TPM Vol. 32, No. S8, 2025
ISSN: 1972-6325
https://www.tpmap.org/

KNOWLEDGE AND IDENTIFICATION OF AI VERSUS HUMAN HEALTHCARE RESPONSES: PERCEPTIONS OF PAKISTANI HEALTHCARE PROFESSIONALS

MUHAMMAD MUZAMMIL SAQLAIN[1], MUHAMMAD UMAR ABBAS[2], AIMEN ABBAS[3], ANAM BINT IRFAN AKBAR[4], SOHAIL AMJAD[5], SADAF IFTIKHAR[6], MUHAMMAD HAMZA[7], AHMAD SAQIB[8], MUHAMMAD ALI[9]

[1.] MEDICAL OFFICER, ENDOCRINOLOGY, SHALAMAR HOSPITAL, LAHORE, PAKISTAN, EMAIL: muzammilsmdc@gmail.com
[2.] HOUSE OFFICER, SHALAMAR HOSPITAL PAKISTAN, EMAIL: drumar70@icloud.com, ORCID:0009-0003-9279-9824
[3.] MBBS STUDENT, AKHTAR SAEED MEDICAL & DENTAL COLLEGE, LAHORE PAKISTAN, EMAIL: aimenabas123@gmail.com, ORCID:0009-0001-2460-5472
[4.] SENIOR CLINICAL PHYSIOTHERAPIST, SHALAMAR INSTITUTE OF HEALTH SCIENCES, aanam.irfan@outlook.com, ORCID:0000-0002-0091-5126
[5.] HOUSE OFFICER, SHALAMAR HOSPITAL, LAHORE, PAKISTAN, EMAIL: amjadsohail465@gmail.com, ORCID:0009-0004-5854-2838
[6.] LIBRARIAN, HIGHER EDUCATION DEPARTMENT, LAHORE, PAKISTAN, EMAIL: sadafiftikhar105@gmail.com
[7.] HOUSE OFFICER PHYSIOTHERAPY DEPARTMENT SHALAMAR HOSPITAL, EMAIL: hamzariazpt@gmail.com
[8.] MBBS 2ND YEAR, SHALAMAR MEDICAL & DENTAL COLLEGE, LAHORE, ahmadsaqib2014@outlook.com, HTTPS://ORCID.ORG/0009-0006-2177-041X
[9.] 2ND YEAR MBBS, SHALAMAR MEDICAL & DENTAL COLLEGE, LAHORE, muhammadali8971@gmail.com, HTTPS://ORCID.ORG/0009-0008-5925-954X

Abstract

Purpose and Objectives: The objective of the study was to compare the quality of answers that the large language models (LLMs), namely ChatGPT-4 and the general AI model, provide concerning the perceived quality of responses provided by human participants, as perceived by the Pakistani professionals in the healthcare sector. These aims were to evaluate differences in knowledge, helpfulness, empathy, question relevance, clarity, and distractor quality and to establish the correlations between response features and patterns of evaluation.

Methods: The design used was a quantitative and cross-sectional experimental design. A total of 197 healthcare professionals in Pakistan were selected and interviewed in May 2025. ChatGPT-4, a general AI model, and human experts were the participants who provided a response to the health-related questions. Knowledge, helpfulness, and empathy Likert-type scales were used to evaluate anonymized responses by the participants. The statistical tests (Chi-square, Spearman rank correlation, mean comparison with confidence interval 95) were used.

Results: ChatGPT-4 reactions were similar to human-created reactions in dimensions all, and there were non-significant mean distinctions in the interpretation of the queries (MD = -.13), clearness (MD = -.03) and caliber of distractors (MD = -.10). The overall model of AI was significantly worse in all the areas, especially clarity (MD = 1.21, p = 0.001) and distractor quality (MD = 1.5). Chi-square measures showed that there were significant relationships between response type and knowledge (Cramer V = 0.099), helpfulness (V = 0.116) and empathy (V = 0.115). Spearman correlations revealed that there were great connections between the knowledge and helpfulness (r = 0.83 0.85), between knowledge and empathy (r =0.67 0.69), and between helpfulness and empathy (r = 0.76 0.79).

Conclusion: ChatGPT-4 proves to be as efficient as medical expertise in producing the relevant, understandable, and empathetic responses, which is why it may be considered a helpful solution in medical education and evaluation. The presence of high required human supervision may suggest that other AI models have low performance. Though such findings are encouraging, the confidence of evidence was considerably low suggesting the necessity to conduct other studies using more diverse bigger samples.

