance with the illustrations as shown on figure 3, then on figure 4 below an attempt to seek assistance was responded with the answer that ChatGPT-3.5 doesn't support such an option.

Fig 3. An example of correct answer and explanations generated by ChatGPT

Countries and institutions relying on ChatGPT-3.5 available free limit their student opportunity to embrace technology and adapt to the growing data driven world. The success of AI in answering theory questions underscores its ability to comprehend and analyze textual information effectively (Brown, 2022). Nevertheless, the limitations observed in diagram interpretation emphasize the current boundaries of AI ChatGPT-3.5 capabilities which is mostly the sole AI available for the poor countries and other institutions of higher learning in the world. We need to appreciate the development of ChatGPT-4o and others as shown in figure 5 to address these limitations. As such this paper shows that previously disadvantaged countries, institutions of higher learning still resorting to free AI will be left behind and disadvantage their students on an opportunity to compete with the world.

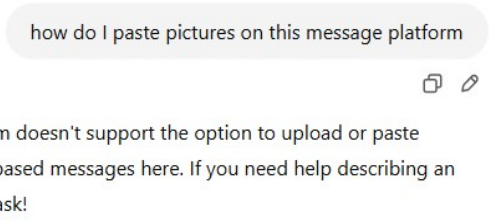
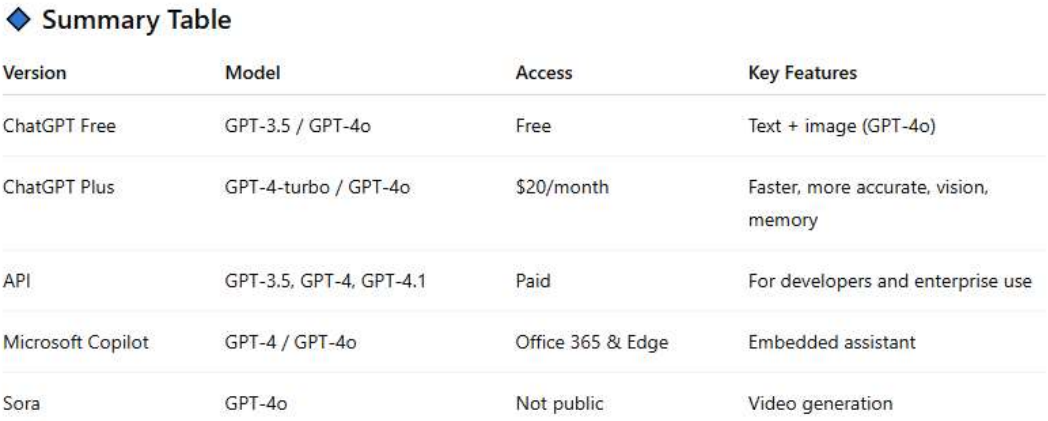


Fig 4. Example of response when ChatGPT was asked to assist with pictures download or pasting

This response is a demonstration of the model's inherent constraints as text only system rather than a mere interface issue. During the period of this study, ChatGPT - 3.5 neither accepted image inputs nor performed visual reasoning. Consequently, when assessment required interpretation of a diagram, the model could not generate a substantive answer because the information needed to respond was contained in an image. This modality mismatch not a deficit in factual knowledge which then explains why ChatGPT – 3.5's achievement on the theory and non-achievement on the two diagrams-based questions. Differently, human students could visually explore the diagram and attempt an interpretation, often securing partial scoring even when their cognitive was undeveloped.



Summary Table

Version	Model	Access	Key Features
ChatGPT Free	GPT-3.5 / GPT-4o	Free	Text + image (GPT-4o)
ChatGPT Plus	GPT-4-turbo / GPT-4o	$20/month	Faster, more accurate, vision, memory
API	GPT-3.5, GPT-4, GPT-4.1	Paid	For developers and enterprise use
Microsoft Copilot	GPT-4 / GPT-4o	Office 365 & Edge	Embedded assistant
Sora	GPT-4o	Not public	Video generation

Fig 5. An example of response when ChatGPT was asked how many free hits there for GPT-4o are

