Paradigm Academic Press
Research and Advances in Education
ISSN 2788-7057
MAY. 2025 VOL.4, NO.4

Evaluating the Performance of Seven Large Language Models (GPT4.5, Gemini, Copilot, Claude, Perplexity, DeepSeek, and Manus) in Answering Healthcare Quality Management Inquiries

Dr. Mohammed Sallam[1,2,3,4], Dr. Johan Snygg[5,6], Dr. Ahmad Hamdan[3,7], Dr. Doaa Allam[1], Dr. Rana Kassem[1] & Dr. Mais Damani[1]

[1] Department of Pharmacy, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai P.O. Box 505004, United Arab Emirates

[2] Department of Management, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai P.O. Box 505004, United Arab Emirates

[3] Department of Management, School of Business, International American University, Los Angeles, CA 90010, United States of America

[4] College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU), Dubai P.O. Box 505055, United Arab Emirates

[5] Department of Management, Mediclinic City Hospital, Mediclinic Middle East, Dubai P.O. Box 505004, United Arab Emirates

[6] Department of Anesthesia and Intensive Care, University of Gothenburg, Sahlgrenska Academy, 41345 Gothenburg, Sweden

[7] Department of Nursing, Mediclinic Welcare Hospital, Mediclinic Middle East, Dubai P.O. Box 31500, United Arab Emirates

Correspondence: Dr. Mohammed Sallam, Department of Pharmacy, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai P.O. Box 505004, United Arab Emirates; Department of Management, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai P.O. Box 505004, United Arab Emirates; Department of Management, School of Business, International American University, Los Angeles, CA 90010, United States of America; College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU), Dubai P.O. Box 505055, United Arab Emirates.

doi:10.63593/RAE.2788-7057.2025.05.005

Abstract

Large language models (LLMs) are increasingly utilized across education, healthcare, and decision support due to their advanced text processing capabilities. This study evaluated the performance of seven LLMs: ChatGPT4.5, Gemini 2.5 Pro, Copilot, Claude 3.7, Perplexity, DeepSeek, and Manus in answering multiple-choice questions related to healthcare quality management. The assessment included 20 validated questions across four domains: organizational leadership (n = 5), health data analytics (n = 5), performance improvement (n = 5), and patient safety (n = 5). Accuracy rates ranged from 70% to 80%, with ChatGPT4.5, Gemini, and Claude achieving 80%; Perplexity and Manus, 75%; and Copilot and DeepSeek, 70%. All models met or exceeded the predefined accuracy threshold of 70%. Descriptive statistics showed a mean of 15.19 correct responses (SD = 0.83) and 5.00 incorrect responses (SD = 0.85) per model, with a combined average of 12.71 responses (SD = 4.46). A Pearson chi-square test indicated no statistically significant differences in accuracy among the models, χ^2 (6, N = 140) = 1.321, P =.971. A Monte Carlo simulation with 10,000 sampled tables

confirmed this result (*P* =.984, 95% CI).

The findings indicated comparable performance across the evaluated AI models in the context of healthcare quality education. These results support the use of large language models as supplementary tools in this domain, while highlighting the need for further evaluation of performance across specific content domains and their applicability in real-world professional training contexts.

Keywords: AI, artificial intelligence, LLMs, healthcare quality management, education tools, multiple-choice questions

1. Introduction

AI-based Large Language Models (LLMs) demonstrate strong potential across various fields and applications due to their capability to deliver detailed responses to complex inquiries (Li, 2023; Ostapuk & Audiffren, 2024). In healthcare and education, AI-based models are also gaining attention for delivering timely, concise, and actionable information in response to standard query requirements (Hristidis et al., 2023; Khlaif et al., 2023; Kung et al., 2023; Sallam, 2023). Notably, LLMs have demonstrated strong performance on professional and academic examinations originally designed for human respondents (A-Abbasi et al., 2024; Johnson et al., 2023; Sallam et al., 2024a; Sallam et al., 2024b).

Despite these advances, concerns remain regarding the precision and reliability of healthcare knowledge generated by LLMs, which poses a critical challenge to their integration into medical education, quality improvement, and other clinical domains (Benítez et al., 2024).

2. Background

Quality management practices within healthcare organizations continue to evolve through the adoption of innovative tools aimed at enhancing service standards (Sallam, 2024); However, historically, healthcare systems have relied on fragmented approaches to quality improvement, resulting in significant variability in workforce competencies, governance structures, and evidence-based practices (Connor et al., 2023; Spath, 2013). To mitigate these challenges, competency-based education in healthcare quality has emerged to strengthen professional knowledge and align with quality goals (Imanipour et al., 2022). Similarly, efforts to standardize quality improvement competencies have extended to integrating quality improvement into healthcare professionals' education (Myers et al., 2022). As Kazana and Dolansky (2021) noted, this integration addresses the alignment of quality initiatives across the healthcare sector.

In addition, the Joint Commission International Accreditation (JCIA) has gained wide application for its emphasis on quality evaluation, patient safety, and enhanced medication management care standards (Alraimi & Al-Nashmi, 2024; Sallam & Hamdan, 2023; Zabin et al., 2024). Its emphasis on aligning healthcare practices with international standards has attracted growing attention from governments, providers, and consumers as a key driver for improving healthcare quality and safety outcomes (Spath, 2013).

