BMC Medical Education

RESEARCH

Open Access

Integrating AI into clinical education: evaluating general practice trainees' proficiency in distinguishing AI-generated hallucinations and impacting factors

Jiacheng Zhou[1,2], Jintao Zhang[1,2], Rongrong Wan[1,2], Xiaochuan Cui[1,2], Qiyu Liu[1,2], Hua Guo[1,2], Xiaofen Shi[1,2], Bingbing Fu[3], Jia Meng[4], Bo Yue[5], Yunyun Zhang[1,2,3,6*†] and Zhiyong Zhang[1,2,3,6*†]

Abstract

Objective To assess the ability of General Practice (GP) Trainees to detect AI-generated hallucinations in simulated clinical practice, ChatGPT-4o was utilized. The hallucinations were categorized into three types based on the accuracy of the answers and explanations: (1) correct answers with incorrect or flawed explanations, (2) incorrect answers with explanations that contradict factual evidence, and (3) incorrect answers with correct explanations.

Methods This multi-center, cross-sectional survey study involved 142 GP Trainees, all of whom were undergoing General Practice Specialist Training and volunteered to participate. The study evaluated the accuracy and consistency of ChatGPT-4o, as well as the Trainees' response time, accuracy, sensitivity (d'), and response tendencies (β). Binary regression analysis was used to explore factors affecting the Trainees' ability to identify errors generated by ChatGPT-4o.

Results A total of 137 participants were included, with a mean age of 25.93 years. Half of the participants were unfamiliar with AI, and 35.0% had never used it. ChatGPT-4o's overall accuracy was 80.8%, which slightly decreased to 80.1% after human verification. However, the accuracy for professional practice (Subject 4) was only 57.0%, and after human verification, it dropped further to 44.2%. A total of 87 AI-generated hallucinations were identified, primarily occurring at the application and evaluation levels. The mean accuracy of detecting these hallucinations was 55.0%, and the mean sensitivity (d') was 0.39. Regression analysis revealed that shorter response times (OR = 0.92, $P = 0.02$), higher self-assessed AI understanding (OR = 0.16, $P = 0.04$), and more frequent AI use (OR = 10.43, $P = 0.01$) were associated with stricter error detection criteria.

†Yunyun Zhang and Zhiyong Zhang contributed equally to this work.

*Correspondence:
Yunyun Zhang
zhangyunyun026133@njmu.edu.cn
Zhiyong Zhang
zhangzhiyong0112@njmu.edu.cn

Full list of author information is available at the end of the article

Conclusions The study concluded that GP trainees faced challenges in identifying ChatGPT-4o's errors, particularly in clinical scenarios. This highlights the importance of improving AI literacy and critical thinking skills to ensure effective integration of AI into medical education.

Keywords ChatGPT-4o generated hallucinations, General practice (GP) trainees, General practice specialist training, Response bias

Introduction

Artificial intelligence (AI) is increasingly being adopted in medical education, particularly in postgraduate training, with significant potential to advance educational practices [1, 2]. AI refers to the development of systems capable of performing tasks that typically require human intelligence, such as language comprehension, image recognition, and decision-making. A subset of AI, large language models (LLMs) focus on processing and generating natural language text. ChatGPT, a generative AI chatbot, is produced through sophisticated fine-tuning of an LLM [3]. The strong performance of GPT-4 in medical tests suggests that LLMs may serve as valuable teaching tools for students who are currently performing at lower levels in these tests [4]. These models demonstrate high accuracy in explaining medical problems and conducting clinical assessments [5, 6]. As evaluation tools, LLMs provide deep insights into examination results, offering educators a real-time understanding of student learning challenges [7]. However, LLMs are trained on large datasets of mixed quality, which can lead to issues such as disparities, biases, and incorrect associations [8–10], commonly referred to as "hallucinations" [11]. These hallucinations may appear logical initially but are often misleading and cannot be entirely eliminated [12]. Clinicians may misinterpret AI-generated recommendations, especially if the reasoning behind the AI's decision-making is unclear or difficult to understand [13]. However, limited research exists on how effectively general practice residents can detect these AI-generated inaccuracies during their training.