Previously disadvantaged countries and under-resourced institutions that rely solely on free versions of AI tools, such as the limited-access ChatGPT-3.5 model (as shown in fig 5), inadvertently expose their students to the full spectrum of emerging technologies. While free ChatGPT-3.5 provides a valuable starting point, it lacks enhanced capabilities, such as multimodal input, memory, real-time reasoning, and broader contextual understanding that are available in more powerful AI models like ChatGPT-4-turbo and ChatGPT-4o. This technological gap restricts students' ability to engage deeply with cutting-edge applications in teaching and learning, research, innovation, and problem-solving. Therefore, these students may fall further behind in a global landscape increasingly shaped by advanced AI tools, widening the digital divide and restrictive opportunities for substantial participation in the knowledge economy.

Figure 6, below, noticeably shows that those relying on free AI may enjoy only restricted messages (3-5 hours) before they have to wait 5 hours to get this advanced opportunity with ChatGPT-4o again. Those students who can afford it may enjoy 80 messages per 3 hours at a consistent fee.

Thus, this study shows that a significant technological gap exists between GPT-3.5 and its more advanced counterparts, such as GPT-4-turbo and GPT-4o, with important implications for higher education institutions and national innovation agendas. Empirical evidence from this study highlights GPT-3.5's notable limitations, particularly its inability to process or interpret visual inputs, such as diagrams, charts, and images. This restriction is particularly striking given the increasing integration of multimodal learning materials in academic environments. As shown in Table 3, GPT-3.5 scored (2.3), showing that the overall validity of questions answered was perfect, indicating that it did not flag, correct, or rewrite any of the assessment items, thus validating the internal consistency of the question set.