Despite such progress, the inherent complexity of healthcare quality assurance, which spans clinical and administrative functions, emphasizes the need for advanced tools, including artificial intelligence, to support improvement initiatives (Sallam et al., 2024c).

In light of these developments, this study aimed to evaluate the potential of seven AI-based LLMs (ChatGPT4.5, Gemini, Copilot, Claude, Perplexity, DeepSeek, and Manus) as supplementary tools in healthcare education and quality management. The evaluation focused on each model's accuracy in answering multiple-choice questions related to core healthcare quality domains (Chen et al., 2022; Newton & Xiromeriti, 2024; Sallam et al., 2024a).

3. Materials and Methods

3.1 Selection of AI Models

Over a year and nine months have passed since the release of GPT-4 on March 14, 2023, and as of 2025, more advanced large language models are now publicly available.

ChatGPT4, Gemini, Copilot, Claude, and Perplexity are considered AI-powered models or generative AI tools designed to process and generate natural language responses. DeepSeek is a new Chinese chatbot powered by artificial intelligence, designed to resemble and function similarly to ChatGPT in appearance and user experience (Sallam et al., 2025). Manus AI, introduced in early 2025, marks a development in autonomous general-purpose artificial intelligence, engineered to perform a wide range of tasks with limited human input (Shen & Yang, 2025). These tools leverage advanced natural language processing (NLP) techniques to assist with various tasks, including answering questions, providing explanations, and generating content (Liu et al., 2024a). They represent specific implementations of generative AI tailored for conversational and problem-solving applications.

3.2 Selection of Healthcare Quality Management Questions

The preparation of questions was guided by the need to comprehensively cover the four key domains of healthcare quality as defined by widely accepted professional standards and best practices in the field (Brandrud et al., 2017; Weheba et al., 2020). These domains included organizational leadership, health data analytics, performance and process improvement, and patient safety. The MCQ questions were utilized with written permission from Ahmad (2022). A collection of 20 questions was selected stepwise (Salam et al., 2020). The selected questions were then validated through a two-step process: first, by a panel of three certified healthcare quality and patient safety experts who reviewed the content for accuracy, relevance, and alignment with real-world scenarios, and second, by piloting the questions with a sample group of healthcare professionals to ensure clarity and applicability (Gottlieb et al., 2023). This validation process ensured the questions provided a solid foundation for evaluating the capability of AI models to address the selected healthcare quality domains (Ali & Zahra, 2024). Table 1 presents the validated multiple-choice questions with correct answers and associated healthcare quality domains.

Table 1. Details and domains of the ten questions

Question number	**Question details and MCQ options**	**Correct answer**	**Domain**
1	The leadership style that is said to motivate employees and optimize the introduction of change is: A. Autocratic B. Consultative C. Participatory D. Democratic	C. Participatory	Organizational Leadership
2	Which of the following is most important to the successful implementation of quality improvement activities? A. Financial commitment and written quality management plan B. Leadership commitment and organization-wide collaboration C. Leadership commitment and financial commitment D. Information management system and department collaboration	B. Leadership commitment and organization-wide collaboration	Organizational Leadership
3	Strategic leadership is linked to success in meeting: A. Budget Requirements B. Intended Objectives C. Governing Body Policy D. Contract Requirements	B. Intended objectives	Organizational Leadership
4	In a crisis, when a manager must make a rapid decision, the most effective leadership style is: A. Consultative B. Participatory C. Autocratic D. Democratic	C. Autocratic	Organizational Leadership
5	Leadership during a lengthy period of crisis in the organization is: A. Based on the leader's position in the organization. B. A participative activity performed by anyone committed to lead. C. Dependent on a set of personal characteristics. D. An autocratic style with decisions made solely by the leader.	B. A participative activity performed by anyone committed to lead	Organizational Leadership
6	Which of the following is most helpful in integrating data collected in the performance improvement process? A. Discussing performance improvement findings with senior management B. Developing a performance improvement prioritization matrix C. Creating a scatter diagram using the data D. Integrating the data based on team consensus	D. Integrating the data based on team consensus	Health Data Analytics

