Large Language Models Evaluation in Answering Multiple Choice Questions in Biochemistry Course

Olena Bolgova, Inna Shypilova, Volodymyr Mavrych

Submitted to: JMIR Medical Education
on: October 06, 2024

Table of Contents

Large Language Models Evaluation in Answering Multiple Choice Questions in Biochemistry Course

Olena Bolgova[1] MD, PhD; Inna Shypilova[2] MD, PhD; Volodymyr Mavrych[1] MD, PhD

[1]College of Medicine, Alfaisal University Riyadh SA
[2]School of Medicine, St Mathews University George Town KY

Corresponding Author:
Volodymyr Mavrych MD, PhD
College of Medicine, Alfaisal University
Al Takhassousi Str
Riyadh
SA

Abstract

Background: Recent advancements in artificial intelligence (AI), particularly in large language models (LLMs), have started a new era of innovation across various fields, with medicine at the forefront of this technological revolution. Many studies indicated that at the current level of development, LLMs can pass different board exams. However, the ability to answer specific subject-related questions requires validation.

Objective: The objective of this study was to conduct a comprehensive analysis comparing the performance of advanced LLM chatbots - Claude (Anthropic), GPT-4 (OpenAI), Gemini (Google), and Copilot (Microsoft), against the academic results of medical students in the medical biochemistry course.

Methods: We used 200 USMLE-style multiple-choice questions selected from the course exam database. They encompassed various complexity levels and were distributed across 23 distinctive topics. The questions with tables and images were not included in the study. The results of 5 successive attempts by Claude 3.5 Sonnet, GPT-4-1106, Gemini 1.5 Flash, and Copilot to answer this questionnaire set were evaluated based on accuracy in August 2024. Statistica 13.5.0.17 (TIBC® Statistica™) was used to analyze the data's basic statistics. Considering the binary nature of the data, the Chi-square test was utilized to compare results among the different chatbots, with a statistical significance level of P<.05.

Results: On average, the selected chatbots correctly answered 81.1±12.8% of the questions, surpassing the students' performance by 8.3% (P=.017). In this study, Claude showed the best performance in biochemistry MCQs, correctly answering 92.5% of questions, followed by GPT-4 (85.1%), Gemini (78.5%), and Copilot (64%). The chatbots demonstrated the best results in the following four topics: Eicosanoids (100%), Bioenergetics and Electron transport chain (96.4±7.2), Hexose monophosphate pathway (91.7±16.7), and Ketone bodies (93.8±12.5). The Pearson Chi-square test indicated a statistically significant association between the answers of all 4 chatbots (P<.001- P<.044).

Conclusions: Our study suggests that different AI models may have unique strengths in specific medical fields, which could be leveraged for targeted educational support in biochemistry courses. This performance highlights the potential of AI in medical education and assessment.

(JMIR Preprints 06/10/2024:67244)
DOI: https://doi.org/10.2196/preprints.67244

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Large Language Models Evaluation in Answering Multiple Choice Questions in Biochemistry Course

Olena Bolgova, Inna Shypilova, Volodymyr Mavrych *

Authors information:

O. Bolgova
College of Medicine, Alfaisal University, Kingdom of Saudi Arabia
(email: obolgova@alfaisal.edu).
https://orcid.org/0009-0002-9496-9754

I. Shypilova
School of Medicine, St Mathews University, Cayman Islands
(email: ishypilova@stmatthews.edu).
https://orcid.org/0009-0000-0707-6997

V. Mavrych*
Alfaisal University, College of Medicine, Kingdom of Saudi Arabia
(corresponding author email: vmavrych@alfaisal.edu).
https://orcid.org/0009-0009-1159-4573

*** -corresponding author**

Abstract

Background:

Recent advancements in artificial intelligence (AI), particularly in large language models (LLMs), have started a new era of innovation across various fields, with medicine at the forefront of this technological revolution. Many studies indicated that at the current level of development, LLMs can pass different board exams. However, the ability to answer specific subject-related questions requires validation.

Objective:

The objective of this study was to conduct a comprehensive analysis comparing the performance of advanced LLM chatbots - Claude (Anthropic), GPT-4 (OpenAI), Gemini (Google), and Copilot (Microsoft), against the academic results of medical students in the medical biochemistry course.

Methods:

We used 200 USMLE-style multiple-choice questions selected from the course exam database. They encompassed various complexity levels and were distributed across 23 distinctive topics. The questions with tables and images were not included in the study. The results of 5 successive attempts by Claude 3.5 Sonnet, GPT-4-1106, Gemini 1.5 Flash, and Copilot to answer this questionnaire set were evaluated based on accuracy in August 2024. Statistica 13.5.0.17 (TIBC® Statistica™) was used to analyze the data's basic statistics. Considering the binary nature of the data, the Chi-square test was utilized to compare results among the different chatbots, with a statistical significance level of *P<.05.*

Results:

On average, the selected chatbots correctly answered 81.1±12.8% of the questions, surpassing the students' performance by 8.3% (*P=.017*). In this study, Claude showed the best performance in biochemistry MCQs, correctly answering 92.5% of questions, followed by GPT-4 (85.1%), Gemini (78.5%), and Copilot (64%). The chatbots demonstrated the best results in the following four topics: Eicosanoids (100%), Bioenergetics and Electron transport chain (96.4±7.2), Hexose monophosphate pathway (91.7±16.7), and Ketone bodies (93.8±12.5). The Pearson Chi-square test indicated a statistically significant association between the answers of all 4 chatbots (*P<.001- P<.044*).

Conclusions:

Our study suggests that different AI models may have unique strengths in specific medical fields, which could be leveraged for targeted educational support in biochemistry courses. This performance highlights the potential of AI in medical education and assessment.

Keywords: *ChatGPT; Claude; Gemini; Copilot; Biochemistry; LLM; Medical Education; Artificial Intelligence;*

Introduction

Recent breakthroughs in artificial intelligence (AI), especially in large language models (LLMs), have started in a new era of innovation across diverse fields, with medicine leading the charge in this technological revolution. The integration of AI into various medical disciplines such as oncology, radiology, and pathology has demonstrated its advancing clinical uses and its potential to revolutionize healthcare delivery. [1, 2, 3]. As new LLMs continue to emerge and evolve, AI is poised to fundamentally reshape our understanding and approach to medicine, offering unprecedented opportunities for improved patient care, diagnostics, and medical education [4].

While academic interest in AI has surged in recent years, integrating AI technologies in educational settings, particularly medicine, has been uneven and fraught with challenges. Among many AI tools available, ChatGPT (Chat Generative Pretrained Transformer) has emerged as a potential game-changer in medical education [5, 6]. This sophisticated language model, powered by advanced neural networks, demonstrates a remarkable ability to interpret prompts and generate human-like responses, making it difficult to distinguish from human-produced language.

