Rima A H, Al-Motassem Y, Ahmed A, Aya N A, Jude K D, Majd I S, Ghaith F A R, Marwan H A. ChatGPT performance on pharmacology examination and board review questions: Implications for medical education and knowledge assessment.Pharmacy Practice 2026 Apr-Jun;24(2): 3488.

https://doi.org/10.18549/PharmPract.2026.2. 3488

Original Research

ChatGPT performance on pharmacology examination and board review questions: Implications for medical education and knowledge assessment

Rima A.Hijazeen, Al-Motassem Yousef, Ahmed Almousa, Aya N Alzoghair, Jude K. Dwairi, Majd I. Sawaqed , Ghaith F. Al-Ryahneh , Marwan H. Ali

Received (first version): 06-Nov-2025 Accepted: 02-Apr-2026 Published online: 29-May-2026

Abstract

Objectives: This study aimed to evaluate ChatGPT's performance on pharmacology exam questions by assessing its accuracy in basic and clinical pharmacology, reasoning processes, and response consistency over time. **Methods:** A dataset of 583 multiple-choice questions from the Pharmacology Examination and Board Review (13th edition) was used. ChatGPT's responses were evaluated for logical justification, use of internal question stem information, and integration of external knowledge. Statistical analyses, including chi-square and McNemar tests, assessed associations and changes in response accuracy over a four-week interval. **Results:** ChatGPT achieved 76.2% accuracy (444/583 questions), demonstrating logical reasoning in 97% of responses. Internal information was used in 99.7% of cases, while external information was incorporated in 98% of correct and 93.5% of incorrect responses ($p = 0.008$). Information errors were the most common reason for incorrect answers. A statistically significant improvement in accuracy upon re-evaluation ($\chi^2 = 37.3$, $p < 0.0001$) was observed, suggesting potential temporal variation in performance. **Conclusion:** ChatGPT meets or exceeds typical passing standards in many educational settings, with evidence of improved response accuracy over time. These findings highlight its capabilities in processing pharmacological content, with potential implications for future research into AI-assisted educational tools.

Keywords: Artificial intelligence, ChatGPT, Medical education, Pharmacology, Reasoning, Multiple-choice questions, Large language model.

Rima A. Hijazeen*. B.Sc. Pharmacy, M.Sc. Clinical Pharmacy, PhD Clinical Pharmacy Practice
Associate Professor in Clinical Pharmacy Practice, The University of Jordan Faculty of Pharmacy Department of Biopharmaceutics and Clinical Pharmacy, Amman 11942, Jordan r.hijazeen@ju.edu.jo
Al-Motassem Yousef. B.Sc. Pharmacy, PhD in Pharmacology and Therapeutics, Professor of Pharmacology and Therapeutics, The University of Jordan, Faculty of Pharmacy, Department of Biopharmaceutics and Clinical Pharmacy, Amman 11942, Jordan. ayousef@ju.edu.jo
Ahmed Almousa. RPh, MSc, PhD, Assistant Professor of Clinical Pharmacy, Department of Biopharmaceutics and Clinical Pharmacy, The Faculty of Pharmacy, University of Jordan. Amman-Jordan. a.almousa@ju.edu.jo
Aya N. Alzoghair. Undergraduate Pharmacy student, The University of Jordan Faculty of Pharmacy, Amman 11942, Jordan. ayaalzoghair3@gmail.com
Jude K. Dwairi. Undergraduate Pharmacy student, The University of Jordan Faculty of Pharmacy Amman 11942, Jordan.jdwairi2002@yahoo.com
Majd I. Sawaqed. Undergraduate Pharmacy student, The University of Jordan Faculty of Pharmacy, Amman 11942, Jordan.majdsawaqed@hotmail.com
Ghaith F. Al-Ryahneh. Undergraduate Pharmacy student, The University of Jordan Faculty of Pharmacy, Amman 11942, Jordan. Ghaithfadi02@gmail.com
Marwan H Ali. Undergraduate Pharmacy student, The University of Jordan Faculty of Pharmacy, Amman 11942, Jordan. marwanalsheikh9@gmail.com

INTRODUCTION

In November 2022, OpenAI introduced the Chat Generative Pre-Trained Transformer (CHAT GPT), a highly advanced large language model (LLM) designed to generate conversational, contextually relevant responses to natural language prompts. Leveraging deep learning algorithms, CHAT GPT has been trained on an extensive corpus of text to achieve state-of-the-art performance in natural language processing tasks[1,2]. It has demonstrated significant potential across various applications, including customer service, content creation, and, notably, in medical and health education[3,4]. As a chatbot, CHAT GPT is specifically fine-tuned on conversational prompts, enabling it to simulate a dialogic interaction that is coherent and contextually adaptive[1,2,5,6].In the medical field, CHAT GPT represents a novel model capable of merging clinical knowledge with conversational engagement[6-8]. This combination has promising applications in personalised patient education and consumer health communication[8]. CHAT GPT has already been explored for potential applications across diverse medical specialities, including endocrinology, hepatology, cardiology, and neurology, as well as in areas of clinical decision-making, medical imaging, and predictive modelling[8-11]. Its ability to generate coherent medical content and respond dynamically to user inputs suggests a significant utility in enhancing healthcare education and interaction.Literature indicates CHAT GPT demonstrates the capability to pass university-level multiple-choice question (MCQ) examinations in disciplines such as law, medicine, and business[12]. Notably, it has achieved scores comparable to those of third-year undergraduate medical students on the United