7	Evaluating the length of stay & outcome data on cardiac catheterization reveals a direct relationship between adverse outcomes & physician practice patterns. This integrated approach involves correlating: A. Case/Care management & finance B. Utilization & quality management C. Finance & utilization management D. Discharge planning & quality improvement	B. Utilization & quality management	Health Data Analytics
8	A surgeon's wound infection rate is 34%. Further examination of which of the following data will provide the most useful information in determining the cause of this surgeon's infection rate? A. Use of prophylactic antibiotics B. Type of anesthesia used C. Mortality rate D. Facility infection rate	A. Use of prophylactic antibiotics	Health Data Analytics
9	The Critical Care QI Team is chartered to improve the admission process to the critical care units. One identified issue, based on preliminary data, relates to admissions by family practice physicians. The medical director drafts the performance measures and criteria for data collection. The critical care nurses collect the data, and the quality management department staff aggregates and displays the data for the team. What key step is missing? A. Collaboration with the medical staff executive committee and family practice department B. Approval of the project by the family practice department C. Data collection and summarization by the medical staff D. Preliminary information proving that assessment is needed	A. Collaboration with the medical staff executive committee and family practice department	Health Data Analytics
10	Based on most quality improvement standards, those responsible for prioritizing data collection to monitor organization-wide performance are: A. The quality council B. The leaders C. Those most knowledgeable about the process D. Those most experienced with statistical analysis	B. The leaders	Health Data Analytics
11	The best evaluation of a performance improvement plan is: A. Process improvement B. Measurable objectives C. Applicable deliverables D. Timeline	B. Measurable objectives	Performance Improvement
12	Quality performance improvement focused on: A. Process B. System C. Individual D. Steps	B. System	Performance Improvement
13	Which of the following is most helpful in integrating data collected in the performance improvement process? A. Discussing performance improvement findings with senior management B. Developing a performance improvement prioritization matrix C. Creating a scatter diagram using the data D. Integrating the data based on team consensus	D. Integrating the data based on team consensus	Performance Improvement
14	All of the following conditions contribute to system improvement except: A. Measuring the performance of processes and their outcomes using valid statistical methods	D. Identifying and responding to individual performance	Performance Improvement

	B. Taking action to improve the way the processes are designed and carried out C. Studying and understanding the complex process that contributes to care D. Identifying and responding to individual performance issues	issues	
15	The best way to evaluate the effectiveness of performance improvement training is through: A. Self-assessment B. Participants' feedback C. Observed behavioral changes D. Post-test results	C. Observed behavioral changes	Performance Improvement
16	Which of the following national patient safety goals applies to everyone in a health care facility? A. Communication B. Medication safety C. Healthcare-associated infection D. Reconcile medication	A. Communication	Patient Safety
17	In assessing the patient safety culture, what should a quality professional do? A. Survey of all employees and physicians B. Survey patients' last 6 months C. Review collected data through incident reporting D. Review post-surgical infection rate data	A. Survey of all employees and physicians	Patient Safety
18	Which of the following are attributes of the culture of safety? A. Transparency & increased patient acuity level B. Error–proof of environment & empowered staff C. Empowered staff & transparency D. Increased patient acuity level & error-proof environment	C. Empowered staff & transparency	Patient Safety
19	Which of the following additional information should be in a patient safety plan? A. Disaster preparedness B. Steps to improve patient satisfaction C. Equipment management D. Efforts to reduce harm	D. Efforts to reduce harm	Patient Safety
20	The most effective way to ensure patient safety as a dimension of performance is to: A. Sponsor a "hotline" for reporting problems B. Focus on processes and minimize individual blame C. Have leaders who commit to and foster a safe culture D. Encourage patients and families to identify risks	C. Have leaders who commit to and foster a safe culture	Patient Safety

3.3 Data Collection and Analysis

The outputs from the seven AI models were documented for each question answered. Data were collected to enable empirical comparison and ensure response consistency across all models. Annotation was performed using Microsoft Excel by categorizing each AI-generated response as 1 (correct) or 2 (incorrect) based on evaluation with the predefined correct answers reviewed and agreed upon by three certified healthcare quality experts. The 20 multiple-choice questions covered four core domains: organizational leadership (n = 5), health data analytics (n = 5), performance improvement (n = 5), and patient safety (n = 5).

To ensure procedural consistency and eliminate input bias, the full set of twenty questions was submitted to each of the seven large language models (ChatGPT4o, Gemini, Copilot, Claude, Perplexity, DeepSeek, and Manus) using the same structured prompt, on the same day (11th May 2025), and in a standardized order. Each full question set was also entered in a new, independent chat session for each model to prevent prior conversation context from influencing responses. The prompt used was:

"*Act as an expert in healthcare quality management and carefully select the most accurate answer (from options A, B, C, or D) for each of the following multiple-choice questions; present your responses clearly in a table format with columns labeled 'Question number,' 'Selected answer,' and 'Answer details'.*"

All questions were submitted in English to maintain consistency across models. This standardized input protocol enabled a fair, replicable, and unbiased model performance evaluation.

The responses were assessed against predefined correct answers. Data processing and analysis were conducted using Microsoft Excel 2021 and IBM Statistical Package for the Social Sciences (SPSS) Version 30.0. Descriptive statistics were used to calculate the mean and standard deviation of correct and incorrect responses. The accuracy of the AI models was assessed against a target passing score of 70%, determined as an internal benchmark agreed upon by the research team to reflect a minimum threshold for acceptable performance. A Pearson chi-square test of independence was conducted to examine statistical differences in accuracy across the models. A Monte Carlo simulation with 10,000 sampled tables was also performed to confirm the robustness of the result, given small expected frequencies in several cells. Statistical significance was defined as $P < 0.05$.

3.4 Ethical Considerations

This study did not involve human participants, identifiable personal data, or clinical interventions. The evaluation focused exclusively on the performance of publicly accessible AI models using predefined and validated educational content.

4. Results

The performance evaluation of the seven generative AI models demonstrated high and relatively consistent accuracy across models. ChatGPT4.5, Gemini 2.5 Pro, and Claude 3.7 each achieved the highest accuracy rate of 80 percent. Perplexity and Manus followed with 75 percent, while Copilot and DeepSeek each recorded 70 percent (see Table 2 and Figure 1).