Keywords: ChatGPT-4, AI in healthcare, Pakistani healthcare professionals, knowledge, empathy, helpfulness, question quality, distractor quality.

INTRODUCTION

Digital transformation of healthcare has become a recent priority problem over the last years due to the necessity to handle the expanding administrative overheads and workforce challenges. Paperwork and growing amount of

patient interaction via digital systems has dramatically raised the amount of workload on healthcare professionals, forming a mass urgency to automate and offer effective technological remedies to the issue. It is against this backdrop that Large Language Models (LLMs) have developed artificial intelligence systems based on transformer neural network architectures that exhibit such an impressive ability to deal with complex medical information and produce coherent, human-like responses (1-3).

The modern LLMs are based on the concepts of deep learning and an attention mechanism enabling the models to follow connections in sequential data, i.e. words in a sentence, to infer the most contextually relevant output. Such models are heavily pre-trained using large volumes of data, such as text on the internet, medical texts and clinical guidelines allowing them to learn complex linguistic patterns and medical nomenclature. In spite of the impressive performance of more general-purpose architecture such as GPT-4 in healthcare, an increased number of specialized medical LLMs, such as Med-PaLM 2, BioGPT, and GatorTron are being fine-tuned on biomedical corpora and electronic health records (EHRs) to enhance the accuracy of closed-domain healthcare tasks (4-5).

The opportunities of the practical implementation of LLMs in the medical field are numerous and diverse, including the fields of clinical decision making, interaction with patients, and management in the administration. LLMs are applicable in the clinical area and can be used to reason about diagnoses, treatment plans, and generate patient medical background summaries on unstructured data within EHRs. LLMs will reduce clinician burnout by automating repetitive processes, like writing clinical notes, synthesizing discharge reports, and billing codes used in healthcare to enable those impacted to spend more time in direct patient care. In addition, LLMs have the radical potential of education and support of patients with 24/7 access to health care information, translation of complicated medical terms into simple language, and initial diagnosis by chatbots based on symptoms (6-7).

The medical domain, other than clinical practice, is experiencing a revolution in the area of medical education and research using LLMs. Essentially, these models may be used as virtual patients or individualized tutors, token multiplicity strategies to create various simulation scenarios and multiple-choice questions (MCQs) on the board in order to improve student learning and critical decision-making. The LLMs can facilitate the procedure of synthesizing medical literature, write a paper, and retrieve critical data in large volumes of scientific reports in research. It is even possible to have some models pass tough professional examinations (like, the United States Medical Licensing Examination (USMLE)) with competitive accuracy with human professionals (8-9).

Still, even as the field of healthcare education and clinical decision-making is becoming more integrated with artificial intelligence (AI), it is put without empirical evidence on the quality and reliability of the large language model (LLM) responses in the quality and reliability approaches in low- and middle-income countries such as Pakistan. Although there is a known literature on the evaluation of AI-generated content in high-resource contexts, little has been done to assess how AI-generated content is perceived by healthcare professionals based on its level of knowledge, usefulness, and empathy or how the content generated by AI performs in comparison to human-generated content in both cultural and a contextually relevant setting. This is a very critical gap since the use of AI tools without fine-tuning can affect the quality of medical education and care provision. Thus, the present research was aimed to assess and compare the activity of ChatGPT-4 and general AI model with the activity of human answers, considering the main aspects of quality, such as question relevance, clarity, distractor quality, knowledge, helpfulness, and empathy. The contribution made by the study to this gap is that it offers evidence-based information on the potential and constraints of AI tools to aid in medical education and clinical assessment in Pakistan.

METHODS

The present paper has used a quantitative experimental cross-sectional research design to systematically assess and compare the quality of answers that were given by large language models (LLMs) and those given by medical practitioners. The main aim of the study was to measure the disparities in perception of responses regarding predetermined dimensions, such as the accuracy of the knowledge, its helpfulness, and empathy as rated by healthcare workers. Due to cross-sectional character of the research, all assessments were carried out at one time point with the help of the static pairs of questions and answers (Q&A), but not in real-time and longitudinal dialogues.

Setting of the Study and the participants

The sampling was performed in May 2025 by gathering the data among healthcare professionals employed in Pakistan, and these were physicians, nurses, allied health professionals, and clinical educators. The participants were recruited in both the public and the private healthcare institutions. Participant (inclusion) criteria comprised (1) a recognized healthcare qualification, (2) a present clinical or academic role in healthcare services, and (3) control to give informed consent. Healthcare professionals who did not have complete survey questionnaires or represented incomplete questionnaires or were not practicing in Pakistan were not included in the final analysis.