https://doi.org/10.18549/PharmPract.2026.2. 3488

States Medical Licensing Examination (USMLE), surpassing the passing threshold in each test[13-15]. Its ability to provide an interactive, on-demand learning environment underscores its potential as a valuable tool in medical education and learning support[14]. However, generative models like CHAT GPT have shown limited success in testing clinical knowledge through question-answering tasks[16,17].This study evaluated the performance of CHAT GPT in pharmacology, a cornerstone of medical education that encompasses pharmacodynamics, pharmacokinetics, and clinical drug therapy. Using a dataset of 583 MCQs sourced from the Pharmacology Examination and Board Review, 13th edition[18], the study examines the model's accuracy, logical justification, and use of intrinsic and extrinsic information in its responses. Additionally, the study explores ChatGPT's potential for iterative performance improvement by re-evaluating questions it initially answered incorrectly, aiming to provide insight into its evolving performance and possible applicability as a future educational tool.

Research Aims and Objectives

To evaluate the performance and educational potential of ChatGPT in pharmacology through an analysis of its accuracy, reasoning patterns, and changes in response accuracy over time.To explore changes in ChatGPT's response accuracy over time by re-assessing its answers to initially incorrect questions.

1. To assess CHAT GPT's accuracy in answering a dataset of 583 pharmacology-related MCQs from the Pharmacology Examination & Board Review, 13th edition.
2. To evaluate the logical justifications provided by CHAT GPT, with a focus on its utilisation of intrinsic information from question stems and extrinsic information from its knowledge base.
3. To explore changes in ChatGPT's response accuracy over time by re-assessing its answers to initially incorrect questions.
4. To explore the potential of CHAT GPT as a supplemental educational resource in pharmacology, based on its accuracy, reasoning patterns, and capacity for knowledge integration.

MATERIALS AND METHODS

Construction of Test Item Dataset

The Pharmacology Examination & Board Review, 13th edition, 2021, was selected as the source of test items due to its widespread use in pharmacology education and its relevance as a preparatory resource for board and licensing examinations. The textbook, authored by pharmacology experts, includes clinically and educationally validated MCQs covering pharmacodynamics, pharmacokinetics, mechanisms of action, and therapeutic applications.

A total of 583 MCQs were selected to create a comprehensive and diverse dataset reflective of both basic and clinical pharmacology. Questions were independently reviewed by a team of pharmacists and pharmacologists to ensure clarity, representativeness, and suitability for analysis by a large language model. Items containing figures (n = 17), visual content, or ambiguous formatting were excluded, as ChatGPT only accepts text-based input. The final dataset spanned a wide range of pharmacological topics intended to evaluate recall, interpretation, and applied reasoning. MCQs are a widely accepted assessment tool in medical education, offering a reliable metric for evaluating cognitive understanding and reasoning abilities in complex subjects like pharmacology[13,19]

Prompt Engineering and Input Standardisation

Prompt engineering refers to the strategic construction of input text to guide large language models (LLMs) like ChatGPT in generating accurate, relevant, and contextually appropriate responses. Because the performance of LLMs is highly sensitive to input phrasing and structure, effective prompt engineering plays a critical role in ensuring reliable outputs.In this study, we designed our prompts with three key principles in mind: (1) clarity of instruction, (2) structured presentation of question components, and (3) elimination of ambiguity. Each MCQ was entered using a standardised format that included a concise narrative stem, a clearly stated question on a new line, and answer choices listed separately. This format ensured that ChatGPT could parse the information efficiently and base its reasoning solely on the textual input provided. Each MCQ was structured with:

1. A narrative stem (clinical scenario or context),
2. A direct question on a new line,
3. Four or five multiple-choice options are presented line by line.

This format was applied uniformly across the dataset to reduce input variability and enhance ChatGPT's interpretative accuracy. A sample of this standardised structure is illustrated in Figure 1. It is important to clarify that while our prompt engineering standardised the structure and clarity of the input, it did not constrain ChatGPT to using only internal information from the prompt. The model, by design, draws on both the input text and its pre-trained knowledge base to generate responses. Our goal was not to prevent external knowledge use, but to analyse when and how such information was integrated. Accordingly, hallucinations were assessed post hoc by identifying instances where ChatGPT introduced information unsupported by the question stem or established pharmacological knowledge.To minimise randomness or overly creative outputs, questions were submitted in a factual, exam-style format, and the default temperature setting of the

https://doi.org/10.18549/PharmPract.2026.2. 3488

Example Question

Please answer this multiple-choice question.