LLM's underlying transformer architecture enables it to excel in natural language understanding, continuously processing and adapting to new information. This adaptability, combined with its vast knowledge base, presents promising opportunities for enhancing teaching and learning methodologies in medical education [7]. AI-powered tools like ChatGPT may be particularly effective in addressing persistent challenges in student engagement, offering interactive and personalized learning experiences that traditional teaching methods often struggle to provide [8].

OpenAI's GPT-4 and GPT-3.5, Google's Gemini, and Anthropic's Claude have emerged as frontrunners, offering unique capabilities and potential medical education and practice applications. As of 2024, the AI landscape in healthcare has become increasingly diverse, with over 20 LLMs available for public use. Among them, four are the most promising.

Anthropic developed Claude, an AI assistant known for its strong natural language understanding and generation capabilities. It has been trained on a wide range of data and is designed to be helpful, harmless, and honest. Claude has shown particular strength in tasks requiring nuanced understanding and ethical reasoning [9].

Created by OpenAI, GPT-4 is the latest GPT (Generative Pre-trained Transformer) series iteration. It represents a significant advancement over its predecessor, GPT-3, with improved language understanding, generation, and reasoning capabilities. GPT-4 has demonstrated impressive performance across various domains, including coding, creative writing, and analytical tasks [10].

Gemini developed by Google AI, Gemini is a multimodal AI model capable of understanding and generating text, images, and other forms of data. It comes in different sizes and is optimized for various tasks and computational requirements. Gemini has shown strong performance in complex reasoning tasks and can understand context across different modalities [11].

Copilot created by GitHub in collaboration with OpenAI, Copilot is an AI pair programmer designed to assist developers by suggesting code completions and entire functions. It is now an integral part of Microsoft Windows. While primarily focused on coding tasks, Copilot's underlying language model has shown capabilities in understanding and generating natural language [12].

Recent studies have highlighted the potential of open-source LLMs in different domains, including medicine [1]. These models offer the advantage of transparency and customizability, allowing researchers and educators to fine-tune them for specific medical education needs. The performance of these models, particularly GPT-4, in handling medical queries has been remarkable. This aligns with a growing body of literature highlighting AI's potential in medical diagnostics and clinical practice. Recent studies have demonstrated how FDA-approved AI devices and algorithms increasingly integrate into medical workflows, reflecting a broader digital transformation in healthcare [1, 3].

For instance, AI has shown significant improvements in diagnostic accuracy for conditions such as breast cancer, where AI-assisted screening has enhanced early detection rates [13]. Moreover, AI's impact extends to interpreting complex medical data, such as echocardiograms and cardiac function assessments, where it has demonstrated high precision and consistency [14]. AI supports early lung cancer detection in oncology, potentially improving patient outcomes through timely interventions [15].

These innovations shift towards the integration of AI into medical practices, promising elevated efficiencies and enhanced patient care across different medical fields. As AI continues to evolve, its role in supporting clinical decision-making and augmenting medical professionals' capabilities is expected to grow exponentially [2].

One primary method for assessing the capabilities of LLMs in knowledge-based fields, including medicine, is their performance on multiple-choice tests [16]. The release of GPT-4 by OpenAI in 2023 marked a significant milestone, demonstrating impressive test-taking abilities across various domains [17]. Similarly, Claude 2 from Anthropic, released in June 2023, has gained attention for its ability to process larger input spaces (up to 100,000 tokens), potentially allowing for a more comprehensive analysis of medical texts and case studies [8].

An interesting study was recently published, which indicated a correlation between GPT-4's performance and the level of medical school courses. This finding suggests that the nature and complexity of the course content influence AI's effectiveness in providing accurate answers. For instance, GPT-4 demonstrated enhanced performance in Cell Biology, likely due to the structured and factual nature of the material, which aligns well with AI's strengths in processing and recalling large amounts of information. However, the study also revealed that AI's performance was lower in courses involving more complex clinical reasoning or practical skills. This discrepancy highlights AI's current limitations in replicating the nuanced decision-making processes and hands-on skills crucial in medical practice [18].

The findings from various studies position LLMs as a valuable supplementary tool in medical education, capable of enhancing learning experiences while acknowledging its limitations [1, 2, 4]. As AI technologies like ChatGPT continue to evolve, there is a pressing need for responsive research and professional development among medical educators. This ongoing evaluation and adaptation will be crucial in maximizing the benefits of AI in learning and teaching while minimizing potential risks.

The high accuracy demonstrated by ChatGPT-4 in answering multiple-choice questions (MCQs) compared to medical students' performance is particularly noteworthy. It suggests that AI could be an effective study aid, helping students review and reinforce their knowledge across various medical subjects. However, it is essential to view AI as a complementary tool rather than a replacement for multiple-choice questions have transformed from their conventional use as assessment tools to become a versatile educational approach in medical curricula. MCQs stimulate students' cognitive abilities and promote active interaction with study materials. By utilizing advanced generative AI-driven language models to address MCQs in medical physiology and other subjects, educators may provide students with an innovative and engaging learning experience, potentially enhancing their grasp of essential medical concepts. traditional teaching methods or human expertise [18].

Multiple-choice questions have transformed from their conventional use as assessment tools to become a versatile educational approach in medical curricula. MCQs stimulate students' cognitive abilities and promote active interaction with study materials. By utilizing advanced generative AI-driven language models to address MCQs in medical physiology and other subjects, educators may provide students with an innovative and engaging learning experience, potentially enhancing their grasp of essential medical concepts. [19]. Integrating AI in MCQ-based learning could allow for personalized question sets tailored to individual students' learning needs, adaptive difficulty levels, and immediate feedback mechanisms. This approach could significantly enhance the efficiency and effectiveness of self-directed learning in medical education [18, 19].

Recent studies have begun to compare the performance of different AI models in medical education contexts. For instance, Claude, an LLM developed by Anthropic, has shown promising results in solving medical MCQs. Some studies have indicated that Claude demonstrated a high frequency of right answers and explanations compared to ChatGPT-3.5 [8, 20]. These comparative studies are crucial in understanding the strengths and limitations of different AI models in medical education. They help educators and researchers identify the most suitable tools for specific learning objectives and contexts within medical curricula.

Despite the promising results, it is important to note the variability in AI performance across different studies and question types. For example, while some studies reported high accuracy rates for ChatGPT in physiology tests [5, 8], others found lower performance rates, particularly as the complexity and difficulty of questions increased [21, 22]. This variability underscores the need for careful consideration when integrating AI tools into medical education. Educators must be aware of these tools' strengths and limitations and ensure they are used appropriately to complement, rather than replace, traditional teaching methods.

As AI advances, future research should focus on exploring its applications in medicine more comprehensively. It includes investigating a broader array of AI models, especially those specialized in particular medical fields or capable of interpreting complex medical imagery. Expanding the diversity of medical subjects on which LLMs are trained is crucial, as well as longitudinal studies tracking the progress and adaptation of chatbots over time [1, 4]. They would provide valuable insights into their learning curves, accuracy improvements, and ability to integrate new medical knowledge. Such research could enhance LLMs' utility in clinical and educational settings, contributing to safer and more effective healthcare delivery [2].