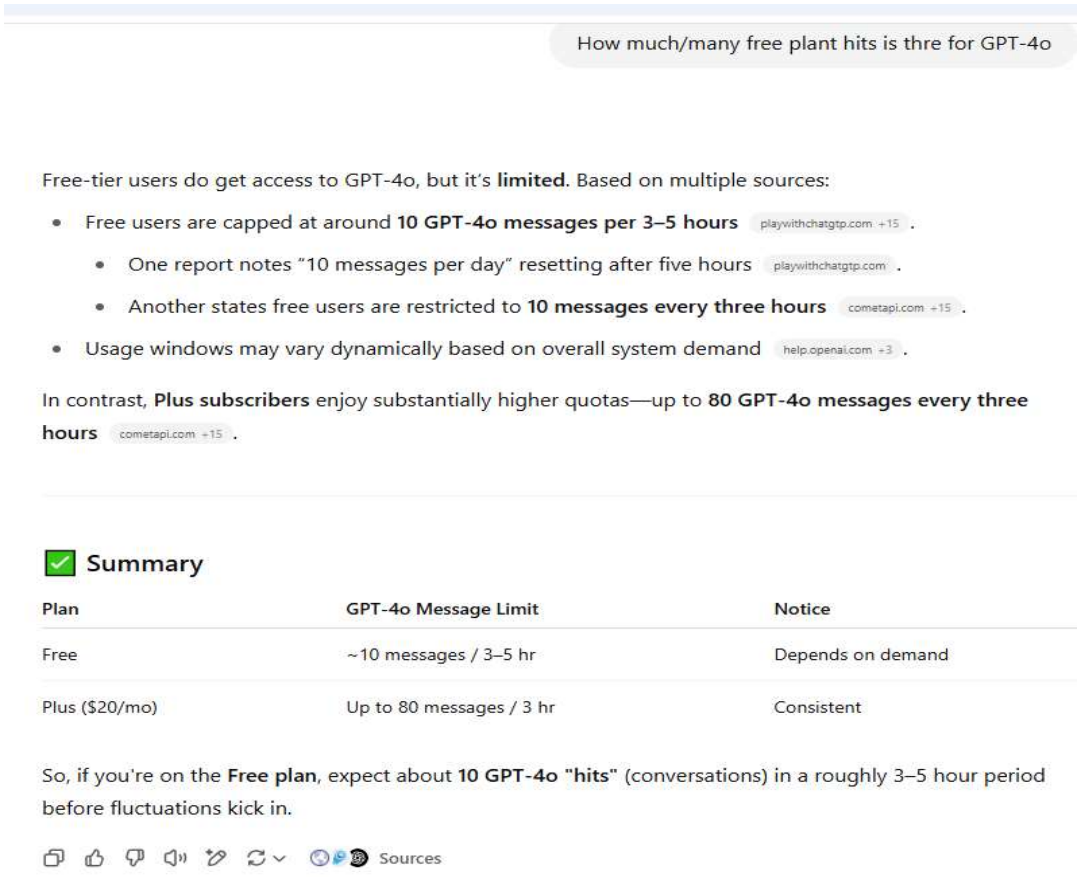


Fig 6. Example of respond when ChatGPT was asked how many free hits there for GPT-4o are

Nevertheless, performance dropped significantly on the two diagram-based questions, which made up 25% of the test set, scoring 2.2, while ChatGPT-3.5 obtained a score of 2.1 due to its restricted capacity for multimodal reasoning. Importantly, no item in the question set fell into the lowest validity category (2.1), further affirming that the observed gap is not due to poor question construction but to the model's technological ceiling. These findings emphasize the urgent need for universities and governments, particularly in the Global South, to rethink their digital transformation strategies and begin actively integrating cutting-edge AI models that can support both textual and visual cognition in educational and research contexts

LIMITATIONS OF THE STUDY

This single institution study examined one introductory module using one assessment, which limits generalizability and prevents conclusions about sustained leaning effects. We did not conduct a formal psychometric evaluation of the instrument for example; internal consistency or test – retest, so score differences should be interpreted cautiously. Since we evaluated the free, text only ChatGPT- 3.5 available at the time, the model could not process diagrams. Consequently, performance on image-based items reflects a modality constraint rather than purely content knowledge. Moreover, large language model (LLM) performance varies substantially across versions, exams, and item formats, which constrains external validity of point estimates from a single model (Gordon et al., 2024; Jin et al., 2024; Liu et al., 2024). Thus, our study focused on test accuracy rather than educational outcomes, recent evidence syntheses emphasize that robust impacts on health – professions learning remain under-documented, and that current LLMs still show important safety and reliability limitations in clinical reasoning contexts (Feigerlova et al., 2025; Hager et al., 2024; Chen et al., 2024).

PRACTICAL RECOMMENDATIONS FOR EDUCATORS IN RESOURCE-CONSTRAINED SETTINGS

Combining AI with Human Feedback

Lecturers in resource-constrained settings should integrate ChatGPT precisely as a supplement to human teaching, not a replacement. For example, educators can use ChatGPT – 3.5 to generate draft quiz questions, explanations, or illustrative examples, but then review and edit these outputs before using them in class. Likewise, students might be encouraged to explore ChatGPT's responses when studying, but educators should discuss these AI-generated answers in class, correcting any errors and highlighting where the AI's reasoning is sound or flawed. The blended approach leverages the efficiency of AI while ensuring a knowledgeable human provides oversight and advance context.

Promote Critical AI Literacy

It is fundamental to train students to use AI tools critically and responsibly. In a low-resource environment where ChatGPT – 3.5 might be one of the few readily available learning aids, students should learn how to question and verify AI-generated content. Hence, lecturers should incorporate short lessons on AI's limitations, for example, explaining that ChatGPT may sound confident but can sometimes produce incorrect information, and highlighting its inability to process certain data types. Openly discussing instances when ChatGPT is wrong and why that happened, educators foster an environment of critical thinking rather than blind trust in AI. This AI literacy includes understanding issues of bias, the importance of citing reliable sources, and the ethical use of AI. Enabling students with these critical skills ensures that, even in under-resourced settings, the introduction of AI in education leads to improved learning rather than misinformation.