Table 2. Performance of Generative AI LLMs in Answering Healthcare Quality Questions

Generative AI Chabot	**Answer**	**N (%)**
ChatGPT4.5	Correct	16 (80%)
	Incorrect	4 (20%)
Gemini 2.5 Pro	Correct	16 (80%)
	Incorrect	4 (20%)
Copilot Think Deeper	Correct	14 (70%)
	Incorrect	6 (30%)
Claude 3.7 Sonnet Answer	Correct	16 (80%)
	Incorrect	4 (20%)
Perplexity	Correct	15 (75%)
	Incorrect	5 (25%)
DeepSeek R1	Correct	14 (70%)
	Incorrect	6 (30%)
Manus	Correct	15 (75%)
	Incorrect	5 (25%)

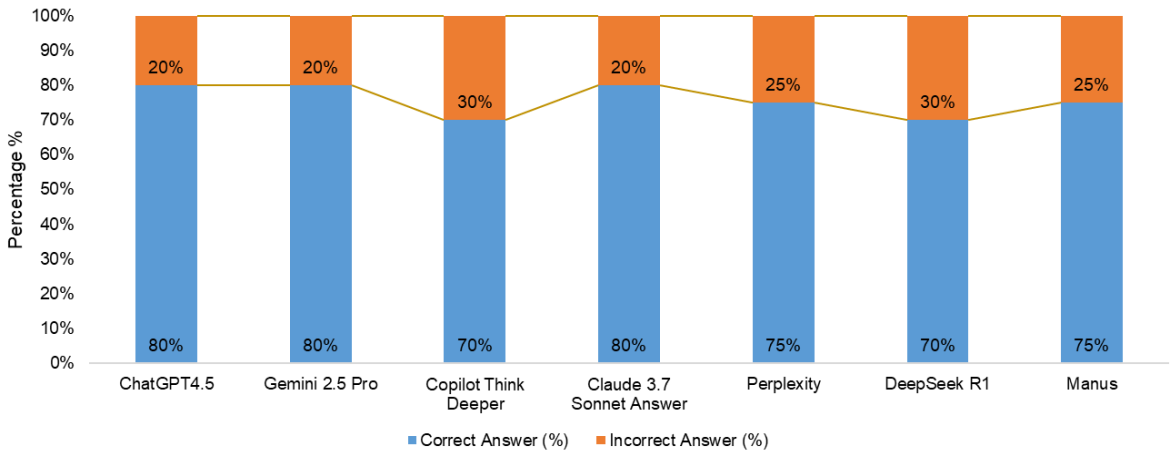

Figure 1. Performance Comparison of Generative AI Models Answering Healthcare Quality Management Questions

Figure 2 illustrates the achieved accuracy scores of each chatbot model relative to the target passing score of 70 percent, showing minor performance differences across models. ChatGPT4.5, Gemini 2.5 Pro, Claude 3.7 Perplexity, and Manus exceeded the benchmark, with the first three achieving the highest score of 80 percent.

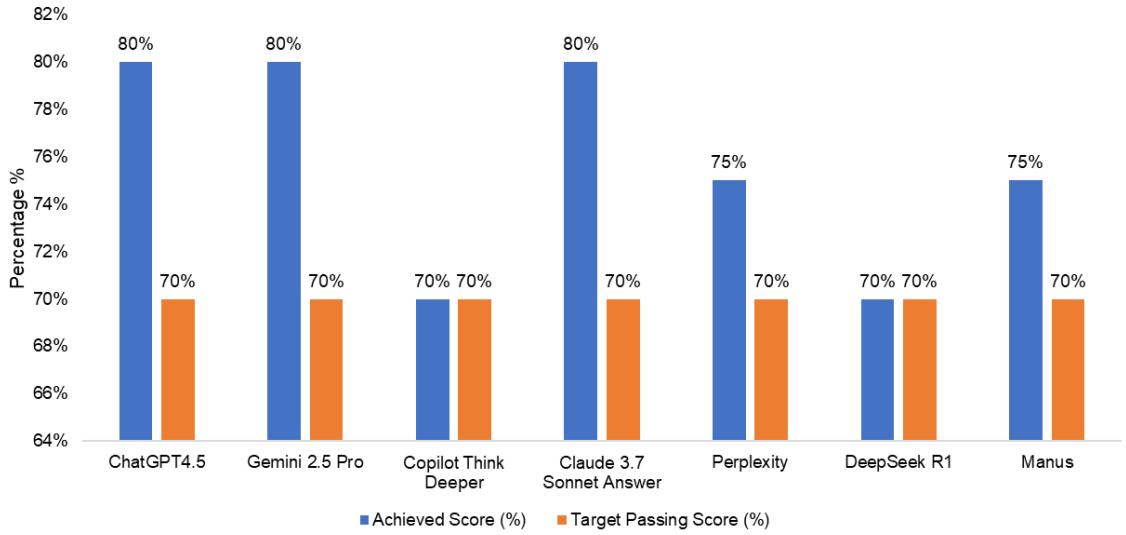

Figure 2. Comparison of Achieved Scores (%) and Target Passing Scores (%) for Generative AI Chatbots in Answering Healthcare Quality Management Questions

Descriptive statistics were used to summarize the responses by answer type across the evaluated large language models (see Table 3). The mean number of correct responses (Answer type 1) was 15.19 (SD = 0.829), while the mean number of incorrect responses (Answer type 2) was 5.00 (SD = 0.853). These findings reflect the overall higher accuracy trend observed across the chatbots, with a total average of 12.71 responses per category (SD = 4.463) out of 140 responses.