It is important for educational strategies to prioritize the integration of LLMs into the curriculum as a vital aspect of the learning process. This integration should enable students to cultivate critical thinking and analytical skills, particularly in understanding the constraints of AI. LLMs have the potential to offer students in-depth knowledge and diverse viewpoints, facilitating a more thorough comprehension of intricate medical concepts. [23]. By utilizing the output of LLMs and working alongside educators to draw upon their existing knowledge, students can actively participate in the learning process. This collaborative approach allows for the refinement of their understanding and insights. The future of medical education depends on the seamless integration of human expertise with AI-powered tools. [3, 19, 23].

The objective of this study was to conduct a comprehensive analysis comparing the performance of advanced LLM chatbots - GPT-4, Claude, Copilot, and Gemini, against the academic results of medical students in biochemistry. This research sought to evaluate the capabilities of AI-powered chatbots across a diverse range of topics covered in a medical biochemistry curriculum.

Methods

Study design

This study focused on a comparative analysis of the capabilities of different AI-driven large language models in the medical biochemistry course. The research included an examination of four chatbots currently available to the public: Claude (Anthropic), GPT-4 (OpenAI), Gemini (Google), and Copilot (Microsoft).

200 scenario-based MCQs with four options and a single correct answer were randomly chosen from the medical biochemistry course's examination database for medical students and validated by two independent experts. The study did not include questions with images and tables. The selected questions encompassed various levels of complexity. They were distributed across 23 distinctive categories: Structural proteins and associated diseases, Globular proteins and hemoglobin, Red blood cells and anemias, Structure and function of amino acids, Structure and function of proteins, Bioenergetics and electron transport chain, Enzymes, Glycolysis and

gluconeogenesis, Glycogen, Signaling mechanisms, Pyruvate dehydrogenase and Krebs cycle, Cholesterol metabolism, Eicosanoids, Fatty acid metabolism, Fructose and Galactose metabolism, Hexose monophosphate pathway, Ketone bodies, Lipoproteins, Lysosomal storage diseases, Amino acid metabolism, Fast and fed state, Heme metabolism, Nitrogen metabolism.

Data collection

For the testing phase, each selected chatbot was required to answer a set of 200 questions, and their performance was evaluated against the responses provided by medical students for the same set of questions. Claude 3.5 Sonnet, GPT-4-1106, Gemini 1.5 Flash, and Copilot proficiency in responding to multiple-choice questions was assessed in the last two weeks of August 2024. An OpenAI paid subscription was obtained to get GPT-4 access.

The results of 5 successive attempts by each chatbot to answer this questionnaire set were evaluated based on accuracy. A total of 40,000 answers from LLMs were analyzed.

Five random answers were generated and analyzed for the same MCQ set utilizing the RAND () function in Microsoft Excel (Microsoft® 365) to compare chatbot results with random guessing.

Data analysis

The answers provided by each LLM were recorded and input into the Microsoft Excel spreadsheet (Microsoft® 365). The data from each (1-5) attempt was matched with the answer key and compared with all previous attempts, finding the percentage of repeated and correct answers among them. After that, a detailed item analysis was performed for each chatbot concerning different question categories.

Statistica 13.5.0.17 (TIBC® Statistica™) was used to analyze the data's basic statistics. Considering the data's binary nature, the Chi-square test was used to compare results among the different chatbots.

Results

According to our data, on average, four selected chatbots accurately answered 81.1±12.0% of 200 multiple-choice questions from the medical biochemistry course. This result was 8.3% (*P=.017*) above the students' average (72.8±12.7%) and almost four times better than randomly generated responses (22.0±2.9%) for the same questions.

There was a significant variation in correct responses among the chatbots. The best result was recorded for Claude (92.5±0.0%), followed by GPT-4 (85.1±1.0%) and Gemini (78.5±0.0%), which were better than the students’ average. Copilot showed the lowest result - 64.0±0.0% (Fig. 1).