RECOMMENDATIONS FOR FUTURE RESEARCH

Future work should replicate these findings across multiple institutions and larger cohorts, incorporate psychometric validation of the test, such as reliability and validity evidence, and track longitudinal learning outcomes (knowledge retention, transfer, and performance) rather than single-sitting accuracy (Feigerlova et al., 2025; Gordon et al., 2024). Given our diagram findings, evaluate multimodal models, for example GPT – 4/4.0 on the same instrument, especially image -based items, while reporting results by item type (Chen et al., 2024; Liu et al., 2024). Therefore, future studies should also compare human-only, AI-only, and human-AI hybrid feedback for formative assessment and explore critical AI literacy interventions to help learners appraise model outputs. Finally, in resource-constrained settings, prioritize low-cost interventions such as created prompts, shared-access labs, faculty micro-training, and clear supervision protocols to alleviate known LLM failure modes in clinical reasoning (Gordon et al., 2024; Hager et al., 2024).

CONCLUSION

In concluding this study, it is crucial to note that it highlights the notable performance of artificial intelligence, specifically ChatGPT-3.5, in answering theoretical questions in an introductory medical science assessment, often surpassing the capabilities of first-year students. Nonetheless, its difficulty in interpreting visual information, such as diagrams, reveals a critical limitation in current AI models. The findings, encapsulated in the proposed title "AI in Medical Education: Excelling in Theory, Navigating Visual Challenges," underscore both the promise and constraints of AI integration in medical education. Likewise, the reliance on free, less capable AI models, particularly in previously disadvantaged countries and under-resourced institutions, risks widening the digital and academic divide. To ensure equitable access and global competitiveness, institutions of higher learning are strongly encouraged to invest in advanced AI tools. Also, the technologies that can enrich learning, foster innovation, and better prepare the students for the demands of a technologically driven future.

ACKNOWLEDGMENTS

Authors acknowledge the university's availability of resources that were used during the study.

INFORMED CONSENT

No student's personal information was disclosed, and their work was not revealed in this paper.

ETHICAL APPROVAL

Though AI wrote the test and answers showed in screen shot are from AI, none of the answers from people / students were used except indication holistically how the group perfumed. The ethical clearance for study is still available with reference, CUT/REIC 2025/000606.

REFERENCES

Al-Worafi, Y.M., Hermansyah, A., Goh, K.W., Ming, L.C. (2023). Artificial intelligence use in university: Should we ban ChatGPT? *preprints.org,* 2023020400. https://doi.org/10.20944/preprints202302.0400.v1

Bin Y, Mandal D (2019) English teaching practice based on artificial intelligence technology. *Journal of Intelligent Fuzzy Systems*, 37: 3381-91. https://doi.org/10.3233/JIFS-179141

Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *1*, 100002. https://doi.org/10.1016/j.caeai.2020.100002

Chen, Y., Huang, X., Yang, F., Lin, H., Lin, H., Zheng, Z., … Li, X. (2024). Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: A comparison study. *BMC Medical Education, 24*, 1372. https://doi.org/10.1186/s12909-024-06309-x

Choi, J.H.; Schwarcz, D.B. (2023). AI Assistance in Legal Analysis: An Empirical Study. *Minnesota Legal Studies Research* Paper No. 23-22. https://doi.org/10.2139/ssrn.4539836

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. I*nternational Journal of Educational Technology in Higher Education*, 20(1), 1-22. https://doi.org/10.1186/s41239-023-00392-8

Dowling, M., & Lucey, B. (2023). ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters*, 53, 103662

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). Opinion Paper:"So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International journal of information management*, *71*, 102642. https://doi.org/10.1016/J.IJINFOMGT.2023.102642

Feigerlova, E., Hani, H., & Hothersall-Davies, E. (2025). A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. *BMC Medical Education, 25*, 129. https://doi.org/10.1186/s12909-025-06719-5

Gao, P. P., Nagel, A., & Biedermann, H. (2020). Categorization of Educational Technologies as Related to. *Pedagogy in basic and higher education: Current developments and challenges*, 167. https://doi.org/10.5772/intechopen.88629

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: "Breakthrough? Or buncombe and ballyhoo?". *Journal of Computer Assisted Learning,* 37(5), 1207-1216. https://doi.org/10.1111/jcal.12577

Gordon, M., Daniel, M., Ajiboye, A., Uraiby, H., Xu, N. Y., Bartlett, R., et al. (2024). A scoping review of artificial intelligence in medical education: BEME Guide No. 84. *Medical Teacher, 46*(4), 446–470.