Your 30-year-old patient has moderately severe new-onset asthma, and you prescribe a highly selective β-2 agonist inhaler to be used when needed. In considering the possible drug effects in this patient, you would note that β-2 stimulants frequently cause

The correct answer is.

(A) Direct stimulation of renin release

(B) Hypoglycemia

(C) Itching due to increased cGMP (cyclic guanine monophosphate) in mast cells

(D) Skeletal muscle tremor

(E) Vasodilation in the skin

ChatGPT's response

The correct answer is (D) Skeletal muscle tremor.

Beta-2 (β2) agonists are commonly used in the treatment of asthma as bronchodilators. While they primarily act on the smooth muscles of the bronchioles to relax and dilate them, they can also have systemic effects. One of the known side effects of β2 agonists is skeletal muscle tremor.

Option (A) Direct stimulation of renin release is not a common effect of β2 agonists. Renin is an enzyme involved in the regulation of blood pressure and is not directly influenced by β2 agonists.

Option (B) Hypoglycemia is not a typical side effect of β2 agonists. In fact, β2 agonists can have a mild hyperglycemic effect by promoting glycogenolysis and inhibiting insulin release.

Option (C) Itching due to increased cGMP in mast cells is not a known effect of β2 agonists. Increased cGMP is associated with vasodilation and smooth muscle relaxation, but it is not directly related to itching.

Option (D) Skeletal muscle tremor is a common side effect of β2 agonists, especially when administered in higher doses or via systemic routes such as oral or intravenous administration.

Option (E) Vasodilation in the skin is not a primary effect of β2 agonists. β2 agonists primarily act on bronchial smooth muscle, leading to bronchodilation, and do not have a significant vasodilatory effect on the skin.

Therefore, the most frequent effect noted with β2 agonists, as described in the scenario, is option (D) Skeletal muscle tremor.

Figure 1. The typical style of each question

free version of ChatGPT was used, which biases the model toward more deterministic and evidence-based responses.All questions were prepared in a single Microsoft Word document in advance by the research team. One researcher was then responsible for conducting all model interactions. Questions were submitted to the free version of ChatGPT (April 15, 2023) using a single OpenAI account. The testing was conducted over one day to minimise model variation and ensure consistency. Questions were grouped by textbook chapter, and each group was submitted in a separate chat session to avoid memory contamination across chapters (Figure 1).

Model Testing Protocol

Approximately 6–10 questions were submitted at a time within a single chat window, and responses were extracted manually. ChatGPT was not given any feedback or information about whether its answers were correct. No custom instructions or memory settings were applied. As the free version of ChatGPT does not allow for modification of model parameters, such as temperature, default OpenAI settings were used throughout the experiment.Each response was recorded in a structured spreadsheet that included:

- Question number
- Correct answer
- ChatGPT answer
- Explanation generated by the model
- Binary indicators for: correctness, logical reasoning, internal information use, external information use
- Error categorisation (if incorrect): logical error, information error, statistical error, or combined error
- Re-evaluation accuracy (4 weeks later, where applicable)

Only one researcher conducted the interaction with the model, thereby minimizing input inconsistencies. No responses were influenced by prior questions, and no cumulative learning occurred, as ChatGPT's free version does not retain memory between sessions

Error Categorisation and Expert Review

Correct responses were analysed using criteria adapted from Gilson et al.[14], including:

- **Logical reasoning:** Whether the explanation reflected sound pharmacological reasoning.
- **Internal information:** Whether the response relied on details from the question stem.
- **External information:** Whether the explanation incorporated relevant background knowledge.

Incorrect answers were further categorised as:

- **Logical error:** Misinterpretation of relevant information.
- **Information error:** Omission or misunderstanding of pharmacological facts.
- **Statistical error:** Misuse or misinterpretation of numerical data.
- **Combined error:** Presence of both logical and information errors.

While the framework for assessing logical reasoning, internal information, and external information was originally adapted for correct responses, we applied it to all ChatGPT outputs—correct and incorrect—to better understand the model's reasoning process regardless of outcome. This approach mirrors human assessment strategies, where a student may interpret a question correctly, draw on appropriate knowledge, and still select the wrong answer. Similarly, ChatGPT at times demonstrated accurate internal and external knowledge integration but failed to select the correct option. By extending the analysis framework to all responses, we aimed to capture these nuances in AI performance and decision-making. All responses were independently reviewed and coded by at least two domain experts (RH, AY, AA, ANA, JD, MS, GR, MA). After initial coding, the research team held consensus meetings to discuss and reconcile discrepancies in categorisation. This process ensured inter-rater reliability and consistency in qualitative assessment.