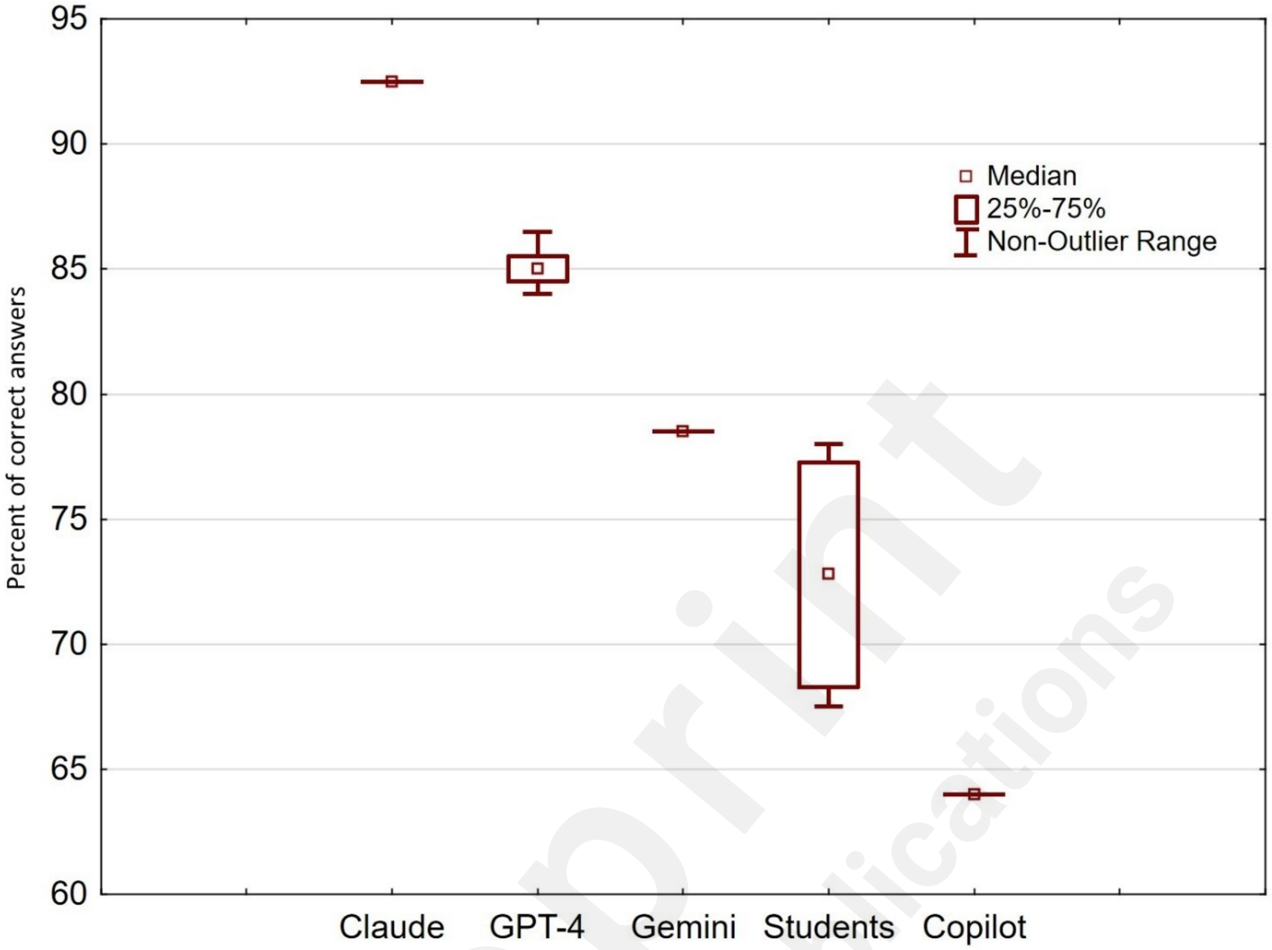


Fig. 1. Percentile of correct answers from different chatbots and students on 200 MCQs from medical biochemistry course.

Interestingly, all chatbots answered 104 questions (52%) correctly in all attempts. General item analysis revealed that Eicosanoids, Bioenergetics and Electron transport chain, Hexose monophosphate pathway, and Ketone bodies were the four best topics, with the average results for all chatbots being 100%, 96.4±7.2, 91.7±16.7 and 93.8±12.5, respectively.

In contrast, the lowest results were recorded for Globular Proteins and Hemoglobin - 58.4±26.4%, Lipoproteins - 64.6±20.3%, and Fructose and Galactose metabolism questions - 65.8±29.9%.

After that, each chatbot's results for all 23 topics were evaluated (Fig. 2).

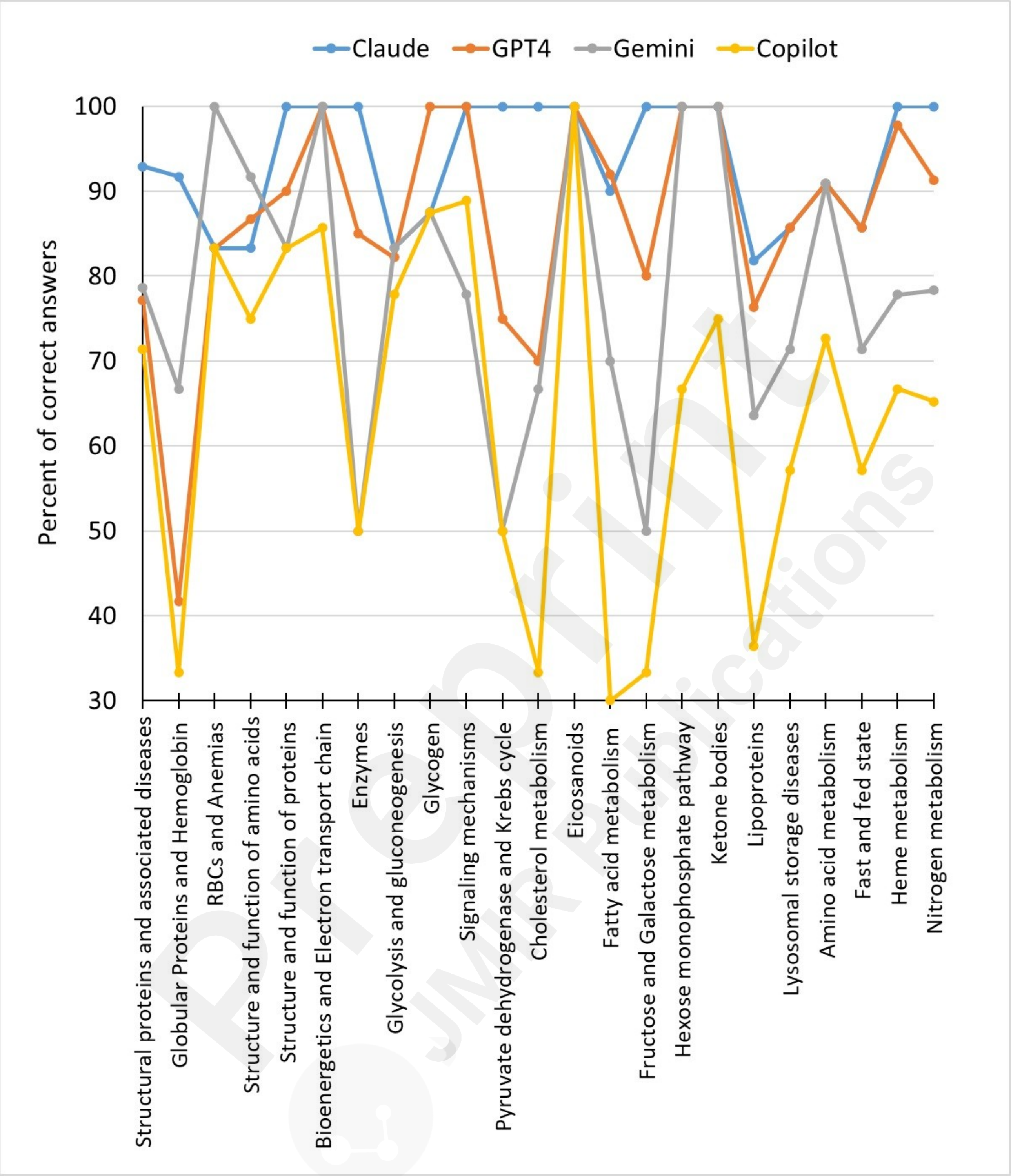


Fig. 2. Evaluation of chatbot performance in different topics of the medical biochemistry course.

Claude

Claude, offered by Anthropic, provided 92.5% correct answers to the set of 200 Biochemistry MCQs. The answers in this second and all subsequent attempts were identical to the first. As the chatbot claims, its knowledge base has not changed between attempts, and it applies the same reasoning to answer each question. It was the best result among the five chatbots, 19.7% better than the student average and 70.5% superior to the random guessing. The item analysis suggested that Claude correctly answered all questions (100%) from the following 12 categories: Structure and function of proteins, Bioenergetics and Electron transport chain, Enzymes, Signaling mechanisms, Pyruvate dehydrogenase and Krebs cycle, Cholesterol metabolism, Eicosanoids, Fructose and Galactose metabolism, Hexose monophosphate pathway, Ketone bodies, Heme metabolism, Nitrogen metabolism. The lowest result, 81.8%, was recorded for the Lipoproteins. For the rest of the topics, the percentile of correct answers was 83.3% - 91.7%.

Claude did not solve only 15 MCQs (7.5%) from the entire questionnaire set. These were comprehensive questions about red blood cells, hemoglobin, enzymes, biotin deficiency, and lipoproteins.