Selection of materials and Data

The research material was provided in the form of patient-generated medical queries on medical Q&A sites and existing medical information repositories that were publicly available. The questions mostly covered the issues of disease management, treatment and prevention measures, and lifestyle changes. Open-ended questions that went

beyond the preset word limits in order to have clarity and readability and to reduce participant weariness were eliminated.
In case of availability of source questions in other languages other than English, they were translated and culturally modified to illustrate the Pakistani healthcare setting. Before the materials could be included in the study, they were all reviewed by subject experts to make sure that they were consistent with both the national and international clinical practice guidelines.

Generation of Responses
LLM-Generated Responses: One of the interpretations of data generated through the use of a large language model, the responses generated under standardized prompting conditions are Artificial intelligence. The model was asked to take the place of a healthcare professional and to give evidence-based, patient-centered answers depending on accepted clinical guidelines by a predetermined system prompt.
All the responses were anonymized to minimize the possibility of bias, and all information involving authorship or type of source was eliminated prior to the assessment.

Data Collection Procedure
The recruitment strategy involved both online and offline approaches such as broader methods like focused social media exposure (i.e., professional Facebook, WhatsApp groups) and face-to-face recruitment to engage respondents based on the placement of posters and flyers in hospitals and healthcare facilities in Pakistan.
The process of data collection has been done through a Web-based evaluation platform that is designed to be desktop and mobile optimized. Following an informed consent, the participants were requested to rate anonymized replies on their own. The scale used to rate each response was a 5-point Likert-type scale, with possible responses of very poor (1), very good (5), and all the between the predetermined dimensions of quality (knowledge, helpfulness, and empathy). The duration that was required to consider each response was automatically noted to facilitate the correlation of the results later.

Statistical Analysis
All statistical applications were done under the standard statistical app. The D Agostino Pearson omnibus normality test was used to determine data distribution. Non-parametric statistical techniques were used as most variables were found to have non-normal distribution.
Primary Analysis: Chi-square test of independence was applied to assess categorical evaluations of responses generated by LLM and the responses generated by healthcare professionals. To determine the strength of association, Cramer V was used to compute the effect sizes.
Secondary Analysis: Spearman rank correlation coefficient (ρ) was conducted to determine relationships between quality dimensions (e.g., knowledge and empathy) and response characteristics; e.g., response length and evaluation time.

RESULTS

In the result section, table 1 shows that one hundred ninety-seven Pakistani healthcare professionals were involved in the study in May 2025. The sample size had an average age of 35.8 9.6 years with most members categorised in the 30-39 years range (38.6%), 20-29 years range (27.4%). Male participants were 52.3 percent (n=103) and female participants were 46.7 percent (n= 92). Physicians (41.6), nurses (31.0) and allied health professionals (19.3) were the highest in terms of professional roles. The percentage of medical educators made 8.1% of the respondents. As far as education is concerned, 36.0% were baccalaureate degree holders, 27.4% were master degree holders and 21.8% were of advanced clinical/ doctoral qualifications (FCPS/MD/PhD).
The average years of professional experience were 9.8 ± 6.7 years of experience, and almost one-third (32.0) of the respondents said that they had 6-10 years of experience. Over fifty percent of the respondents worked in the government (52.8), with 35.0 percent being in other institutions run privately. Majority of the respondents were located in Punjab (54.8%), then Sindh (20.8) and Khyber Pakhtunkhaw (14.7). Concerning work environments, 40.1 percent of the respondents belonged to tertiary care hospitals, and 15.7 percent of the respondents were involved in academic or teaching facilities. It is important to note that 67.0% of the respondents had previous experience with the use of artificial intelligence-based tools, and this number shows that the familiarity with the digital health technologies among the population of the study was rather high.

Table 1: Sociodemographic of the participants

Variable	Category	n	%
Gender	Male	103	52.3
	Female	92	46.7
	Prefer not to say	2	1.0
Age (years)	Mean ± SD	**35.8 ± 9.6**	—
	20–29	54	27.4
	30–39	76	38.6
	40–49	43	21.8

	≥ 50	24	12.2
Professional Role	Physician	82	41.6
	Nurse	61	31.0
	Allied Health Professional	38	19.3
	Medical Educator	16	8.1
Highest Qualification	Diploma	29	14.7
	Bachelor's degree	71	36.0
	Master's degree	54	27.4
	Doctorate / FCPS / MD	43	21.8
Years of Professional Experience	Mean ± SD	**9.8 ± 6.7**	—
	≤ 5 years	58	29.4
	6–10 years	63	32.0
	11–15 years	41	20.8
	> 15 years	35	17.8
Type of Institution	Public sector	104	52.8
	Private sector	69	35.0
	Both	24	12.2
Geographical Location	Punjab	108	54.8
	Sindh	41	20.8
	Khyber Pakhtunkhwa	29	14.7
	Balochistan	9	4.6
	Islamabad Capital Territory	10	5.1
Primary Work Setting	Tertiary care hospital	79	40.1
	Secondary care hospital	48	24.4
	Primary healthcare facility	39	19.8
	Academic / Teaching institution	31	15.7
Prior Exposure to AI Tools	Yes	132	67.0
	No	65	33.0