Statistical Analysis

Statistical analysis was performed using IBM SPSS Statistics for Windows, Version 25.0 (IBM Corp., Armonk, NY). Categorical variables, such as response correctness (correct or incorrect) and the presence or absence of logical reasoning, internal information, and external information, were analysed using chi-square tests to assess associations between response correctness and these qualitative metrics. Incorrect responses were categorised as logical errors, information errors, or combined logical and information errors. The association between error types and response correctness was evaluated using chi-square tests to determine significant relationships. To explore potential variation in ChatGPT's performance over time, a follow-up evaluation of 139 initially incorrect responses was conducted four weeks after the original interaction. The same submission protocol was used, without any feedback or correction provided between sessions. McNemar's test was applied to assess whether the proportion of correct answers had changed significantly. All statistical tests were two-tailed, with a p-value < 0.05 considered statistically significant. This analysis provided a comprehensive evaluation of CHAT GPT's performance, reasoning quality, and potential for iterative learning.

AI Version and Limitations

The study utilised the free version of ChatGPT available on April 15, 2023 (likely GPT-3.5, as defined by OpenAI's release timeline). Temperature parameters and backend model settings were not adjustable in this version. As such, the results represent a one-time snapshot of ChatGPT's pharmacological reasoning capabilities. Longitudinal or version-controlled testing was not performed. While some improvement was noted in the follow-up, it cannot be attributed to self-learning, as no direct feedback was provided and the model's internal updates are not transparent to users.

RESULTS

ChatGPT achieved an accuracy of 76.2% (444 out of 583 questions) on pharmacology questions, which meets or exceeds typical passing standards in many educational settings. This level of performance suggests the model's potential to support formative assessment and self-directed learning and reflects the model's ability to consistently apply logical reasoning and integrate both internal and external information when formulating answers. A detailed analysis showed logical reasoning in 97% of all responses and significant use of relevant external knowledge in correct answers. Moreover, a statistically significant improvement upon re-evaluation of previously incorrect answers after four weeks (23.7% improvement, $p < 0.0001$) suggests potential for iterative performance enhancement. While the model's capabilities indicate promise for use in educational settings, this study primarily assessed performance on pharmacology MCQs and did not directly evaluate its application as a teaching tool. Therefore, any educational utility should be interpreted as a possible future direction supported by performance data, rather than a confirmed outcome of this investigation.

Overall Performance

CHAT GPT's accuracy level achieved demonstrated its ability to consistently employ logical reasoning and effectively integrate internal and external information, as outlined in the qualitative analysis. The model correctly answered 76.2% (444/583) of the questions, with 23.8% (139/583) answered incorrectly. Most errors were attributed to gaps in foundational knowledge, though the model showed potential for improvement upon re-evaluation. These results highlight CHAT GPT's capability to address the majority of pharmacology-related questions accurately, demonstrating its utility in supporting medical education. A summary of the performance is provided in Table 1.

Qualitative Breakdown of Responses

To understand the characteristics of CHAT GPT's responses, a qualitative analysis examined logical reasoning, internal information utilisation, and external knowledge integration. Table 2 evaluates ChatGPT's answer quality across three metrics: logical reasoning, internal information, and external information. Logical reasoning was present in 97% of responses, regardless of correctness. A highly significant association was found between answer correctness and logical reasoning ($\chi^2 = 59.33$, df = 1, $p < 0.0001$). CHAT GPT utilised internal information in 99.7% (581/583) of responses, with a significant association between correctness and internal information ($\chi^2 = 6.41$, df = 1, $p = 0.011$), confirming statistical significance at the 0.05 level.This framework was applied across all responses to capture reasoning patterns broadly, not only in correct answers but also to identify how internal and external knowledge was used when errors occurred.Logical reasoning was evident in 96.9% of all responses, with 100% of correct answers demonstrating clear justification for the selected option. Among incorrect responses, 87.1% still contained logical reasoning, indicating that errors often stemmed from information gaps rather than illogical conclusions. Internal information, derived directly from the question stem, was utilised in 99.7% of responses. External information, which integrates knowledge beyond the question content, was present in 98% of correct responses and 93.5% of incorrect ones, highlighting its role in achieving accuracy. Chi-square analysis confirmed significant associations between response correctness and these metrics ($P < 0.05$).

Error Analysis

Incorrect responses were categorised into logical errors, information errors, and combined logical and information errors. Information errors accounted for the majority (61.2%) of incorrect responses, reflecting a lack of essential pharmacological knowledge in certain areas. Logical errors were observed in 31.6% of incorrect answers and typically involved the misapplication of principles, such as suggesting beta-blockers for asthmatic patients. Combined errors, involving both information gaps and flawed reasoning, were the least frequent at 7.2% keeping in mind that statistical errors were the least frequent at 0%. The significant association between error type and correctness ($\chi^2 = 127.54$, df = 3, $p < 0.0001$) underscores the need to address these specific deficiencies in future iterations of the model.