General practitioners play a pivotal role in managing patient health and early disease diagnosis, serving a diverse population across various specialties. As AI becomes more integrated into clinical environments, general practitioners will encounter clinical content and question-and-answer data spanning multiple disciplines. This requires careful interpretation of the information while remaining cautious of potential AI biases [14, 15]. Furthermore, studies show that GPT's accuracy in addressing general practice-related queries often falls below the passing mark, with significant discrepancies across different fields and sources. This highlights the need for general practitioners, especially GP trainees to possess not only a strong foundation in medical knowledge and skills but also a reliable cognitive framework and the ability to identify erroneous information [16, 17]. However, there is a lack of research on how GP trainees recognize and manage hallucinations generated by large language models (LLMs). Therefore, understanding how clinicians and trainees perceive and address these inaccuracies is crucial for developing targeted training programs and ensuring the safe and effective application of LLMs in primary care settings.

Our study evaluated GP trainees' ability to recognize and manage hallucinations generated by ChatGPT-4o. We also investigated factors influencing this ability, such as AI usage frequency, clinical experience, and familiarity with AI. This research explored the key factors that enhance trainees' ability to identify inaccuracies generated by ChatGPT-4o during General Practice Specialist Training, aiming to provide valuable insights for developing future educational frameworks that integrate AI into clinical training.

Methods

Study design

Our research utilized an electronic questionnaire (Supplementary Materials 1) to assess the ability of GP trainees to identify hallucinations generated by ChatGPT-4o. ChatGPT-4o's performance was evaluated through a cross-sectional survey, with human experts analyzing and determining the characteristics of each question. The generated responses and explanations from ChatGPT-4o were then collected and evaluated for accuracy and consistency, comparing the alignment between its answers and explanations to identify any hallucinations. The 'stimulus input' consisted of 50.0% incorrect and 50.0% correct answers, after which GP trainees made a judgment and decision, followed by the 'electronic questionnaire' as the response output. Ethical approval was obtained for this study, and all participants provided informed consent before completing the questionnaire (Supplementary Fig. 1).

Participant settings

The study involved four centers located in the southern and northern regions of China. The recruitment announcement targeted GP trainees undergoing General Practice Specialist Training who were willing to participate in ChatGPT-4o generated responses. To prevent participants from completing tasks such as reviewing materials independently, the tasks were conducted in group settings for each center. Participants from Wuxi

People's Hospital in Jiangsu, First Affiliated Hospital of Jiamusi University, Second Affiliated Hospital of Harbin Medical University, and Second Affiliated Hospital of Qiqihar Medical University in Heilongjiang undergoing standardized general practice training voluntarily participated in the study. Exclusion criteria: Individuals who were absent due to illness, left on the day of the study, or were unwilling to participate were excluded. Additionally, participants who did not complete the training as scheduled or provided inaccurate data were also excluded. The remaining participants were included in the research.

Selection of LLMs

ChatGPT and its LLM counterparts are the tools most frequently mentioned by researchers when asked to provide the most impressive or useful examples of AI in science. ChatGPT also ranked the highest among researchers as the most used AI in science. There are now hundreds of versions of GPT models, and our experiments conducted up to the submission date indicate that GPT-4o outperforms GPT-4, Claude, and Gemini 1.0, in terms of clinical question response accuracy and cognitive functionality [5, 6, 18–21]. Considering the growing user base, the development and application of ChatGPT have advanced significantly, with an increasing number of scholars adopting GPT as their primary tool for daily use and research focus [18], setting it apart from other LLMs [19]. The tension between its broad applicability and the occurrence of hallucinated content led us to choose GPT-4o [20]. Additionally, we selected GPT-4o because our question bank includes image content, which was only supported by GPT-4o at the time this research was conducted.