Table 3. Mean Number of Responses by Answer Type (Correct vs. Incorrect) Across All Chatbots

Answer response type	Mean	N	Standard Deviation (SD)
1	15.19	106	.829
2	5.00	34	.853
Total	12.71	140	4.463

N: Number of responses.

In addition to overall performance, Table 4 presents the number of correct answers per domain for each model. While all models achieved full scores in the organizational leadership domain, variation was observed in health data analytics, performance improvement, and patient safety, particularly with lower accuracy in selected safety and data interpretation questions.

Table 4. Model Accuracy by Domain (Number of Correct Answers out of 5 per Domain)

Model	**Organizational Leadership**	**Health Data Analytics**	**Performance Improvement**	**Patient Safety**	**Total (out of 20)**
ChatGPT4.5	5/5	4/5	4/5	3/5	16/20
Gemini 2.5 Pro	5/5	4/5	4/5	3/5	16/20
Copilot Think Deeper	5/5	3/5	3/5	3/5	14/20
Claude 3.7 Sonnet	5/5	4/5	4/5	3/5	16/20
Perplexity	5/5	3/5	3/5	4/5	15/20

DeepSeek R1	5/5	3/5	3/5	3/5	14/20
Manus	5/5	4/5	3/5	3/5	15/20

A Pearson chi-square test of independence was conducted to examine the relationship between the type of generative AI chatbot and answer accuracy (correct vs. incorrect). The result was not statistically significant, χ^2 (6, N = 140) = 1.321, *P* = .971, indicating comparable performance across the seven chatbot models. To ensure robustness of the result due to small expected frequencies in several cells, a Monte Carlo simulation based on 10,000 sampled tables was performed. The simulation confirmed the finding, yielding a non-significant *P*-value of .984 with a 95% confidence interval. These results indicate no significant differences in accuracy among the evaluated AI models (see Table 5).

Table 5. Chi-Square and Monte Carlo Test Results for Differences in Accuracy Among Generative AI Models

				Monte Carlo Sig. (2-sided)		
					95% Confidence Interval	
	Value	df	*P*-value	Significance	Lower Bound	Upper Bound
Pearson Chi-Square	1.321[a]	6	.971	.984[b]	.981	.986
Likelihood Ratio	1.314	6	.971	.984[b]	.981	.986
Fisher-Freeman-Halton Exact Test	1.462			.984[b]	.981	.986
N of Valid Cases	140					

a. 7 cells (50.0%) have expected count less than 5

b. Based on 10000 sampled tables with starting seed 2000000

df: Degrees of freedom

5. Discussion

To the best of our knowledge, this is the first study to evaluate the performance accuracy of seven publicly available large language models across four domains of healthcare quality management.

5.1 AI Chatbot Accuracy and Domain-Specific Performance

The performance evaluation revealed that all seven generative AI chatbots met or exceeded the target passing score of 70%. However, a comparison of generative AI models in answering questions demonstrated slightly varying levels of accuracy across the evaluated models (Liu et al., 2024b). ChatGPT4.5, Gemini 2.5 Pro, and Claude 3.7 Sonnet Answer achieved the highest accuracy with 80% correct responses, followed by Perplexity and Manus at 75%, while Copilot Think Deeper and DeepSeek R1 scored 70%. These results showed that the top-performing models slightly outperformed others, but the differences were minor.

The results suggest that the evaluated LLMs exhibit similar accuracy levels when applied to healthcare quality management multiple-choice questions, with no statistically significant variation in their performance. These findings indicate that LLMs, regardless of brand or model, may provide comparable levels of assistance in educational settings. However, while not statistically significant, the observed accuracy differences suggest areas where specific models may require further refinement to enhance reliability in healthcare decision-making.

The analysis revealed that patient safety was the most challenging domain, with all models struggling on multiple questions, suggesting limitations in their ability to interpret complex safety-related concepts. Performance improvement and health data analytics also showed inconsistencies, with several models providing incorrect responses, particularly in data-driven and quality improvement scenarios. In contrast, organizational leadership was the least problematic, as most models performed well. These findings highlighted the need for further refinement in AI models, particularly in domains requiring deeper contextual understanding and clinical reasoning (Ainingkun et al., 2025). Also, AI models can supplement healthcare education but should not replace expert review in high-stakes decisions.

The degree of trust in accuracy and capability will be crucial in shaping the scope and speed of AI LLM adoption in future certification, education, and decision-making processes (Sallam et al., 2024c; Waldock et al., 2024).