Re-Evaluation of Incorrect Responses

To evaluate CHAT GPT's capacity for iterative learning, 139 responses initially answered incorrectly were re-assessed after a four-week interval. During this re-evaluation, the model corrected 33 of its previous mistakes, representing a 23.7% increase in accuracy for these questions (Table 3). The McNemar test revealed that this improvement was statistically significant ($\chi^2 = 37.3$, $p < 0.0001$), suggesting that ChatGPT's performance may vary across time points. However, since the model did not receive any corrective feedback, and its internal updates are not publicly disclosed, these changes cannot be attributed to self-learning. The majority of responses (106/139) remained incorrect, highlighting the limitations of using static evaluation to interpret underlying model dynamics.

DISCUSSION

This study evaluated ChatGPT's performance on pharmacology-related MCQs, focusing on its accuracy, reasoning patterns, and potential for performance improvement over time. These findings offer preliminary insights into capabilities that may support future integration into educational tools, particularly in pharmacology. The principal findings revealed an accuracy rate of 76.2%, with 444 out of 583 questions answered correctly, underscoring the model's substantial improvement over earlier iterations of generative models. CHAT GPT consistently demonstrated logical reasoning (96.9% of responses) and frequently relied on internal information from the question stem (99.7%) to generate its answers. While this indicates the model's ability to interpret input contextually, such reliance may reflect the structure of the MCQs themselves and does not necessarily equate to deep pharmacological understanding. Correct responses frequently incorporated external knowledge (98%), underscoring the importance of integrating broader data in achieving high accuracy. These findings position CHAT GPT as a promising tool for interactive and efficient learning in pharmacology, particularly in environments with limited access to traditional educational resources. Additionally, the model exhibited a modest potential for iterative improvement , correcting 23.7% of initially incorrect answers upon re-evaluation, further highlighting the potential for evolving performance across different time points. However, such changes may reflect updates to the model or variations in its internal reasoning, rather than a true self-learning process.

Comparison with Prior Models

The performance of CHAT GPT represents a significant advancement over earlier large language models (LLMs) in medical education. Previous models, such as OpenQA and GPT-3.0, demonstrated limited contextual understanding and reasoning capabilities. For example, Ha et al. (2019) reported a 29% accuracy using the OpenQA dataset for United States Medical Licensing Examination (USMLE) Step 1 and Step 2 questions[20]. Similarly, Jin et al. achieved a 36.7% accuracy by combining information retrieval techniques with neural networks on 12,723 USMLE-style questions[21]. These earlier models relied heavily on retrieval-based methods, lacking the ability to synthesise information coherently or provide logical justifications for their answers.In contrast, ChatGPT overcomes many of the limitations observed in earlier large language models by demonstrating enhanced reasoning and contextual understanding. Studies by Gilson et al.[14] and Kung et al.[15] showed that ChatGPT achieved passing scores on USMLE Step 1 and Step 2 exams, highlighting its ability to answer complex, high-stakes medical questions. Our study contributes to this body of evidence by examining ChatGPT's performance on pharmacology-specific MCQs and by conducting a detailed qualitative analysis of its reasoning processes and knowledge integration. The model's consistent use of internal and external information, along with its coherent justifications, reflects capabilities that may be relevant for future research on educational applications, pending further validation in learner-based settings.

Qualitative Insights into Performance

The qualitative analysis of CHAT GPT's performance highlighted its consistent reliance on logical reasoning and internal information, with logical reasoning present in 100% of correct responses and 87.1% of incorrect ones, indicating that errors stemmed primarily from domain-specific knowledge gaps rather than reasoning flaws. This aligns with findings by Kung et al., who reported similar trends in CHAT GPT's logical consistency[15]. Internal information was universally utilised (99.7%), while correct responses more frequently incorporated external knowledge (98%) compared to incorrect ones (93.5%), suggesting that broader knowledge integration significantly enhances accuracy. These results are consistent with studies such as Gilson et al.[14] and Sharma et al.[22], which demonstrated ChatGPT's ability to provide logical justifications even in incorrect responses. Chen et al., further noted that while CHAT GPT excels in certain reasoning tasks, it struggles with complex biomedical reasoning, underscoring the need for improvements in domain-specific knowledge[2]. Collectively, these findings emphasise the importance of both internal and external information integration in CHAT GPT's educational utility.