Determining characteristics of questions

A cross-sectional survey using ChatGPT (GPT-4o) simulated responses for the 2024 Intermediate Attending Physician Examination in General Medicine [21]. The responses consisted of four sections: three with 100 single-choice questions each and one with 18 clinical cases (86 multiple-choice questions), totally 386 questions. Miller's pyramid, proposed in 1989, outlines four levels of medical education assessment: knowledge (knowing), comprehension (knowing how), application (how), and evaluation (doing) [22]. Key assessment strategies for evaluating learners' clinical reasoning abilities were outlined, spanning knowledge acquisition in real-world applications in the clinical setting [23]. At the same time e-assessment methods could replace currently conventional methods [24]. This study categorized the original questions according to Miller's pyramid to assess specific medical competencies and analyzed how different cognitive levels affect the accuracy of ChatGPT-4o generated responses and participants' ability to identify errors. Three senior GPs evaluated these conversations. In cases where the two GPs disagreed, a third GP resolved the tie [25].

Prompts for ChatGPT answers and explanations

Prompt engineering has highlighted the potential of LLMs as effective tools in clinical medicine [26]. To optimize the responses generated by these models, we applied the Relevance, Objectives, Tasks (ROT) framework tailored to the context and specific goals of the task [27].

Relevance clearly defines the domain of the query, such as "Assume the role of an experienced clinician". *Objectives* articulate the aim of the query, such as, "This is a mid-level medical examination; respond to the questions provided with your best effort". *Tasks* specify the exact task to be performed, such as "Offer detailed explanations". This framework was designed to improve the relevance and quality of the responses generated by LLMs in clinical scenarios, thereby enhancing their effectiveness as tools for medical education and practice.

Evaluation of ChatGPT-4o's hallucinations by experts

The criteria for selecting experts, as referred to in one study, were as follows [25]:

1. Have a background in general medicine.
2. Hold a national senior professional title or higher.
3. Serve as an associate professor or higher at a medical school.
4. Have more than 10 years of clinical teaching experience.

Before evaluating the responses from ChatGPT-4o, consistency training was conducted. The training content included the definition of AI hallucination and how to evaluate the hallucinations. The responses were categorized into four types based on the correctness of the answers and explanations. The number of hallucinations and final decision-making were based on the two experts mentioned. In cases of disagreement, a third expert was involved in voting, and the final evaluation result was determined by a majority vote (two votes in favor). The flowchart of the process for expert evaluation of GPT-generated responses was shown in (Supplementary Fig. 2).

Detection of reponses by GP trainees

The errors and correct responses were numbered and randomly assigned using the random function, with the subjects categorized into: "Basic knowledge," "Related professional knowledge," and "Professional knowledge" as subjects 1–3, and "Professional practice" as subject 4. The participants completed the questions using a

standardized approach, and their responses were collected through an electronic survey. Signal detection theory (SDT) was employed to analyze the experimental data [28]. Each participant selected one of the three options (Correct, Incorrect, or Uncertain) to assess the accuracy of each response and explanation. The "hit" refers to correctly identifying correct choice when it's present (HR), while a "false alarm" is incorrectly detecting an incorrect answer when it's absent (FAR). A "miss" is failing to detect the correct answer when it's present, and a "correct rejection" is correctly identifying the absence of the answer. Sensitivity (d') measured the participant's ability to correctly identify true positives. This was calculated as Z (Hit Rate) − Z (False Alarm Rate). Response bias (β) evaluated the tendency of participants to favor one response option over others. Likelihood ratio (β) is related to the hit rate and false alarm rate; it can be computed using the odds ratio. The ratio of correct answers to incorrect hallucinated answers was 1:1, and the judgment of student responses was also a basic binary decision in signal-detection tasks. For this project, a value below the threshold of 1 indicated that the ability to correctly identify signals was weaker than random selection, reflecting a more lenient and trusting response tendency toward GPT responses. Conversely, a value above the threshold of 1 suggests a tendency toward a cautious judgment of GPT responses, so individuals were categorized into two groups depending on whether the β value is equal to, greater than, or less than 1.

Statistical analysis

All statistical analyses were performed using IBM SPSS Statistics, version 26.0. For normally distributed data, the t-test was used for two groups and one-way ANOVA for multiple comparisons, using least significant difference (LSD) or Dunnett's test for pairwise comparisons. For categorical data, chi-square tests and Bonferroni correction were used to adjust for multiple comparisons among group rates. Binary logistic regression analysis was performed to explore the factors influencing the participants' response bias. All statistical tests were two-sided, with a p-value below 0.05 considered statistically significant.