5.2 Study Strengths, Limitations, and Recommendations for Future Studies

This study offers several strengths, including a comprehensive comparison of seven widely available large language models—ChatGPT4.5, Gemini 2.5 Pro, Copilot Think Deeper, Claude 3.7 Sonnet, Perplexity, DeepSeek R1, and Manus—in the context of healthcare quality education. The inclusion of validated questions across four core domains and the use of standardized inputs contributed to the reliability and reproducibility of the evaluation. The findings highlight the consistent accuracy of the evaluated models, with all achieving or exceeding the internal benchmark score of 70 percent. The observed performance consistency across chatbots suggests that LLMs may be viable supplementary tools for supporting foundational knowledge in healthcare quality management.

Using accuracy rates as objective performance metrics provided quantifiable insights into model outputs, while including diverse AI platforms strengthened the generalizability of the findings (Sarker, 2022). Moreover, the structured question input and domain-level breakdown enabled a focused assessment of chatbot performance across specific content areas, identifying potential variation in domain-specific accuracy.

However, the study has certain limitations. The number of multiple-choice questions was limited to 20, which may constrain the scope of assessment and overlook broader functional capabilities of the models. The evaluation focused solely on accuracy and did not examine the quality or completeness of chatbot explanations. Additionally, while statistical testing revealed no significant differences among the models, the analysis did not explore model performance variability across individual domains using inferential methods.

As Perkins and Pregowska (2024) noted, overreliance on AI-generated outputs without human oversight may also impact critical thinking and introduce bias embedded in model training data. Furthermore, the findings reflect a specific snapshot in time and may not account for ongoing model updates or changes in API behavior.

Future research should explore the integration of LLMs into real-world healthcare education and assess their effectiveness in supporting professionals in more complex, context-dependent decision-making scenarios (Blacker et al., 2024). Upcoming studies may also benefit from including a larger and more diverse set of questions, comparing chatbot responses with human experts, and evaluating response quality, reasoning depth, and user-friendliness (Strachan et al., 2024). Investigating domain-specific performance using inferential statistics and the models' role in blended learning environments may provide additional insights. Future studies could also incorporate higher-order reasoning assessments to evaluate how well LLMs address conceptual and context-rich healthcare challenges (Mondorf & Plank, 2024).

6. Conclusion

This study evaluated the accuracy of seven large language models, including ChatGPT4.5, Gemini, Copilot, Claude, Perplexity, DeepSeek, and Manus in answering multiple-choice questions related to healthcare quality management. All models met or exceeded the predefined accuracy threshold of 70 percent, with performance ranging from 70 to 80 percent. ChatGPT4.5, Gemini, and Claude achieved the highest accuracy rates at 80 percent, followed by Perplexity, Manus at 75 percent, and Copilot and DeepSeek at 70 percent. The findings indicated comparable levels of accuracy across the models, with no significant differences observed in response distribution. This study highlighted the need for ongoing refinement of AI models to enhance reliability and contextual accuracy in healthcare quality education. Future research should examine the integration of LLMs into professional training and evaluate their effectiveness in real-world decision-support environments.

Funding

This research received no external funding.

Authors Contribution

Conceptualization: M. Sallam

Methodology: M. Sallam

Validation: Johan Snygg and A. Hamdan

Resources, data curation, and visualization: M. Sallam, D. Allam, R. Kassem, and M. Damani

Review and editing: M. Sallam, Johan Snygg, A. Hamdan, D. Allam, R. Kassem, and M. Damani

Supervision: M. Sallam

Reading and approving the final manuscript: All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors thank Mediclinic Middle East healthcare quality management experts who validated the twenty questions used in this study. Their expertise and contributions were instrumental in ensuring the objectivity and relevance of the evaluation.

Abbreviations

AI: Artificial Intelligence

NLP: Natural Language Processing

LLMs: Large Language Models

ChatGPT: A Chatbot Based on Generative Pre-Trained Transformer Large Language Model

JCIA: Joint Commission International Accreditation

MCQs: Multiple-Choice Questions

SPSS: Statistical Package for the Social Sciences

χ^2: Chi-Square test

References

A-Abbasi, H., Al-Qudheeby, M., Kheyami, Z. A., Khalil, R., Khamees, N., Hijjawi, O., Sallam, M., & Barakat, M., (2024). Cross-Linguistic Evaluation of Generative AI Models for Diabetes and Endocrine Queries. *Jordan Medical Journal*, *58*(4), 311-326. https://doi.org/10.35516/jmj.v58i4.3369

Ahmad, M., (2022). Multiple Choice Questions (MCQs) for CPHQ. In *QHorizon CPHQ Preparatory Material* (pp. 0-27). Qmentum Consulting. https://www.researchgate.net/publication/371990757_Multiple_Choice_Questions_MCQs_for_CPHQ

Ainingkun, X., Jingxue, T., Dehua, H., & Haixia, L., (2025). Comparative Study on Response Efficacy of Generative Artificial Intelligence Large Language Model for Elderly Diabetes Mellitus. *Journal of Innovations in Medical Research*, *4*(2), 66-76. https://doi.org/10.63593/JIMR.2788-7022.2025.04.008

Ali, K., & Zahra, D., (2024). Ten tips for effective use and quality assurance of multiple-choice questions in knowledge-based assessments. *Eur J Dent Educ*, *28*(2), 655-662. https://doi.org/10.1111/eje.12992