Error Analysis

Analysis of CHAT GPT's incorrect responses revealed that information errors were the most prevalent, accounting for 61.2% of mistakes. These errors primarily stemmed from gaps in pharmacological knowledge or incorrect recommendations, such as prescribing antibiotics for viral infections. Logical errors contributed to 31.6% of incorrect responses and typically involved the misapplication of principles, such as recommending beta-blockers for asthmatic patients. Combined errors, involving both logical and informational deficiencies, represented 7.2% of the incorrect answers. The significant association between error type and response correctness ($\chi^2 = 127.54$, df = 3, $p < 0.0001$) highlights the importance of targeted refinements in CHAT GPT's domain-specific knowledge and reasoning frameworks. These findings align with prior research by Antaki et al, which observed similar challenges in ophthalmology-related tasks, where errors were often attributed to incomplete understanding of specialised concepts[23]. Comparable studies also highlight these limitations. For example, Gilson et al., identified that while CHAT GPT demonstrates strong logical reasoning, its domain-specific knowledge in areas such as medical licensing exam questions remains incomplete[14]. Similarly, Chen et al. reported that while CHAT GPT excels in general reasoning tasks, it faces challenges with more complex biomedical reasoning, mirroring the findings from this study[2].Unlike CHAT GPT, earlier models such as GPT-3.0 exhibited even more frequent logical errors due to their limited contextual understanding, as highlighted

https://doi.org/10.18549/PharmPract.2026.2. 3488

Table 1. CHAT GPT[a] Performance		
Response	Frequency	Percent
Correct	444	76.2
Incorrect	139	23.8
Total	583	100

[a]ChatGPT: Chat Generative Pre-Trained Transformer

by Jin et al., who found that neural network approaches achieved limited success on complex medical tasks[24]. These comparisons underscore the incremental but substantial advancements made by CHAT GPT in reducing errors related to logical reasoning while continuing to face challenges with information gaps. Addressing these deficiencies in future iterations could significantly enhance the accuracy and applicability of the model in medical disciplines.

Comparison with Similar Studies

The observed performance change over time of CHAT GPT, demonstrated by a statistically significant improvement of 23.7% upon re-evaluation of initially incorrect responses ($\chi^2 = 37.3$, $p < 0.0001$), aligns with findings from recent studies evaluating iterative refinement capabilities in LLMs. For instance, Gao et al. (2024) introduced a Self-Evolving GPT framework, enabling models to autonomously learn and improve their performance through iterative adjustments, showing substantial gains in various natural language processing tasks[25]. Similarly, Madaan et al. proposed the Self-Refine method, allowing LLMs to generate self-feedback and refine outputs iteratively. This approach significantly enhanced performance across diverse tasks, further highlighting the feasibility of improved response accuracy upon re-evaluation in LLMs[26]. These findings reinforce the importance of incorporating dynamic, real-time feedback mechanisms in LLMs to enable continuous learning and adaptation. While CHAT GPT's architecture exhibits modest potential for iterative improvement, it remains constrained by its reliance on static updates from OpenAI. Incorporating frameworks similar to those proposed by Gao et al.[25] and Madaan et al.[26] could significantly expand its utility and effectiveness in educational contexts.

Implications for Medical Education

CHAT GPT's ability to deliver immediate, evidence-based responses with logical explanations positions it as a valuable tool for interactive and self-directed learning. By simulating clinical scenarios and supporting problem-solving tasks, the model aligns with modern educational strategies that emphasize critical thinking and the application of knowledge in real-world settings. Additionally, CHAT GPT's accessibility and ease of use make it a practical resource for students in resource-limited environments. The model's potential to evaluate essays and provide feedback on grammar, structure, and content further broadens its applicability, enabling educators to focus on fostering deeper conceptual understanding.However, ethical considerations must be addressed, including the risks of over-reliance on AI tools and the potential propagation of misinformation. Educators should integrate CHAT GPT as a complement to traditional methods, ensuring that its outputs are critically evaluated and used to enhance, rather than replace, conventional teaching. Moreover, this study opens several avenues for future research. One promising direction is to empirically evaluate the impact of CHAT GPT-integrated learning on student performance through controlled interventions. For example, comparing pharmacology exam scores between students who engage in AI-guided, self-learning assignments and those following traditional study methods could provide valuable insight into its effectiveness. Longitudinal studies may also help assess whether sustained use of CHAT GPT fosters deeper conceptual understanding and critical thinking over time. Further research is needed to explore the educational effectiveness of CHAT GPT through direct student interaction and learning assessments. Additionally, its accessibility and ease of use may offer value in resource-limited environments.Beyond theoretical implications, ChatGPT could be practically deployed in medical curricula as an AI-assisted quiz practice tool for pharmacology students, a system for generating automated formative feedback on MCQs, and a resource for faculty to build or expand question banks efficiently. These applications could make pharmacology learning more interactive, support spaced repetition, and free up faculty time for higher-order teaching activities.