Results

Characteristics of GP trainees and identification of hallucinations by exports

After the initial screening of 142 collected questionnaires (from a total of 148 GP trainees, with 6 on leave), five questionnaires with all correct answers were excluded. The participants whose average age was 25.93, consisted of 46.7% males and 81.8% undergraduates; 45.3% were from Jiangsu. The group comprised 41 participants from the class of 2021, 43 from 2022, and 53 from 2023. Among them, 73 individuals successfully passed the Occupational Medical Examination. Regarding familiarity with and usage of AI, 52.6% were unfamiliar with AI; 35.0% had never used AI; and only 8.0% used AI frequently in academic research. On a 0–10 scale, with higher scores indicating greater satisfaction, the average satisfaction score for AI usage was 7.0. ChatGPT-4o's overall accuracy was 80.8%, which slightly decreased to 80.1% after human verification. However, the accuracy for professional practice (Subject 4) was only 57.0%, and after human verification, it dropped further to 44.2% (Supplementary Table 1). Of the 386 questions answered by ChatGPT-4o, 312 were correct, and 74 were incorrect. Of the 312 correct answers, 299 provided accurate explanations. However, 13 explanations were deemed incorrect because of common sense inconsistencies, calculation errors, or logical flaws (Fig. 1). Regarding knowledge, comprehension, and application levels, the performance of ChatGPT-accuracy after expert verification was 89.3%, 86.6%, and 77.1%, respectively. For the evaluation level, it was 30.4%.

Recognition of ChatGPT-4o generated hallucinations by GP trainees

Our study found that the median hit, false alarm, correct rejection, and miss rates of in GP Trainees identifying generative hallucinations were 86.0%, 74.0%, 26.0%, and 14.0%, respectively. Overall accuracy was 55.0% at a sensitivity of 0.39. A comparison of accuracy in identifying different subjects and cognitive levels among GP Trainees showed that recognition accuracy for Subject 4 was significantly lower than both the overall accuracy ($P < 0.05$) and the accuracy for Subjects 1–3 ($P < 0.001$). As the cognitive level increased, the accuracy of identification gradually decreased. The accuracy at the evaluation level was the lowest among the four cognitive levels ($P < 0.001$), and the difference was statistically significant (Fig. 2A). The sensitivity analysis for identifying various subjects and cognitive levels among GP Trainees revealed that recognition sensitivity for Subject 4 was lower than that for Subjects 1–3 ($P < 0.05$). As cognitive levels increased, the sensitivity to identification gradually improved. Sensitivity regarding understanding was stronger than for the basic knowledge level, whereas sensitivity for the application and evaluation aspects was higher than that for the knowledge and understanding levels ($P < 0.001$) (Fig. 2B). A comparison of response tendencies among GP Trainees in identifying different subjects and cognitive levels showed that they applied stricter criteria when evaluating Subject 4 than Subjects 1–3 ($P < 0.05$). There were no statistically significant differences in the selection criteria across different cognitive levels (Fig. 2C). Further statistical analysis of the number of individuals selecting different criteria showed that the number of those choosing neutral and strict standards for Subject 4 was