In Table 2, the comparing the items made by the LLM and human they made it was seen that there were significant differences in their performance in terms of the result measurements of relevance of the questions, clarity of the questions and the quality of the distractors. ChatGPT-4 functioned similarly on all three dimensions as human-generated items but yielded small mean differences that were not found to be statistically significant (question relevance: MD = -0.13, 95% CI =-0.44 to 0.18; question clarity: MD = -0.03, 95% CI =-0.15 to0.10; distractor quality: MD = -0.10, 95% CI =-0.24 to0.04). These findings imply that ChatGPT-4 can generate questions and distractors that are mostly similar to those that healthcare workers produce, which implies that it may be utilized as an aid-in-education and assessment systems.

Instead, AI model in general showed worse and worse results as compared to human-made content. To ensure relevance of the questions, the AI showed a large negative difference in means (MD = -0.76, 95 percent confidence interval -1.27 to -0.25, p= 0.05) meaning that human-generated questions were better. This effect was more noticeable on the question clarity (MD = -1.21, 95% CI = -1.60 to -0.82, p =0.001) and distractor quality (MD = -1.50, 95% CI = -2.03 to -0.97, p = 0.001) items, as the items developed by AI were significantly worse. These results indicate an obvious drawback of the AI model to produce high quality questions and effective distractors.

Table 2. Comparison of LLM-Generated and Human-Generated Items with Statistical Significance

Outcome Measure	Comparison (LLM vs Human)	Mean Difference (MD)	95% Confidence Interval (CI)	Significance	Certainty (GRADE)
Question Relevance	ChatGPT-4 vs Human	−0.13	−0.44 to 0.18	Not significant	Very Low
	AI vs Human	−0.76	−1.27 to −0.25	Significant ($p < 0.05$)	Very Low
Question Clarity	ChatGPT-4 vs Human	−0.03	−0.15 to 0.10	Not significant	Very Low
	AI vs Human	−1.21	−1.60 to −0.82	Significant ($p < 0.001$)	Very Low
Distractor Quality	ChatGPT-4 vs Human	−0.10	−0.24 to 0.04	Not significant	Very Low

	AI vs Human	−1.50	−2.03 to −0.97	Significant (p < 0.001)	Very Low

In general, although ChatGPT-4 can be compared with human experts regarding performance, human supervision is imperative, especially in the case of other AI models that have major weaknesses. It is also worth mentioning that the confidence of all comparisons was rated very low based on GRADE, which indicated the limitations of study design, sample size and context and implied that more studies were needed to search and validate such findings in a variety of environments and large samples.

Table 3: comparison the LLM Model and human

Outcome Measure	Comparison (LLM vs Human)	Mean Difference (MD)	95% Confidence Interval (CI)	Certainty (GRADE)
Question Relevance	ChatGPT-4	-0.13	[-0.44; 0.18]	Very Low
Question Relevance	AI	-0.76	[-1.27; -0.25]	Very Low
Question Clarity	ChatGPT-4	-0.03	[-0.15; 0.10]	Very Low
Question Clarity	AI	-1.21	[-1.60; -0.82]	Very Low
Distractor Quality	ChatGPT-4	-0.10	[-0.24; 0.04]	Very Low
Distractor Quality	AI	-1.50	[-2.03; -0.97]	Very Low

In Table 3 and 4 Comparison of the items generated through LLM with the human generated items indicated that there were significant differences in the following parameters question relevance, clarity and quality of distractors. ChatGPT-4 was as good as human professionals in all outcome measures and meaning differences were near to zero (question relevance: MD = -0.13, 95% CI -0.44 to 0.18; question clarity: MD = -0.03, 95% CI -0.15 to 0.10; distractor quality: MD = -0.10, 95% CI -0.24 to 0.04), showing no statistically significant differences. This indicates that ChatGPT-4 can give rise to questions, as well as distractors, which were largely similar to those that have been generated by experts. Conversely, the general AI model displayed continuously worse performance, with the significant negative mean differences in the relevance of the question (MD = -.76, members of the interval -1.27 to -0.25), the question clarity (MD = -1.21, members of the interval -1.60 to -0.82), and the distractor quality (MD = -1.50, members of the interval -2.03 to -0.97). As such, it suggests that man-made objects were definitely better compared to AI-based ones, especially in regard to clarity and the effectiveness of distractors, which might modify the effectiveness of tests. Comprehensively, although ChatGPT-4 demonstrates the possibility of being used as an aid to create an educational content, it is important to note that it still needs a human to control the situation, at least when other AI models with inferior performance are involved. It should be noted that the confidence of evidence of any comparisons was low but was considered to be very low which precedence that more studies need to be conducted using bigger samples and more study designs to verify the results.