GPT-4

The results of five successive ChatGPT-4 (Open AI) attempts to answer the set of 200 Biochemistry MCQs showed 85.1±1.0% correct answers on average. The best result of its five attempts was 86.5%, 13.7% better than the average for medical students, and 64.5% above the random guessing. The fourth attempt was the most successful; the results of the other four attempts were close to 85% (84% - 85.5%).

Table 1 shows the coincidence generated by GPT-4 answers with the previous attempts; it was 91.5% - 94.5% among them, and the coincidence of correct answers was 81% - 83.6%.

TABLE 1

The results of 5 successive attempts of ChatGPT-4 to answer the set of 200 MCQs: correct answers, the coincidence of the answers with an earlier attempt, and the coincidence of the correct answers with a previous attempt (%)

Attempt number:	1	2	3	4	5
Correct answers	85.5	85	84	86.5	84.5
Coincidence with 1		92	91.5	93	90.5
Coincidence corrects with 1		82.5	81.5	83	81.5
Coincidence with 2			92	91	91.5
Coincidence corrects with 2			81.5	82.5	81.5
Coincidence with 3				94.5	93
Coincidence corrects with 3				83.6	81
Coincidence with 4					92
Coincidence corrects with 4					81

158 questions (79%) were answered correctly across all five attempts and considered a solid knowledge area for GPT-4. Most of these MCQs were recall questions, but some were complex and required critical thinking. The item analysis indicated that the best 6 categories with 100% correct answers were Bioenergetics and Electron transport chain, Glycogen, Signaling mechanisms, Eicosanoids, Hexose monophosphate pathway, and Ketone bodies. The lowest result was recorded for Globular Proteins and Hemoglobin questions - only 41.7% of the correct answers. For the rest of the topics, the percentile of correct answers was 77.1% - 97.8%.

GPT-4 did not answer only 17 MCQs (8.5%) from the entire questionnaire set in any one out of all five attempts. These were more comprehensive questions about defective proteins, oxygen saturation, anemia, amino acids, Glycogen, Glycolysis and gluconeogenesis, and lipoproteins.

Gemini

Google recently introduced Gemini as a successor to Bard. The results of five attempts by Gemini to answer the set of 200 Biochemistry MCQs showed 78.5% correct answers, 5.7% above the average for medical students, and 56.5% above the random answers. Unlike Bard, five successive attempts from Gemini were similar; the same answers were received.

The item analysis of these 157 correct answers shows that Gemini did the best (100% accurate) for questions in the following 5 categories: RBCs and Anemias, Bioenergetics and Electron transport chain, Eicosanoids, Hexose monophosphate pathway, and Ketone bodies. Most of these MCQs were recall questions. The lowest 50% results were recorded for the following 3 categories: Enzymes, Pyruvate dehydrogenase and Krebs cycle, Fructose and Galactose metabolism. Gemini's responses in other topics were at 63.6% - 91.7% interval. Gemini did not answer 43 MCQs (21.5%) from the entire questionnaire set, which were comprehensive questions mostly about proteins, enzymes, the Krebs cycle, fatty acid, fructose, and galactose metabolism.

Copilot

Microsoft's Copilot can accept only up to 2000 characters in the prompt, so only 2 to 7 MCQs can be answered at a time, which is inconvenient to work with. The results received on the first try were not different from four successive attempts, so there was zero variation among all 5 five attempts. Copilot generated 64% accurate answers for the same set of 200 MCQs from the Biochemistry course, 8.8% lower than the average medical student but 42% better than the random guessing.

The item analysis of these 126 correct answers indicated that these MCQs were mostly recall questions. The best result (100%) was shown for the Eicosanoids category and the lowest for Fatty acid metabolism - only 30% of correct answers. Copilot's responses in other topics vary from 33.3% to 88.9%. Copilot did not answer 72 MCQs (36%) from the questionnaire set. These questions concerned proteins, hemoglobin, amino acids, enzymes, fatty acids, pyruvate dehydrogenase, Krebs cycle, and fast and fed state.

Pearson Chi-square test results

Table 2 shows the results of the Pearson Chi-square test, which we employed due to the binary nature of the data to compare the performance of the different AI-driven chatbots against each other.

TABLE 2

Pearson Chi-square test results to compare the performance of Claude, GPT-4, Gemini, and Copilot against each other

LLMs	Chi-square	df	*P*
Claude x GPT-4	19.7253.5	1	0.00001
Claude x Gemini	6.085452	1	0.01363
Claude x Copilot	4.054054	1	0.04407
GPT4 x Gemini	33.07590	1	0.00000
GPT4 x Copilot	15.99813	1	0.00006
Gemini x Copilot	23.50351	1	0.00000

The null hypothesis was rejected because the p-value for all chatbots was less than alpha *P=.05*. So, there is a statistically significant association between the answers of all four chatbots.

Discussion

Principal Findings

Medical education is rapidly evolving, with artificial intelligence playing an increasingly significant role. In this context, evaluating AI efficacy and relevancy to results is crucial, particularly given the precision and depth of understanding required in medical practice. AI-driven LLMs like ChatGPT, Claude, Copilot, and Gemini have been compared against medical students in various studies, revealing both the strengths and limitations of AI in medical education. These comparisons show how AI can enhance human learning while also highlighting areas where it may not measure up. Research into AI's role in medical training has uncovered intriguing possibilities and important constraints. [1, 5, 7].

Multiple choice questions form a cornerstone of assessment in medical education. Analyzing these questions is vital as it allows educators to assess their effectiveness in testing higher-order thinking and clinical reasoning skills, ensuring that assessments accurately reflect the competencies required for medical practice [18]. While LLMs have demonstrated impressive capabilities in answering queries and simulating scenarios, the depth and breadth of their understanding, particularly concerning MCQs in medical exams, still requires thorough evaluation [19].

Recent studies have shown that LLMs, specifically GPT-4, often outperform medical students on MCQ items in board and licensing exams. This finding underscores the significance of MCQs in medical licensing exams, extensively utilized in crucial assessments worldwide. Examples include the Peruvian National Licensing Medical Examination, the United States Medical Licensing Examination (USMLE), the United Kingdom Medical Licensing Assessment (UKMLA), and the Australian Medical Council (AMC) Exam [20, 24, 25, 26]. The widespread use of MCQs is attributed to their effectiveness in evaluating higher-order skills through complex clinical scenarios, analysis, and problem-solving. These questions assess students' ability to integrate information, reflecting real-world challenges and shaping competent professionals. It is well correlated with the results of our study, which have shown that selected 4 chatbots answered correctly to 81.1±12.0% of 200 questions from the medical biochemistry course, which is 8.3% above the students' average.