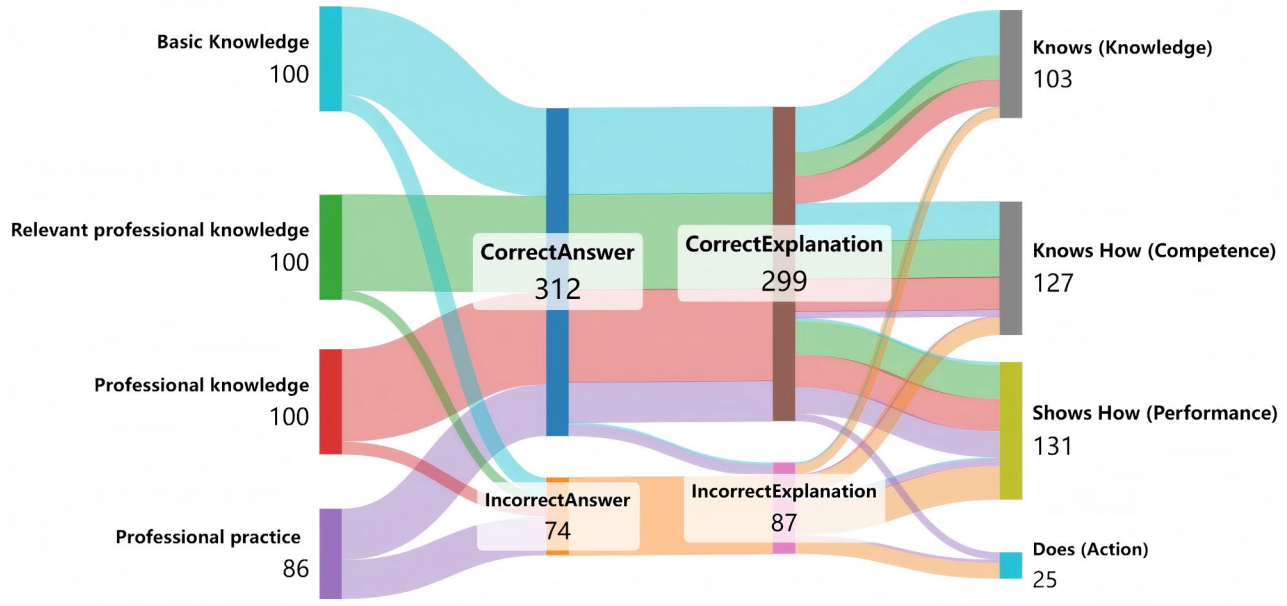

Fig. 1 ChatGPT-4o responses categorization into four types based on the correctness of the answer and the explanation

significantly higher than those choosing Subjects 1–3 and the overall total ($P<0.001$, all). The results of the classification by different cognitive levels indicated that more residents chose neutral and strict standards at the evaluation level than at the application level, with a statistically significant difference ($P<0.001$) (Fig. 2D).

Factors impacting response bias in hallucination identification

This study investigated the factors influencing GP Trainees' choices of liberal, neutral, or conservative criteria for assessing ChatGPT-4o generated responses. Among the variables analyzed, only AI-usage frequency showed significant differences ($P=0.02$). Factors such as age, sex, degree, institution, grade, familiarity, use cases, and user satisfaction had no significant impact. The participants who met stricter criteria tended to use AI more frequently (Table 1). Binary regression analysis identified response time, AI familiarity, and AI usage frequency as key factors affecting the choice of neutral and stricter criteria. Specifically, shorter response times, less familiarity with AI, and frequent AI usage were associated with stricter criteria ($P=0.02$, $P=0.04$, $P=0.01$, respectively) (Table 2).

Discussion

A total of 137 participants were recruited from four centers spanning the northern and southern regions. General Practice Specialist Training in China is a key program designed to cultivate skilled GPs for primary healthcare. GP Trainees undergoing this training encounter AI-generated clinical data and support information throughout chronic disease management [29, 30], necessitating the integration, understanding, and coordination of patient information from all healthcare professionals, environments, and timelines [15].

Our research confirmed the existence of hallucinations in ChatGPT-4o generated responses. Moreover, the probability of hallucinations varied across different cognitive levels. Specifically, the probability of hallucinations at the evaluative level was 69.6% and ranged from 10.7 to 22.9% at other cognitive levels. For single-text discharge summaries, the AI can function without hallucinations [31]. In scenarios involving creative tasks, hallucinations were more likely to occur [32]. Similarly, Aljamaan et al. found that AI chatbots have higher hallucination scores with complex prompts, especially regarding the relevance of the prompt keywords, which is similar to our findings. Our study also showed that in complex clinical scenarios, such as differential diagnosis or treatment planning, the chance of hallucinations is higher than that in basic knowledge queries [33]. Huang et al. showed that AI chatbots are prone to generating hallucinations, particularly in the medical field. Given the risk of hallucinations, it is important to verify all the answers and recommendations generated by these models [34]. This suggests that the use of ChatGPT in medical education for complex clinical cases requires further optimization of the dataset and expert review. Goddard also emphasized the need for extra caution when using information generated by ChatGPT in the biomedical field because it may produce hallucinations and provide inaccurate medical information [35].