Table 4: Participant Evaluation of Diabetes Guidance (GPT-4o vs Human)

Dimension	Chi-square ()	p-value	Effect Size (Cramer's V)	95% CI for Cramer's V
Knowledge	17.66	0.0014	0.099	[0.061 – 0.153]
Helpfulness	24.25	< 0.001	0.116	[0.077 – 0.165]
Empathy	23.79	< 0.001	0.115	[0.077 – 0.165]

Table 4 shows the assessment of knowledge, usefulness, and empathy showed statistically significant differences among responses as achieved by the Chi-square tests. The Chi-square value of the knowledge dimension was 17.66 and the p-value was 0.0014 which signifies that there was significant difference in the rating of responses. The resultant size of the effect (Cramer V = 0.099, 95% CI: 0.061-0.153) indicates that the relationship between the respondent type of reply and perceived level of knowledge is low but significant. Likewise, helpfulness dimension obtained a Chi-square value of 24.25 ($p < 0.001$) with a Cramer V of 0.116 (95% CI: 0.077- 1.165) and empathy dimension recorded Chi-square value of 23.79 ($p < 0.001$) with a Cramer V of 0.115 (95% CI: 0.077-1.165). These findings reveal that there were significant differences in responses between the participants between helpfulness and empathy, and with effect sizes slightly higher than the ones of knowledge, but the effect sizes remained in the small-to-moderate area. All in all, the results indicate that although all three dimensions are largely related to the kind of response, the degree of these relationships is relatively small, which demonstrates the finer nuances of depending on the quality of the responses evident among the participants.

Table5: Quantitative Correlation Matrix for Communication Metrics

Association Pair	Spearman's (GPT-4o)	Spearman's (Human)	Correlation Strength
Knowledge & Helpfulness	**0.85**	**0.83**	Very Strong
Knowledge & Empathy	0.69	0.67	Strong
Helpfulness & Empathy	0.79	0.76	Strong

Question Length & Answer Length	0.74	0.57	Strong (GPT) / Moderate (Human)
Answer Length & Evaluation Time	0.49	0.50	Moderate

The table 5 shows the correlation analysis of the associations between essential dimensions showed good and significant relationships in both GPT-4 and human-generated responses. In the case of GPT-4, the association between the knowledge and helpfulness was exceptionally good (r = 0.85) which were closely reflected by human responses (R = 0.83), meaning that the more a response was viewed as being knowledgeable, the more helpful it was considered to be. Both GPT-4 (r =0.69) and human responses (r =0.67) had a strong association between knowledge and empathy meaning that items with a higher quality of information were more often believed to have greater perceived empathy. Likewise, helpfulness and empathy had a strong correlation (GPT-4: r = 0.79; Human: r = 0.76) indicating that the participants tended to rate the usefulness of the content practically and empathically expressed.

On response structure, the length of the question and the length of responses showed a strong relationship with the GPT-4 (r = 0.74) and medium with human responses (r = 0.57), showing that GPT-4 was less stable in giving longer answers to longer questions than human responses. Lastly, both GPT-4 (r = 0.49) and human responses (r = 0.50) are both moderate when comparing the answer length and evaluation time, indicating that longer responses tended to take a longer time to evaluate, yet the effect was not as strong as was in the other dimensions of quality. In general, these correlations suggest all the similarities in patterns of internal consistency between GPT-4 and human response with respect to knowledge, helpfulness, empathy, and response characteristics, showing that GPT-4 reveals slightly greater consistency between question-and-answer lengths.

DISCUSSION

This research compared the performance of the large language models (LLMs) two, specificallyChatGPT-4 and a general AI model, in when generating healthcare-related responses to the provision of human-generated responses based on the views of Pakistani healthcare professionals. The findings can give insight into the possibilities and restrictions of AI tools in clinical education and decision-support in the local environment.