Another comprehensive study compared the results of 4 LLMs across 163 questions from sample NBME clinical subject exams. The results were striking: GPT-4 achieved a perfect score of 100% (163/163), significantly outperforming GPT-3.5, Claude, and Bard. GPT-3.5 scored 82.21% (134/163), Claude 84.66% (138/163), and Bard 75.46% (123/163). The statistical superiority of GPT-4 was evident, with no significant differences observed among the other three models [27]. Interestingly, while GPT-4 excelled across all subject exams, the different models demonstrated variable strengths. GPT-3.5 performed best in family medicine and obstetrics/gynecology, Claude in surgery, and Bard in surgery and neurology. The surgery exam yielded the highest average score across all models, while family medicine had the lowest. GPT-4's exceptional performance may be attributed to its extensive training data, which exceeded 45 terabytes by September 2021, despite not being specifically fine-tuned for medical data [10].

Our data contradicts this clinical study and suggests that GPT-4 did well (85.1% correct answers) but is not currently the most proficient chatbot for biochemistry questions. The best result was recorded for Claude, with an impressive 92.5% of the correct answers. Gemini took third place with 78.5% of correct answers, which is still above the student's average (72.8±5.2%) for the same questions. The lowest result was recorded for Copilot - 64%.

These findings highlight the potential of LLMs in medical education and practice. Their ability to tackle complex medical questions opens doors to innovative clinical decision support, research, and education applications. However, it is worth noting that GPT-4, the only LLM in this study not available for free, could be less accessible to a broad range of students, potentially limiting its widespread use in educational settings.

Several studies have evaluated ChatGPT's performance in biochemistry. One study examined GPT-3.5's potential as a self-study adjunct for medical students in biochemistry, using 200 questions. ChatGPT provided correct answers to 58% of the biochemistry questions. While this performance allowed it to pass the university's medical biochemistry exam, the study suggests there is room for improvement in GPT-3.5 as a comprehensive and reliable self-learning tool [28].

Another study focused on ChatGPT's ability to address higher-order questions in medical biochemistry. Using GPT-3.5, researchers conducted an online cross-sectional study presenting 200 randomly selected, complex reasoning questions from an institutional question bank, classified according to CBME curriculum modules. Two expert biochemistry academicians evaluated responses on a 0-5 scale. The AI achieved a median score of 4.0 (Q1=3.50, Q3=4.50), which was comparable to a hypothetical value of 4 (p=0.16) but significantly lower than the maximum of 5 (p=0.001). These results suggest that GPT-3.5 shows promise as an effective tool for addressing complex questions in medical biochemistry, demonstrating its potential in handling higher-order thinking tasks in this field [29].

Our research confirms that GPT-4 has significant improvements and is superior to GPT-3.5. Our data suggest that GPT-4 responded right to 84% - 86.5% of MCQs, and 79% answered correctly across all five attempts.

The implications of AI's performance in medical education extend beyond mere test-taking abilities. LLMs can answer complex medical questions that raise important questions about the future of medical education and topics in which LLMs demonstrate proficiency, so they may be used to assist students. The detailed analysis of MCQs in our study revealed that questions from 4 topics are well answered by all chatbots: Eicosanoids, Bioenergetics, Electron transport chain, and Ketone bodies. In contrast, the lowest results were recorded for Globular Proteins and Hemoglobin, Lipoproteins, and Fructose and Galactose metabolism questions. However, there was a significant difference in the 4 LLMs performances. Claude showed the most impressive results and answered all questions (100%) from 12 categories: Structure and function of proteins, Bioenergetics and Electron transport chain, Enzymes, Signaling mechanisms, Pyruvate dehydrogenase and Krebs cycle, Cholesterol metabolism, Eicosanoids, Fructose and Galactose metabolism, Hexose monophosphate pathway, Ketone bodies, Heme, and Nitrogen metabolism.

In conclusion, the rapid advancements in AI technology, particularly in medical education, present opportunities and challenges. While LLMs have shown impressive capabilities in answering medical exam questions, it is crucial to remember that medical education encompasses more than just knowledge acquisition. Clinical skills, empathy, ethical decision-making, and the ability to navigate complex healthcare systems are all integral parts of medical training that current AI models may not fully capture.

As we progress, we must continuously evaluate AI's role in medical education, ensuring that it complements rather than replaces human expertise. The goal should be to harness AI's potential to enhance medical education and improve patient care while maintaining the critical human elements of healthcare practice. Future research should focus on integrating AI into medical curricula, using AI to personalize learning experiences, and preparing future healthcare professionals for an AI-augmented medical landscape.

Limitations

This study’s findings on different chatbot proficiencies are limited to multiple-choice questions from the biochemistry course, which may not represent other medical questions or contexts. In addition, the sample size of 200 questions, excluding questions with images or tables, may not capture the full range of difficulty levels or content areas.

LLMs receive regular updates, which result from training on inputs and tuning so that they may provide different answers depending on the testing date.

Another limitation is that GPT -4, which performed well, is not freely available, potentially limiting its applicability in widespread educational settings.

Conclusions

Large Language Models like ChatGPT, Claude, Copilot, and Gemini have impressive capabilities in answering MCQs, often outperforming medical students. In this study, the selected chatbots correctly answered 81.1±12.0% of 200 medical biochemistry questions, surpassing the students' average by 8.3% (*P=.017*). This performance highlights the potential of AI in medical education and assessment. Different LLMs exhibit varying strengths in different topics of medical biochemistry courses. In this study, Claude showed the best performance, answering 92.5% of questions correctly, followed by GPT-4 (85.1%), Gemini (78.5%), and Copilot (64%). This variability suggests that different AI models may have unique strengths in specific medical fields, which could be leveraged for targeted educational support. The strong performance of LLMs in answering complex medical questions raises important considerations for the future of medical education. While AI demonstrates proficiency in knowledge-based assessments, it is crucial to remember that medical training encompasses more than just information recall. Clinical reasoning, empathy, ethical decision-making, and navigating healthcare systems remain essential components that current AI models may need to capture fully.

Acknowledgments

The authors thank the Dean of the College of Medicine at Alfaisal University, Prof. Khaled Al-Kattan, and the Head of the Department of Anatomy and Genetics, Prof. Paul Ganguly, for their support in this research.

Conflicts of Interest

None declared.