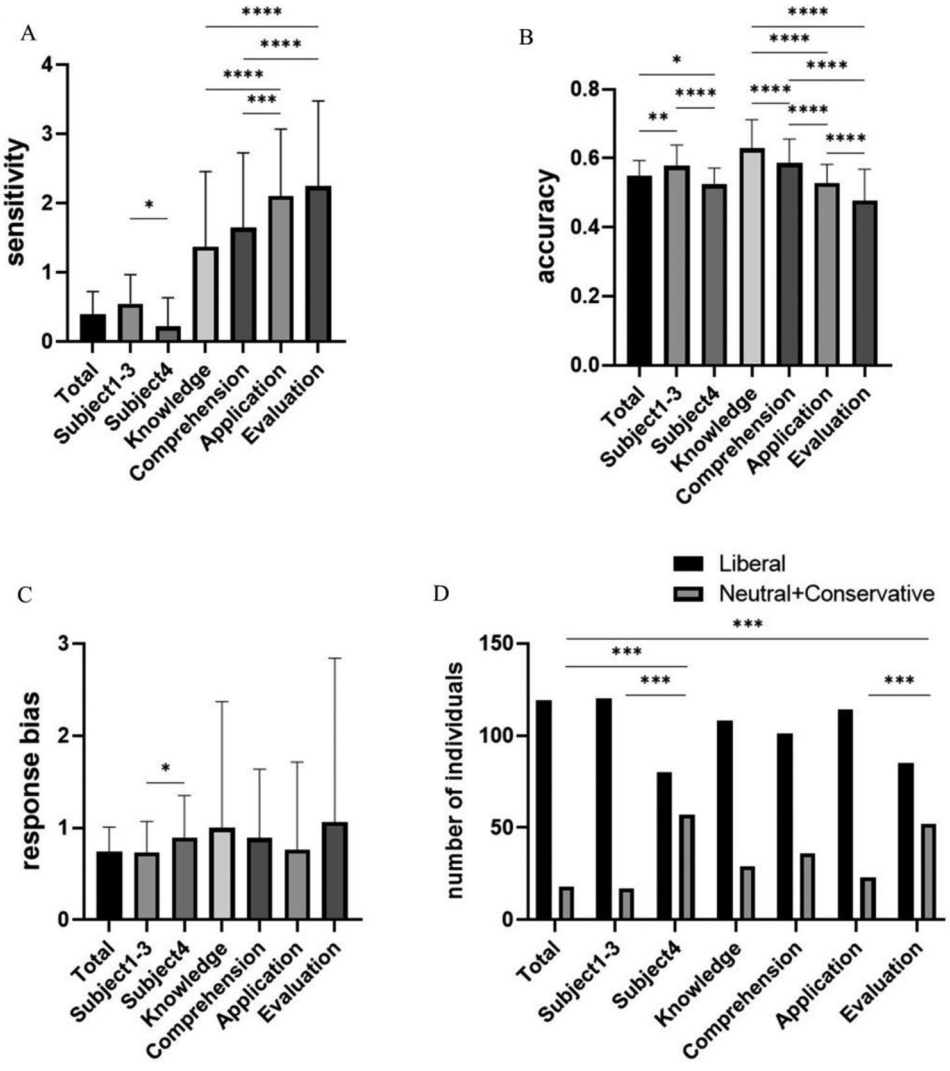

Fig. 2 (**A**-**D**) Sensitivity, accuracy and propensity to response of hallucinations among GP trainees. The analysis focused on the differences in sensitivity (Fig. 2A), accuracy (Fig. 2B), and response tendencies (Fig. 2C and D) between the overall group and individual subjects, as well as across different cognitive levels(*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$)