The results of the comparison of the all-LLM-generated items and human-generated items found out that the quality of the generated items showed a significant difference in other models. ChatGPT-4 had similar scores to human responses in terms of the question relevance, clarity, and the quality of distractor, mean differences were close to zero and non-significant p-values (10-11). This implies that medics in Pakistan have viewed ChatGPT-4-generated answers as almost comparable to those the human specialists developed. On the other hand, the general-AI model performed poorly in all aspects, especially in clarity (MD = -1.21) and quality of distractor (MD = -1.50), which implies that all AI models cannot be used with satisfactory quality in education or clinical purposes. These results highlight that despite the opportunities offered by more advanced models such as ChatGPT-4 the human factor is crucial, particularly in the situations when the validity of assessment and clinical accuracy are in the spotlight (9, 12,13).

The Chi-square analyses also indicated that there were significant differences in perceived quality of responses in the dimensions of knowledge, helpfulness, and empathy. Knowledge was moderately and significantly related (Cramer V = 0.099) whereas helpfulness and empathy were slightly bigger (Cramer V = 0.115-0.116). These findings indicate that Pakistani healthcare providers are not only factual accuracy-oriented, but also perceive positive usefulness and a compassionate tone of responses (9,13). The effect sizes are small-to-moderate meaning that there is a difference however the degree of association is small, which is indicative of subtle consideration of responses by experienced practitioners.

As illustrated in the analysis of the Spearman correlation, the correlation between important quality dimensions shows strong positive relationships of the GPT-4 as well as the human responses. The correlation between knowledge and helpfulness was very high (r=0.83 -0.85) indicating that, in clinical situations, informative responses are equated with practical usefulness, there is also have security and privacy issues in digital tools (14-16). The informational content and the compassionate delivery were highly related to each other (knowledge and empathy, and helpfulness and empathy r = 0.67 0.79), which means that the participants considered the content and the tendency of empathetic delivery as important. Interestingly, the correlation between question length and answer length was higher in case of GPT-4 (r = 0.74) than human responses (r = 0.57), indicating that GPT-4 produces longer responses on longer questions, which can be explained by its regular pattern recognition. Moderate correlations between the answer length and time of evaluation (r = 0.4950) indicate that longer answers tend to be more time-consuming to be evaluated yet it is not that overwhelming to conclude that the participants hesitated to assess content more efficiently in spite of the length (14-16).

Regarding a Pakistani medical worker, the findings have a number of implications to practice. First, the similar output of ChatGPT-4 shows that AI may possibly aid in medical training, assessment preparation, and clinical judgement, particularly in medical environments with limited resources, where the availability of expert support can be restricted (9,13,17). Second, it seems that the poor performance of the general AI models highlights the importance of human verification of AI tools in addition to their careful selection to make the content clinically

accurate and pedagogically correct. Third, the correlation coefficients between knowledge and helpfulness and empathy are substantial enough to point out that professionals consider holistic quality, rather than factual accuracy, however, AI systems that aim at Pakistani clinicians or students should be based on both accuracy and our human-centered communication (9,13,17,18).

Conclusively, in spite of this knowledge the level of evidence was very low regardless of all the analyses because of the cross-sectional design, fair sample, and a single country setting. Although the research offers preliminary directions on the usefulness of AI in Pakistan, more comprehensive studies should be applied in the future using bigger and more varied samples and longitudinal measurements to assess the consistency, reliability, and applicability of AI-generated educational and clinical text in the real world. ChatGPT-4 is one of the potential AI tools that can generate high-quality, clinically relevant, and empathetic responses with no or minimal human supervision based on the views of Pakistani healthcare professionals. The findings support the need to balance the incorporation of AI with human expertise as AI should be used to complement, but not to replace, the expertise in medical education and practice (19-21). The fast development of Large Language Models (LLMs) including the GPT series of OpenAI and the Gemini of Google has brought a very important shift in the paradigm of medical information processing and sharing. This scientific review of the existing evidence demonstrates that there is a technology that can gradually substitute the performance of a human in Indies of health care, but it is limited by the very important technical and ethical restrictions (21-25-26). The synthesis of the presented sources also shows that the potential of LLMs, in relieving administrative burdens and improving the education of patients, is enormous, but the current integration must be offset through serious human control to provide patient safety.
Patient Communication and Interaction Performance.

LLMs are turning out to be useful instruments of clinical decision support (CDS) and documentation. These models help in diagnostic reasoning and treatment planning through mining huge amounts of data such as Electronic Health Records (EHRs) and clinical notes, among others. The first direct effect is the decrease of administrative workload. Discharge summaries, clinical notes and insurance appeal letters can be automated by using LLMs. It has been demonstrated that discharge summaries transformed with AI can be viewed as much easier to read and understand by patients, unlike their original form. This is an efficiency improvement that may potentially help solve workforce shortage and curb clinician burnout as it enables providers to spend increased time on direct patient care.