Limitations

While CHAT GPT's performance is promising in pharmacology-related MCQs, achieving a 76.2% accuracy rate with consistent logical reasoning and effective knowledge integration, several limitations warrant consideration. This study reflects a static evaluation of the model's capabilities at a specific point in time, with its performance expected to evolve alongside technological advancements and updates. Therefore, while some improvement was observed upon re-evaluation of previously incorrect responses, this study cannot confirm whether such changes reflect self-learning, external updates to the model, or random variation. To avoid misinterpretation, it is important to note that the 23.7% improvement seen in the re-evaluation of previously incorrect questions likely reflects variability in model outputs across sessions and possible backend updates from OpenAI, rather than true self-learning or cumulative knowledge acquisition. The lack of transparency in how generative models update or access information limits the ability to determine causality. This study did not stratify ChatGPT's performance by content domain or pharmacological sub-topic. Future research should explore which types of questions (e.g., pharmacokinetics vs. therapeutics) yield higher accuracy to better identify the model's strengths and limitations within the curriculum. Moreover, the dataset was confined to pharmacology-specific MCQs from the Pharmacology Examination and Board Review

Table 2. Qualitative reasoning characteristics across all responses

Metric	Overall (n=583), n (%)	Correct (n=444), n (%)	Incorrect (n=139), n (%)
Logical reasoning			
TRUE	565 (96.9)	444 (100.0)	121 (87.1)
FALSE	18 (3.1)	0 (0.0)	18 (12.9)
Internal information			
TRUE	581 (99.7)	444 (100.0)	137 (98.6)
FALSE	2 (0.3)	0 (0.0)	2 (1.4)
External information			
TRUE	565 (96.9)	435 (98.0)	130 (93.5)
FALSE	18 (3.1)	9 (2.0)	9 (6.5)
Reason for incorrect answer			
Logical error		--[b]	44 (31.6)
Information error		--	85 (61.2)
Logical and information errors		--	10 (7.2)
Statistical error		--	0 (0.0)

[a] ChatGPT: Chat Generative Pre-Trained Transformer.
[b] Not applicable

13th edition, limiting the generalisability of the findings to other areas of medical education. Furthermore, the exclusion of visual components, such as figures and diagrams commonly used in clinical training, leaves gaps in understanding the model's effectiveness in interpreting visual data.Although the reliance on standardised prompt engineering ensured consistency, it may not fully reflect the variability of real-world student interactions with AI tools, thereby limiting insights into its practical applications. Moreover, CHAT GPT employs a passive model update relying on periodic updates from OpenAI rather than dynamic, real-time feedback or interaction. This constraint reduces the model's adaptability to the evolving needs of learners and diverse educational contexts, highlighting the need for enhancements to its responsiveness and flexibility. It is important to note that while this study evaluated the performance of CHAT GPT, it did not investigate actual student learning outcomes. Future research should address how students interact with such tools, how they interpret correct versus incorrect responses, and what educational benefits or risks may result.

Future Directions

To fully harness GPT potential in medical education, future research should explore its applications in broader fields, including anatomy, pathology, and clinical case-based assessments, to assess its adaptability across diverse disciplines. Enhancing the model with multimedia capabilities to process visual data such as radiographs, histological slides, and anatomical images would address current limitations and increase its relevance in clinical training. Developing real-time feedback and adaptive learning frameworks could enable GPTs to refine responses dynamically, fostering a more engaging and personalised educational experience. These advancements would better align the model with evolving user needs and modern educational demands.Ethical considerations must guide its integration, ensuring that over-reliance on AI does not undermine critical thinking or problem-solving skills. Safeguards to prevent misinformation and clear guidelines for responsible use are essential for sustainable implementation. By addressing these challenges and advancing its capabilities, GPT could become a transformative tool in medical education,

Table 3. Performance Improvement After Re-Evaluation

Response Type	Initially Correct (First Attempt)	Improved Correct (Second Attempt)	Remained Incorrect
Questions	444	33	106

enriching learning experiences, supporting educators, and enhancing student outcomes.

CONCLUSIONS

CHAT GPT demonstrated significant promise as a tool for medical education, particularly in pharmacology. Achieving a high accuracy rate of 76.2%, the model consistently employs logical reasoning, integrates internal and external knowledge effectively, and provides immediate, evidence-based responses. These capabilities suggest potential for supporting interactive learning environments, particularly those that emphasize problem-solving and reflective thinking. By offering students an accessible and engaging resource, CHAT GPT supports mastery of complex pharmacological concepts.However, the model's reliance on static learning mechanisms and gaps in domain-specific knowledge highlight areas for improvement. Addressing these limitations could substantially enhance its versatility and reliability in medical education. With continued refinement, CHAT GPT has the potential to evolve into a virtual medical mentor, fostering deeper understanding, critical thinking, and knowledge retention for students and educators alike.

AUTHORS' CONTRIBUTIONS

Rima A. Hijazeen: conceptualization, data curation, formal analysis, investigation, methodology, project administration, supervision, validation, visualization, writing – original draft, writing – review and editing.

Al-Motassem Yousef : conceptualization, data curation, project administration, supervision, validation, writing – review and editing.