References

1. Liu PR, Lu L, Zhang JY, Huo TT, Liu SX, Ye ZW: Application of Artificial Intelligence in Medicine: An Overview. Curr Med Sci. 2021, 41:1105-1115. 10.1007/s11596-021-2474-3
2. Garcia-Vidal C, Sanjuan G, Puerta-Alcalde P, Moreno-García E, Soriano A: Artificial intelligence to support clinical decision-making processes. EBioMedicine. 2019, 46:27-29. 10.1016/j.ebiom.2019.07.019
3. Ellahham S: Artificial Intelligence: The Future for. Diabetes Care. Am J Med. 2020, 133:895-900. 10.1016/j.amjmed.2020.03.033
4. Singhal K, Azizi S, Tu T, et al.: Large language models encode clinical knowledge [published correction appears in. Nature. 2023, 620:19-10. 10.1038/s41586-023-06291-2
5. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS: ChatGPT Knowledge Evaluation in Basic and Clinical Medical Sciences: Multiple Choice Question Examination-Based Performance. Healthcare (Basel. 2023:2046-2023. 10.3390/healthcare11142046
6. Bolgova O., Shypilova I., Sankova L., Mavrych V: How Well Did ChatGPT Perform in Answering Questions on Different Topics in Gross Anatomy?. European Journal of Medical and Health Sciences. 2023, 5:94-100.
7. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R: Artificial Intelligence in Medical Education: Comparative Analysis of ChatGPT, Bing, and Medical Students in Germany. JMIR Med

Educ. 20239, 46482-2023. 10.2196/46482

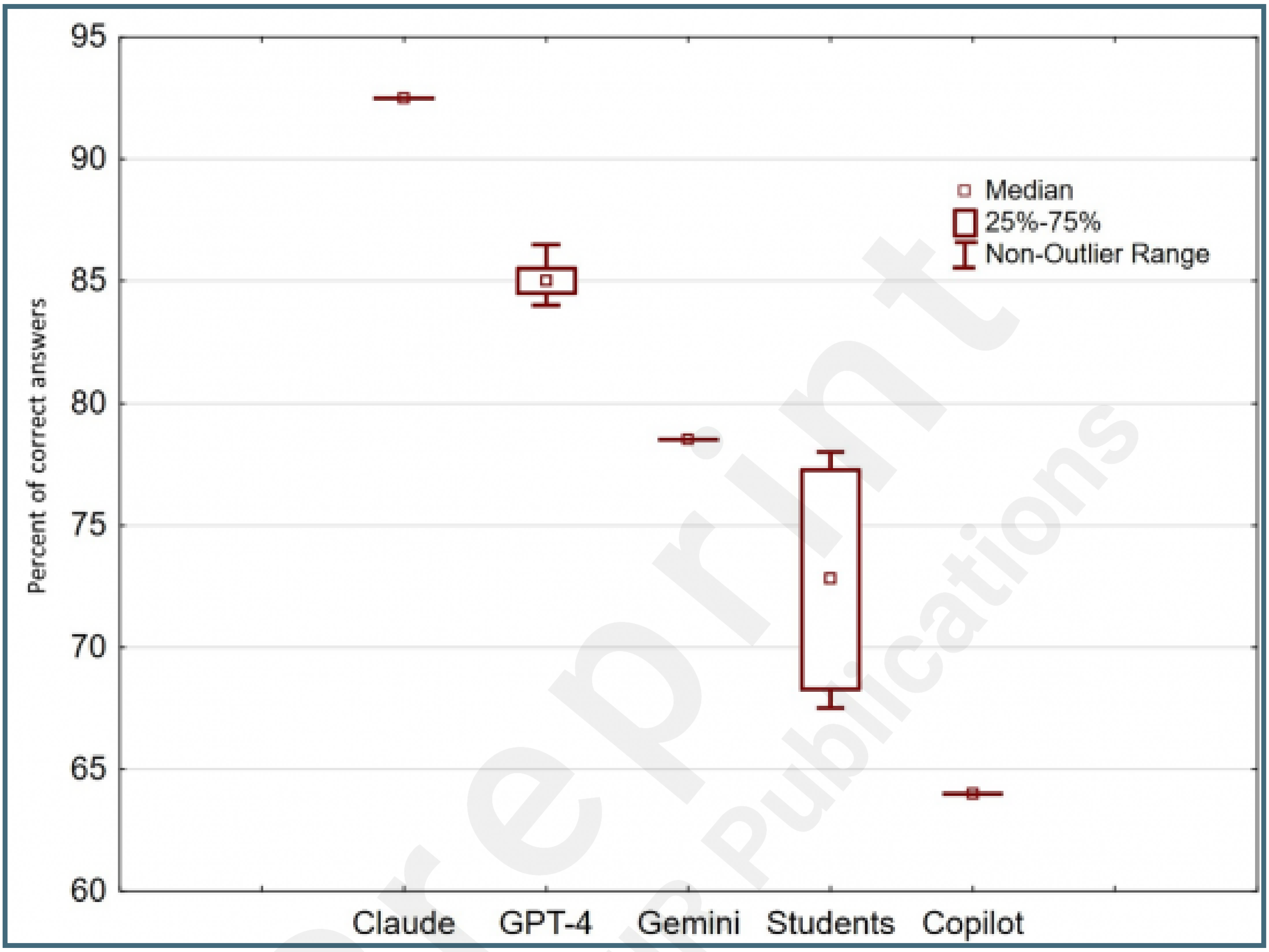
8. Agarwal M, Goswami A, Sharma P: Evaluating ChatGPT-3.5 and Claude-2 in Answering and Explaining Conceptual Medical Physiology Multiple-Choice Questions. Cureus. 2023159, 46222-2023. 10.7759/cureus.46222
9. Welcome to Claude. (2024). Accessed: September 6. (2024). https://docs.anthropic.com/en/docs/welcome.
10. GPT-4 Turbo and GPT-4. (2024) Accessed: September 6. (2024). https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4.
11. Gemini models. (2024) Accessed: September 6. (2024). https://ai.google.dev/gemini-api/docs/models/gemini.
12. Copilot for Microsoft 365. (2024) Accessed: September 6. (2024). https://learn.microsoft.com/en-us/office365/servicedescriptions/office-365-platform-service-description/microsoft-365....
13. Retamero JA, Gulturk E, Bozkurt A, et al.: Artificial Intelligence Helps Pathologists Increase Diagnostic Accuracy and Efficiency in the Detection of Breast Cancer Lymph Node Metastases. Am J Surg Pathol. 2024, 48:846-854. 10.1097/PAS.0000000000002248
14. Zhang J, Xiao S, Zhu Y, et al.: Advances in the Application of Artificial Intelligence in Fetal Echocardiography. J Am Soc Echocardiogr. 2024, 37:550-561. 10.1016/j.echo.2023.12.013
15. Chao HS, Tsai CY, Chou CW, et al.: Artificial Intelligence Assisted Computational Tomographic Detection of Lung Nodules for Prognostic Cancer Examination: A Large-Scale Clinical Trial. Biomedicines. 2023111, 147-2023. 10.3390/biomedicines11010147
16. Mavrych V, Bolgova O: Evaluating AI performance in answering questions related to thoracic anatomy. MOJ Anat Physiol. 2023, 10:55-59. 10.15406/mojap.2023.10.00339
17. Brin D, Sorin V, Vaid A, et al.: Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep. 2023, 13:16492. 10.1038/s41598-023-43436-9
18. Bharatha A, Ojeh N, Fazle Rabbi AM, et al.: Comparing the Performance of ChatGPT-4 and Medical Students on MCQs at Varied Levels of Bloom's Taxonomy. Adv Med Educ Pract. 2024, 15:393-400. 10.10.2147/AMEP.S457408
19. Goyal M, Agarwal M, Goel A: Interactive Learning: Online Audience Response System and Multiple Choice Questions Improve Student Participation in Lectures. Cureus. 2023157, 42527-2023. 10.7759/cureus.42527
20. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, et al.: Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: a cross-sectional study. J Educ Eval Health Prof. 2023, 20:30. 10.3352/jeehp.2023.20.30
21. Gilson A, Safranek CW, Huang T, et al.: How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment [published correction appears in. JMIR Med Educ. 20242710575941021965759420239, 45312-2023. 10.2196/45312
22. Friederichs H, Friederichs WJ, März M: ChatGPT in medical school: how successful is AI in progress testing?. Med Educ Online. 2023, 28:2220920. 10.1080/10872981.2023.2220920
23. Lin Z: Why and how to embrace AI such as ChatGPT in your academic life. R Soc Open Sci. 2023108, 230658-2023. 10.1098/rsos.230658
24. Kung TH, Cheatham M, Medenilla A, et al.: Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS digital health, 2, e0000198. 2023 (ed):
25. Lai UH, Wu KS, Hsu TY, Kan JK: Evaluating the performance of ChatGPT-4 on the United Kingdom Medical Licensing Assessment. Frontiers in medicine. 2023, 10:1240915.
26. Kleinig, O., Gao, C., & Bacchi, S. (2023: This too shall pass: the performance of ChatGPT-3.5, ChatGPT-4 and New Bing in an Australian medical licensing examination. The. Medical journal of Australia. 219:237.