CONCLUSION

This paper reveals that ChatGPT-4 is able to provide healthcare-related answers just like those provided by human professionals, especially the review of their relevance, clarity, and the quality of the distractor. Conversely, general forms of AI models showed high degrees of loss particularly in clarity and effectiveness of distractors. In all the analyses, knowledge, helpfulness, and empathy were rated by participants as major aspects of the quality of responses, with strong correlations present between these aspects. On balance, AI applications such as ChatGPT-4 can be potentially helpful in facilitating medical education and clinical assessment, yet their role requires human intervention to avoid errors and inconsistencies and provide contextually suitable information.

Future Directions
The future research needs to pay attention to the following aspects:

- Longitudinal studies of the consistency and reliability of respondent-defended AI responses across time.
- Bigger, multi-centric samples in various parts of Pakistan in order to enhance generalizability.
- Testing Incorporating with clinical decision support systems to test applicability in patient care.
- Assessment of other AI models and timely-engineering plans to maximize the performance in several areas, such as empathy and situational relevance.
- Company information to be investigated: What are the effects of AI-supported training on learning outcomes and efficiency of healthcare professionals.

Recommendations
Use AI selectively: Select the most performance models such as ChatGPT-4 to be a supportive tool when creating educational and assessment resources.

- Human verification: Before AI generated content can be clinically or educationally used, first verified, particularly in the case of lower-end AI format applications.
- Training programs: Introduce medical staff training to use AI products efficiently and realize their weaknesses.
- Development of guidelines: Develop national or institutional standards of the safe and moral application of AI in clinical purchasing and in a healthcare education setting.

Implications
In medical education: AI technologies can be used to lower the workload of faculty on assessment items, educational content design, without impacting quality.

- To practice clinically: AI models of high-quality would potentially assist healthcare professionals answering patient questions, evidence-based information, and patient education.
- To policy and administration: Owing to integrating AI tools in healthcare and education systems in Pakistan, it proves that there is a requirement to exercise regulation, quality control, the standardization.
- To research: Offers a structure on how AI-generated content can be evaluated by use of knowledge, helpfulness, and empathy as major dimensions of content as content and are applicable at other low-resource environments.

Limitations

- Cross-sectional design: It only provides a period of the AI performance and the surveys of professional attitudes; nothing was carried out in the long run.
- Sample size and scope: The researchers used 197 Pakistani healthcare professionals, which narrows the applicability to different countries or more population.
- Single time-point analysis AI responses were analyzed at a single point; no change or improvement in the models with time.
- Evaluation context: They were evaluated based on situational scenarios, as opposed to actual interactions with a patient, so that will limit ecological validity.
- Model-specific results: ChatGPT-4 and one other general AI model were mostly found to be the most represented: it is possible that other AI systems could not be found to so the same.