https://doi.org/10.1080/0142159X.2024.2314198

Gupta, B.B., Gaurav, A., Panigrahi, P.K., and Arya, V. (2023). Analysis of artificial intelligence-based technologies and approaches on sustainable entrepreneurship. *Technological Forecasting and Social Change*. 186(Part B). 122152. https://doi.org/10.1016/j.techfore.2022.122152

Hager, P., Jungmann, F., Holland, R., et al. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine, 30*, 2613–2622. https://doi.org/10.1038/s41591-024-03097-1

Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. *Contemporary Educational Technology,* 15(2), ep421. https://doi.org/10.30935/cedtech/ 13036. https://doi.org/10.30935/cedtech/13036

Hargreaves, E. (2005). Assessment for learning? Thinking outside the (black) box. *Cambridge Journal of Education*, *35*(2), 213-224. https://doi.org/10.1080/03057640500146880

https://beta.openai.com/overview

https://kejarcita.id/

https://openai.com/chatgpt.

Jin, H. K., Lee, H. E., & Kim, E. (2024). Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: A systematic review and meta-analysis. *BMC Medical Education, 24*, 1013. https://doi.org/10.1186/s12909-024-05944-8

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), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Liu, M., Okuhara, T., Chang, X., Shirabe, R., Nishiie, Y., Okada, H., & Kiuchi, T. (2024). Performance of ChatGPT across different versions in medical licensing examinations worldwide: Systematic review and meta-analysis. *Journal of Medical Internet Research, 26*, e60807. https://doi.org/10.2196/60807

Lo, C.K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*. 13(4), 410, 1-15. https://doi.org/10.3390/educsci13040410

Memarian, B., & Doleck, T. (2023). Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, *5*, 100152. https://doi.org/10.1016/j.caeai.2023.100152

Mhlanga, D. (2023). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. https://doi.org/10.2139/ssrn.4354422

Moore, S., Nguyen, H. A., Bier, N., Domadia, T., & Stamper, J. (2022). Assessing the quality of student-generated short answer questions using GPT-3. In I. Hilliger, P. J. MunozMerino, T. D. Laet, A. Ortega-Arranz & T. Farrell (Eds.), Educating for a new future: *Making sense of technology-enhanced learning adoption* (pp. 243-257). Springer. https://doi.org/10.1007/978-3-031-16290-9_18

Moore, S., Nguyen, H.A., Chen, T., & Stamper, J. (2023). Assessing the Quality of MultipleChoice Questions Using GPT-4 and Rule-Based Methods. In O. Viberg, I. Jivet, P. K. Munoz-Merino, M. Perifanou & T. Papathoma (Eds.), *Responsive and Sustainable Educational Features* (pp. 229-245). Springer. https://doi.org/10.1007/978-3-031-42682-7_16

Moshirfar, M., Altaf, A. W., Stoakes, I. M., Tuttle, J. J., & Hoopes, P. C. (2023). Artificial Intelligence in Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering StatPearls Questions. *Cureus*, *15*(6), e40822. https://doi.org/10.7759/cureus.40822

Olney, A.M., Pavlik Jr, P.I., & Maass, J.K. (2017, June). Improving reading comprehension with automatically generated cloze item practice. I*n International Conference on Artificial Intelligence in Education* (pp. 262-273). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-61425-0_22

Suchman, L. (2023). The uncontroversial 'thingness' of AI. *Big Data & Society*, *10*(2). https://doi.org/10.1177/20539517231206794 (Original work published 2023)

Van Vaerenbergh, S., & Pérez-Suay, A. (2022). A classification of artificial intelligence systems for mathematics education. In *Mathematics education in the age of artificial intelligence: How artificial intelligence can serve mathematical human learning* (pp. 89-106). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-86909-0_5

Yang SJ, Ogata H, Matsui T, Chen NS. Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*. 2021 Jan 1; 2:100008. sciencedirect.com. https://doi.org/10.1016/j.caeai.2021.100008

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., ... & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*, *2021*(1), 8812542. https://doi.org/10.1155/2021/8812542