Alraimi, A. A., & Al-Nashmi, M. M., (2024). The interactive effect of applying the management-centered standards of Joint Commission International (JCI) and practicing administrative control in improving the quality of health services: a study on three Yemeni hospitals seeking accreditation. *Journal of Hospital Management and Health Policy*, *8*. https://doi.org/10.21037/jhmhp-24-47

Benítez, T. M., Xu, Y., Boudreau, J. D., Kow, A. W. C., Bello, F., Van Phuoc, L., Wang, X., Sun, X., Leung, G. K., Lan, Y., Wang, Y., Cheng, D., Tham, Y. C., Wong, T. Y., & Chung, K. C., (2024). Harnessing the potential of large language models in medical education: promise and pitfalls. *J Am Med Inform Assoc*, *31*(3), 776-783. https://doi.org/10.1093/jamia/ocad252

Blacker, S. N., Chen, F., Winecoff, D., Antonio, B. L., Arora, H., Hierlmeier, B. J., Kacmar, R. M., Passannante, A. N., Plunkett, A. R., Zvara, D., Cobb, B., Doyal, A., Rosenkrans, D., Brown, K. B. J., Gonzalez, M. A., Hood, C., Pham, T. T., Lele, A. V., Hall, L., ... Isaak, R. S., (2024). An Exploratory Analysis of ChatGPT Compared to Human Performance With the Anesthesiology Oral Board Examination: Initial Insights and Implications. *Anesthesia & Analgesia*. https://doi.org/10.1213/ane.0000000000006875

Brandrud, A. S., Nyen, B., Hjortdahl, P., Sandvik, L., Helljesen Haldorsen, G. S., Bergli, M., Nelson, E. C., & Bretthauer, M., (2017). Domains associated with successful quality improvement in healthcare — a nationwide case study. *BMC Health Services Research*, *17*(1), 648. https://doi.org/10.1186/s12913-017-2454-2

Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C., (2022). Two Decades of Artificial Intelligence in Education Contributors, Collaborations, Research Topics, Challenges, and Future Directions. *Educational Technology & Society*, *25*(1), 28-47. https://www.jstor.org/stable/48647028

Connor, L., Dean, J., McNett, M., Tydings, D. M., Shrout, A., Gorsuch, P. F., Hole, A., Moore, L., Brown, R., Melnyk, B. M., & Gallagher-Ford, L., (2023). Evidence-based practice improves patient outcomes and healthcare system return on investment: Findings from a scoping review. *Worldviews on Evidence-Based Nursing*, *20*(1), 6-15. https://doi.org/10.1111/wvn.12621

Gottlieb, M., Bailitz, J., Fix, M., Shappell, E., & Wagner, M. J., (2023). Educator's blueprint: A how-to guide for developing high-quality multiple-choice questions. *AEM Education and Training*, *7*(1), e10836. https://doi.org/10.1002/aet2.10836

Hristidis, V., Ruggiano, N., Brown, E. L., Ganta, S. R. R., & Stewart, S., (2023). ChatGPT vs Google for Queries Related to Dementia and Other Cognitive Decline: Comparison of Results. *J Med Internet Res*, *25*, e48966. https://doi.org/10.2196/48966

Imanipour, M., Ebadi, A., Monadi Ziarat, H., & Mohammadi, M. M., (2022). The effect of competency-based education on clinical performance of health care providers: A systematic review and meta-analysis. *Int J Nurs Pract*, *28*(1), e13003. https://doi.org/10.1111/ijn.13003

Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., Chang, S., Berkowitz, S., Finn, A., Jahangir, E., Scoville, E., Reese, T., Friedman, D., Bastarache, J., van der Heijden, Y., Wright, J., Carter, N., Alexander, M., Choe, J., ... Wheless, L., (2023). Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Res Sq*. https://doi.org/10.21203/rs.3.rs-2566942/v1

Kazana, I., & Dolansky, M., (2021). Quality improvement: Online resources to support nursing education and practice. *Nursing Forum*, *56*(2), 341-349. https://doi.org/10.1111/nuf.12533

Khlaif, Z. N., Mousa, A., Hattab, M. K., Itmazi, J., Hassan, A. A., Sanmugam, M., & Ayyoub, A., (2023). The Potential and Concerns of Using AI in Scientific Research: ChatGPT Performance Evaluation. *JMIR Med Educ*, *9*, e47049. https://doi.org/10.2196/47049

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V., (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Li, X., (2023). Research on methods and applications of question answering system in the context of ChatGPT. *Proceedings of the 5th International Conference on Computing and Data Science*.