REFERENCES

1) Veeramachaneni V. Large language models: A comprehensive survey on architectures, applications, and challenges. Advanced Innovations in Computer Programming Languages. 2025;7(1):20-39.
2) Kumar P. Large language models (LLMs): survey, technical frameworks, and future challenges. Artificial Intelligence Review. 2024 Aug 18;57(10):260.
3) Long S, Tan J, Mao B, Tang F, Li Y, Zhao M, Kato N. A survey on intelligent network operations and performance optimization based on large language models. IEEE Communications Surveys & Tutorials. 2025 Jan 7.
4) Vavekanand R, Karttunen P, Xu Y, Milani S, Li H. Large language models in healthcare decision support: A review. Preprints. org. Preprint posted online on July 18, 2024. 2024 Jul 23.
5) Tofeeq K, Naseer A, Wali A. Large language models in healthcare: a systematic evaluation on medical Q/A datasets. Health Information Science and Systems. 2025 Nov 21;14(1):2.
6) Sarkar PR, Kudapa SP. Systematic Review of Stress And Burnout Interventions Among US Healthcare Professionals Using Advanced Computing Approaches. Journal of Sustainable Development and Policy. 2024 Dec 24;3(04):101-32.
7) Pavuluri S, Sangal R, Sather J, Taylor RA. Balancing act: the complex role of artificial intelligence in addressing burnout and healthcare workforce dynamics. BMJ Health & Care Informatics. 2024 Aug 24;31(1):e101120.
8) Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR medical education. 2023 Feb 8;9(1):e45312.
9) Jaleel A, Aziz U, Farid G, Bashir MZ, Mirza TR, Abbas SM, Aslam S, Sikander RM. Evaluating the potential and accuracy of ChatGPT-3.5 and 4.0 in Medical Licensing and In-Training Examinations: systematic review and meta-analysis. JMIR Medical Education. 2025 Sep 19;11(1):e68070.
10) Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT–3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. JAAOS-Journal of the American Academy of Orthopaedic Surgeons. 2023 Dec 1;31(23):1173-9.
11) Liu J, Gu J, Tong M, Yue Y, Qiu Y, Zeng L, Yu Y, Yang F, Zhao S. Evaluating the agreement between ChatGPT-4 and validated questionnaires in screening for anxiety and depression in college students: a cross-sectional study. BMC psychiatry. 2025 Apr 10;25(1):359.
12) Soddu M, De Vito A, Madeddu G, Nicolosi B, Provenzano M, Ivziku D, Curcio F. Assessing the Accuracy, Completeness and Safety of ChatGPT-4o Responses on Pressure Injuries in Infants: Clinical Applications and Future Implications. Nursing Reports. 2025 Apr 14;15(4):130.
13) Jaleel A, Aziz R, Farid G, Bashir MZ. The impact of ChatGPT on academic integrity in medical education: a developing nation perspective. InFrontiers in Education 2025 May 21 (Vol. 10, p. 1554444). Frontiers Media SA.
14) Farid G, Warraich NF, Iftikhar S. Digital information security management policy in academic libraries: A systematic review (2010–2022). Journal of Information Science. 2025 Aug;51(4):1000-14.
15) Jo E, Song S, Kim JH, Lim S, Kim JH, Cha JJ, Kim YM, Joo HJ. Assessing GPT-4's performance in delivering medical advice: comparative analysis with human experts. JMIR Medical Education. 2024 Jul 8;10(1):e51282.
16) Beaulieu-Jones BR, Berrigan MT, Shah S, Marwaha JS, Lai SL, Brat GA. Evaluating capabilities of large language models: performance of GPT-4 on surgical knowledge assessments. Surgery. 2024 Apr 1;175(4):936-42.

17)Grilo A, Marques C, Corte-Real M, Carolino E, Caetano M. Assessing the quality and reliability of chatgpt's responses to radiotherapy-related patient queries: Comparative study with gpt-3.5 and gpt-4. JMIR cancer. 2025 Apr 16;11(1):e63677.
18)Ullah R, Shaikh MS, Shahani N, Lone MA, Fareed MA, Zafar MS. Comparing ChatGPT and Dental Students' Performance in an Introduction to Dental Anatomy Examination: A Cross-Sectional Study. European Journal of Dentistry. 2025 May 13;17.
19)Fatima A, Shafique MA, Alam K, Ahmed TK, Mustafa MS. ChatGPT in medicine: A cross-disciplinary systematic review of ChatGPT's (artificial intelligence) role in research, clinical practice, education, and patient interaction. Medicine. 2024 Aug 9;103(32):e39250.
20)Wang X, Sanders HM, Liu Y, Seang K, Tran BX, Atanasov AG, Qiu Y, Tang S, Car J, Wang YX, Wong TY. ChatGPT: promise and challenges for deployment in low-and middle-income countries. The Lancet Regional Health–Western Pacific. 2023 Dec 1;41.
21)Umar M, Ali V, Shamim L, Musharaf I, Hafsa R, Ahsan MU, Ahmad O, Sabhan LB, Saeed M, Ahmed S, Iftikhar S. Transforming healthcare with large language models: Current applications, challenges, and future directions—a literature review. Journal of Intelligent Medicine. 2025.
22)Yeasmin S, Semi MM, Rony MK, Das S, Sabeena AA, Rahman R, Biswas B, Ahmed F, Hossain A. Artificial Intelligence for Mental Health Monitoring: A Solution for Digital Behavioral Health Care and Education—An Umbrella Review. Health Science Reports. 2026 Jan;9(1):e71703.
23)Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. Journal of Medical Systems. 2024 Feb 17;48(1):22.
24)Rane N, Choudhary S, Rane J. Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. Journal of Applied Artificial Intelligence. 2024 Mar 20;5(1):69-93.
25)Ong JC, Chang SY, William W, Butte AJ, Shah NH, Chew LS, Liu N, Doshi-Velez F, Lu W, Savulescu J, Ting DS. Ethical and regulatory challenges of large language models in medicine. The Lancet Digital Health. 2024 Jun 1;6(6):e428-32.
26)Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LT, Li YC. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. Journal of Biomedical Science. 2025 May 7;32(1):45.