Our study further examined the ability of GP Trainees to recognize ChatGPT-4o generated hallucinations. The results of this study indicate that although the hit rate was as high as 86.0%, the false alarm rate was as high as 74.0%, and correct rejection rate was as low as 26.0%. The accuracy for Subject 4 was lower than for Subjects 1–3. The sensitivity for Subject 4 was lower than that for Subjects 1–3. For different cognitive levels, the accuracy of identification decreased as cognitive level increased. The results are different from those for accuracy in that sensitivity increased as cognitive level increased. Specifically, sensitivity for application and evaluation was higher than that for knowledge and comprehension. The reason for this might be that accuracy is equal to the hit rate plus the correct rejection rate. Sensitivity is equal to the hit rate minus the false alarm rate. The residents struggled to evaluate ChatGPT-4o generated responses for two main reasons: (i) the gap between theoretical knowledge and clinical experience and (ii) over-reliance on automated tools [36–38]. When AI provided correct information, GPs significantly improved their accuracy in diagnosing skin lesions. However, when AI provided incorrect information, most GPs failed to correctly identify and reject the erroneous diagnosis. GPs with dermatology knowledge were more effective at rejecting AI's incorrect insights [39]. They tended to depend on automated systems for problem solving, which limited their ability to think critically, thereby overlooking the limitations of the ChatGPT-4o generated answers in relying too heavily on them. ChatGPT received high ratings for usability, accuracy, completeness, and usefulness in residency training in one study in China [40].

Our findings on response tendencies revealed that participants applied stricter standards when evaluating

Table 1 Factors impact response Bias of identification of hallucinations among GP trainees

Characteristic	Liberal (*n* = 119)	Neutral+ Conservative (*n* = 18)	Value (t or X^2)	*P*
age	25.87 ± 2.10	26.33 ± 2.01	-0.87	*0.39*
Gender (male%)	54(45.0%)	10(55.6%)	0.65	*0.42*
Degree(undergraduate%)	96(80.7%)	16(88.9%)	0.71	*0.40*
Institution (Jiangsu%)	54(45.4%)	8(44.4%)	0.01	*0.94*
Grade (2021%)	37(31.2%)	4(22.2%)	1.18	*0.55*
Occupational Physician Examination (Unpassed%)	52(43.7%)	12(66.7%)	3.31	*0.07*
Knowledge of AI (unknown%)	63(52.9%)	9(50.0%)	0.05	*0.81*
Application of AI Utilized (no%)	38(31.9%)	11(61.1%)	5.80	*0.02**
AI use cases (Academic%)	10(8.4%)	1(5.6%)	6.82	*0.15*
User satisfaction	7.09 ± 1.52	6.13 ± 1.12	2.22	*0.05*
Response time	72.82 ± 6.84	67.61 ± 15.21	1.43	*0.17*

Table 2 Binary logistics regression analysis of factors impact neutral and Conservative response Bias

Factor	B	Exp(B)	95%CI	*P*
Response Time	-0.08	0.92	0.86–0.99	0.02*
Knowledge of AI				
Known	-1.09	0.16	0.03–0.88	0.04*
unknown		1[Reference]		
Application of AI				
Utilized	2.35	10.43	1.57–62.05	0.01*
unused		1[Reference]		
Age	0.05	1.06	0.78–1.42	0.73
Gender				
female	0.54	1.72	0.54–5.44	0.36
male		1[Reference]		
Institution				
Heilongjiang	0.38	1.45	0.34–6.22	0.62
Wuxi/Jiangsu		1[Reference]		
Degree				
graduate	0.38	1.46	0.25–8.54	0.68
undergraduate				
Grade (2021)				
2023	0.22	1.25	0.16–9.98	0.84
2022	-0.23	0.79	0.10–6.20	0.79
2021		1[Reference]		
Occupational Physician Examination				
pass	1.19	3.30	0.58–18.64	0.18
not passed		1[Reference]		