27. Abbas, A., Rehman, M. S., & Rehman, S. S. : Comparing the Performance of Popular Large Language Models on the National Board of Medical Examiners Sample Questions. Cureus. 2024, 16:55991-10. 10.7759/cureus.55991
28. Surapaneni KM, Rajajagadeesan A, Goudhaman L, et al.: Evaluating ChatGPT as a self-learning tool in medical biochemistry: A performance assessment in undergraduate medical university examination. Biochem Mol Biol Educ. 2024, 52:237-248. 10.1002/bmb.21808
29. Ghosh A, Bir A: Evaluating ChatGPT's Ability to Solve Higher-Order Questions on the Competency-Based Medical Education Curriculum in Medical Biochemistry. Cureus. 2023154, 37023-2023. 10.7759/cureus.37023

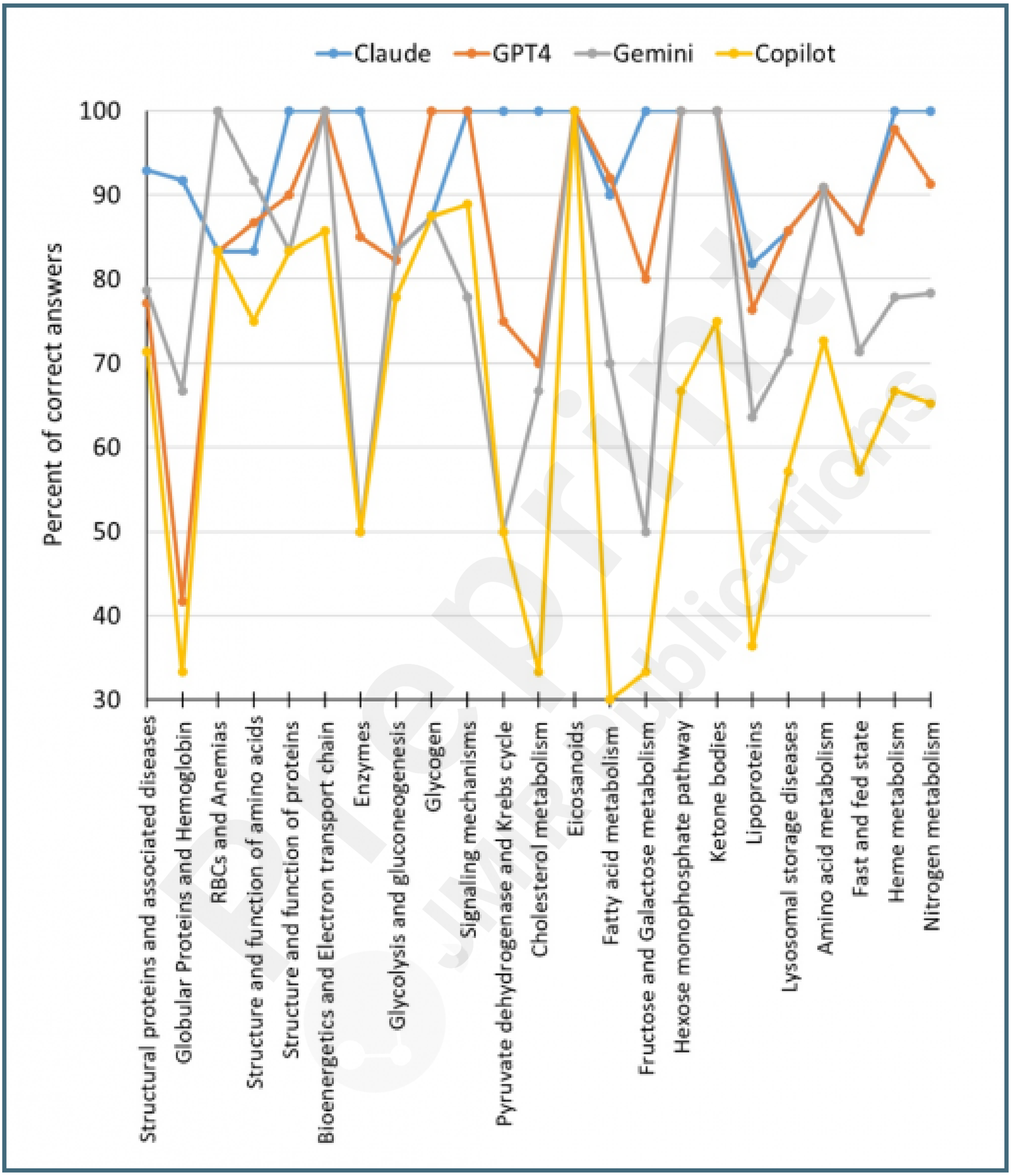
Supplementary Files

Figures

Percentile of correct answers from different chatbots and students on 200 MCQs from medical biochemistry course.



Evaluation of chatbot performance in different topics of the medical biochemistry course.


Claude
GPT4
Gemini
Copilot
Percent of correct answers
Structural proteins and associated diseases
Globular Proteins and Hemoglobin
RBCs and Anemias
Structure and function of amino acids
Structure and function of proteins
Bioenergetics and Electron transport chain
Enzymes
Glycolysis and gluconeogenesis
Glycogen
Signaling mechanisms
Pyruvate dehydrogenase and Krebs cycle
Cholesterol metabolism
Eicosanoids
Fatty acid metabolism
Fructose and Galactose metabolism
Hexose monophosphate pathway
Ketone bodies
Lipoproteins
Lysosomal storage diseases
Amino acid metabolism
Fast and fed state
Heme metabolism
Nitrogen metabolism