Ahmed Almousa: project administration, supervision, validation, writing – review and editing.

Aya N Alzoghair: conceptualization, data curation, project administration, supervision, validation, writing – review and editing.

Jude K. Dwairi: data curation, project administration, validation, writing – review and editing.

Majd I. Sawaqed: data curation, project administration, validation, writing – review and editing.

Ghaith F. Al-Ryahneh: data curation, project administration, validation, writing – review and editing.

Marwan H Ali: data curation, project administration, validation, writing – review and editing

CONFLICTS OF INTERESTS

The authors declare no competing interests

References

1. Biswas S. CHAT GPT and the Future of Medical Writing. Radiology. 2023;307(2):e223312. doi:10.1148/radiol.223312.
2. Chen Y, Zhao C, Yu Z, McKeown K, He H. On the relation between sensitivity and accuracy in in-context learning. Findings Assoc Comput Linguist EMNLP. 2023:155–167. doi:10.18653/v1/2023.findings-emnlp.12.
3. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good Boolean query for systematic review literature search? Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023:1426–1436. doi:10.1145/3539618.3591703.
4. Guo B, Zhang X, Wang Z, Jiang M, Nie J, Ding Y, Wang F, Chen J, Zhang S. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. arXiv [Preprint]. 2023.
5. OpenAI. Optimizing language models for dialogue. OpenAI Blog. 2022.
6. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Christiano P, Leike J, Lowe R. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022;35:27730–27744.
7. Moons P, Van Bulck L. ChatGPT: Can artificial intelligence language models be of value for cardiovascular nurses and allied health professionals? Eur J Cardiovasc Nurs. 2023;22(7):e9. doi:10.1093/eurjcn/zvad022.
8. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, Jiang D, Lee DH, Lee TH, Cheung R, Nguyen MH. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023;29(3):721–732. doi:10.3350/cmh.2023.0089.
9. Sng GGR, Tung JYM, Lim DYZ, Bee YM. Potential and pitfalls of ChatGPT and natural-language artificial intelligence models for diabetes education. Diabetes Care. 2023;46(5):e78–e80.
10. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. Am J Obstet Gynecol. 2023;228(6):696–705. doi:10.1016/j.ajog.2023.03.009.
11. D'Amico RS, White TG, Shah HA, Langer DJ. I asked ChatGPT to write an editorial about how we can incorporate chatbots into neurosurgical research and patient care…. Neurosurgery. 2023;92(4):663–664.
12. Hammer A. ChatGPT can pass the US medical licensing exam and the bar exam. Mail Online. 2023;23(23).

13. Amin Z, Khoo HE. Basics in Medical Education. Singapore: World Scientific Publishing; 2003.
14. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D, Hartman B, Moeller J, Kearney K, Xu Y. How does ChatGPT perform on the medical licensing exams? The implications of large language models for medical education and knowledge assessment. medRxiv [Preprint]. 2022.
15. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.
16. Gutiérrez BJ, McNeal N, Washington C, Chen Y, Li L, Sun H, Wang X, Lin C, Ji H, Xie Q, Yu X. Thinking about GPT-3 in-context learning for biomedical IE? Think again. Findings Assoc Comput Linguist EMNLP. 2022:4526–4541.
17. Logé C, Ross E, Yaw D, Dadey A, Jain S, Saporta A, Aloufi H, Alabi A, Ching C, Gordon T. Q-Pain: A question answering dataset to measure social bias in pain management. Harvard Dataverse. 2023.
18. Katzung BG, Trevor AJ. Pharmacology Examination & Board Review. 13th ed. New York: McGraw-Hill Education; 2021.
19. Buckwalter JA, Schumacher R, Albright JP, Cooper RR. Use of an educational taxonomy for evaluation of cognitive performance. J Med Educ. 1981;56(2):115–121.
20. Ha T, Yaneva V. Evaluating the performance of OpenQA on USMLE Step 1 and Step 2 questions. arXiv [Preprint]. 2019.
21. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. Information retrieval and neural networks for medical question answering: Performance on USMLE-style questions. arXiv [Preprint]. 2021.
22. Sharma A, Patel V, Singh H, Kumar N. ChatGPT performance on USMLE: A step toward AI-assisted learning. arXiv [Preprint]. 2023.
23. Antaki F, Cahill M, Gaudet V, Shah AS, Darvishian F. Evaluating ChatGPT in ophthalmology: Performance and limitations. JAMA Ophthalmol. 2023;141(7):577–584.
24. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci. 2021;11(14):6421.
25. Gao L, Xu M, Zhou X, Wang Y. Self-Evolving GPT: Autonomous improvement through iterative refinement. arXiv [Preprint]. 2024.
26. Madaan A, Muqeeth M, Yazdanbakhsh A, Chen X, Yao S, Zhou D. Self-Refine: Iterative refinement with self-feedback in large language models. arXiv [Preprint]. 2023.