Liu, M., Okuhara, T., Chang, X., Shirabe, R., Nishiie, Y., Okada, H., & Kiuchi, T., (2024b). Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. *J Med Internet Res*, *26*, e60807. https://doi.org/10.2196/60807

Liu, M., Okuhara, T., Dai, Z., Huang, W., Okada, H., Furukawa, E., & Kiuchi, T., (2024a). Performance of Advanced Large Language Models (GPT-4o, GPT-4, Gemini 1.5 Pro, Claude 3 Opus) on Japanese Medical Licensing Examination: A Comparative Study. https://doi.org/10.1101/2024.07.09.24310129

Mondorf, P., & Plank, B., (2024). Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models — A Survey. ArXiv, abs/2404.01869. https://doi.org/10.48550/arXiv.2404.01869

Myers, J. S., Kin, J. M., Billi, J. E., Burke, K. G., & Harrison, R. V., (2022). Development and validation of an A3 problem-solving assessment tool and self-instructional package for teachers of quality improvement in healthcare. *BMJ Qual Saf*, *31*(4), 287-296. https://doi.org/10.1136/bmjqs-2020-012105

Newton, P., & Xiromeriti, M., (2024). ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. *Assessment & Evaluation in Higher Education*, *49*(6), 781-798. https://doi.org/10.1080/02602938.2023.2299059

Ostapuk, N., & Audiffren, J., (2024). Tasks for LLMs and Their Evaluation. In A. Kucharavy, O. Plancherel, V. Mulder, A. Mermoud, & V. Lenders (Eds.), *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation* (pp. 65-72). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-54827-7_6

Perkins, M., & Pregowska, A., (2024). The role of artificial intelligence in higher medical education and the ethical challenges of its implementation. *AIH*, *2*(1). https://doi.org/10.36922/aih.3276

Salam, A., Yousuf, R., & Bakar, S. M., (2020). Multiple Choice Questions in Medical Education: How to Construct High Quality Questions. *International Journal of Human and Health Sciences (IJHHS)*, *4*, 79. https://doi.org/10.31344/ijhhs.v4i2.180

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

Sallam, M., (2024). Hospital Pharmacy Operations Management: Synergizing Lean Efficiency and Six Sigma Precision for Optimal Service Quality — An Action Research From United Arab Emirates. (Publication Number 31485273) Doctoral Dissertation, International American University. ProQuest Dissertations & Theses Global, United States. https://www.proquest.com/docview/3142695951?utm_medium=email&utm_source=transaction

Sallam, M., Al-Mahzoum, K., Almutawaa, R. A., Alhashash, J. A., Dashti, R. A., AlSafy, D. R., Almutairi, R. A., & Barakat, M., (2024a). The performance of OpenAI ChatGPT-4 and Google Gemini in virology multiple-choice questions: a comparative analysis of English and Arabic responses. *BMC Research Notes*, *17*(1), 247. https://doi.org/10.1186/s13104-024-06920-7

Sallam, M., Al-Mahzoum, K., Alshuaib, O., Alhajri, H., Alotaibi, F., Alkhurainej, D., Al-Balwah, M., Barakat, M., & Egger, J., (2024b). Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic. *BMC Infectious Diseases*, *24*, 799. https://doi.org/10.1186/s12879-024-09725-y

Sallam, M., Al-Mahzoum, K., Sallam, M., & Mijwil, M. M., (2025). DeepSeek: Is it the End of Generative AI Monopoly or the Mark of the Impending Doomsday? *Mesopotamian Journal of Big Data*, *2025*, 26-34. https://doi.org/10.58496/MJBD/2025/002

Sallam, M., & Hamdan, A., (2023). Examining the Influence of Joint Commission International (JCI) Accreditation Surveys on Medication Safety Practices: A Cross-Sectional Study from Mediclinic Welcare Hospital in Dubai, UAE. *Journal of Integrated Health*, *2*(4), 68-79. https://doi.org/10.51219/JIH/Mohammed-Sallam/13

Sallam, M., Snygg, J., & Sallam, M., (2024c). Assessment of Artificial Intelligence Credibility in Evidence-Based Healthcare Management with "AERUS" Innovative Tool. *Journal of Artificial Intelligence, Machine Learning and Data Science*, *1*(4), 9-18. https://doi.org/10.51219/JAIMLD/mohammed-sallam/20

Sarker, I. H., (2022). AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems. *SN Computer Science*, *3*(2), 158. https://doi.org/10.1007/s42979-022-01043-x

Shen, M., & Yang, Q., (2025). From Mind to Machine: The Rise of Manus AI as a Fully Autonomous Digital Agent. Preprint. https://doi.org/10.48550/arXiv.2505.02024

Spath, P., (2013). *Introduction to Healthcare Quality Management*. Health Administration Press. https://books.google.ae/books?id=yQgXmAEACAAJ

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C., (2024). Testing theory of mind in large language models and humans. *Nat Hum Behav*, *8*(7), 1285-1295. https://doi.org/10.1038/s41562-024-01882-z

Waldock, W. J., Zhang, J., Guni, A., Nabeel, A., Darzi, A., & Ashrafian, H., (2024). The Accuracy and Capability of Artificial Intelligence Solutions in Health Care Examinations and Certificates: Systematic Review and Meta-Analysis. *J Med Internet Res*, *26*, e56532. https://doi.org/10.2196/56532

Weheba, G., Cure, L., & Toy, S., (2020). Perceived dimensions of healthcare quality in published research. *International Journal of Healthcare Management*, *13*(sup1), 357-364. https://doi.org/10.1080/20479700.2018.1548156

Zabin, L. M., Shayeb, B. F., Kishek, A. A. A., & Hayek, M., (2024). Nursing perception towards the impact of JCI accreditation on the quality of care in a university hospital in Palestine: a cross-sectional study. *BMC Nursing*, *23*(1), 695. https://doi.org/10.1186/s12912-024-02353-6