Subject 4. While no significant differences were observed in response tendencies across various cognitive levels. Overall, they tended to favor more lenient standards. A comparison of the participants' choices revealed a greater tendency to adopt neutral or strict standards when assessing tasks at higher cognitive levels. Univariate and regression analyses further identified shorter response times, limited familiarity with ChatGPT-4o, and more frequent use of ChatGPT-4o, as key factors influencing the preference for neutral or strict evaluation standards. GP trainees tended to apply more lenient standards when evaluating ChatGPT-4o generated responses in general scenarios but adopted stricter criteria in complex clinical settings. This behavior likely stems from higher trust in AI in non-critical situations, increased risks and responsibility in complex cases, and the need for more careful evaluation to avoid potential harm [41]. Additionally, GP trainees may lack experience in ChatGPT-4o assessments, making them more cautious in high-stakes scenarios. This highlights the importance of educating GP trainees about appropriately using and critically assessing ChatGPT-4o, especially in high-risk contexts. From the psychological perspective of reward and punishment mechanisms, in general scenarios, the cost of applying lenient standards to assess ChatGPT-4o generated responses is low, with minimal consequences, which reduces the perceived punishment. However, in complex clinical scenarios, the high risk of severe consequences such as misdiagnosis or treatment errors increases the perception of punishment, prompting students to apply stricter standards to avoid negative outcomes [16, 42]. This cautious approach acts as a protective mechanism by reducing the moral and professional burden of error. Key risk factors included frequent AI use and a limited understanding of how it works, which led to more stringent evaluation standards. This underscores the importance of understanding the fundamentals and limitations of AI, reducing overreliance on AI, and enhancing clinical reasoning and critical thinking [43–45]. ChatGPT-4o (a type of LLMs) has the potential to support GPs by analyzing data and offering new insights, aiding in earlier disease identification and optimizing diagnostic processes. However, ChatGPT-4o (a type of LLMs) should not replace GP diagnostic skills but rather augment them. Before widespread implementation, it requires thorough evaluation to ensure it enhances health outcomes without increasing patient anxiety or burdening healthcare budgets [14].

This study also has several notable strengths. The study population consisted of GP trainees undergoing General Practice Specialist Training, a critical phase in medical education. Our research not only evaluated the performance of ChatGPT-4o but also examined the GP trainees' ability to identify hallucinations generated by ChatGPT-4o, offering a comprehensive perspective on its application in clinical settings. Participants tended to apply lenient recognition standards, demonstrating low sensitivity to ChatGPT-4o's responses, which underscores the need for specialized training to enhance

critical thinking and evaluation skills. Additionally, there is a clear need to strengthen clinical practice to improve the ability to handle complex cases. Furthermore, GP trainees are more vulnerable to hallucinations than experienced GPs, though they may benefit more from ChatGPT's support [13, 42]. This highlights the importance of medical educators ensuring that GP trainees understand the limitations of Chatbot tools and the necessary precautions when integrating them into clinical practice. Moreover, the low recognition accuracy suggests the need for stricter usage guidelines for ChatGPT-4o and clearer markers of uncertainty within the system to better distinguish valid recommendations from potential errors.

Limitations

Despite some results, this study also had several limitations. First, the reliance on a specific AI system (ChatGPT-4o) may limit the generalizability of the findings to other AI models. Second, the assessment, conducted through closed-ended questions, ranging from basic knowledge to complex clinical scenarios, may not fully reflect real clinical situations. Third, the basic knowledge section offered an objective and accurate assessment of the model's performance. In complex cases, hallucination evaluations were conducted by three experts reviewing the answers and annotations. Prior to the assessment, we conducted consistency training without employing any additional scales for evaluation. Finally, although participants were selected from both southern and northern training units, a small number of individuals cannot represent the diverse educational levels over the country.

Conclusion

GP trainees have limited ability to recognize ChatGPT-4o generated hallucinations, particularly in complex clinical scenarios. GP trainees may overestimate their comprehension of clinical scenario recommendations generated by ChatGPT-4o. Therefore, in medical education, assessing medical students' ability to identify hallucinations is crucial for laying a solid foundation for designing relevant training programs.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-025-06916-2.

Supplementary Material 1

Supplementary Material 2: Flowchart of the Study.

Supplementary Material 3: Flowchart of Senior General Practitioners' Evaluation of ChatGPT-4o's responses.

Supplementary Material 4

Acknowledgements
We are very grateful to all the residents who participated in this research.

Author contributions
Yunyun Zhang and Zhiyong Zhang had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Conceive and design: Jiacheng Zhou, Xiaochuan Cui, Yunyun Zhang and Zhiyong Zhang. Acquisition, analysis, or interpretation of data: All authors. Statistical analysis: Jintao Zhang and Yunyun Zhang. Drafting of the manuscript: Jiacheng Zhou and Yunyun Zhang. Obtained funding: Yunyun Zhang and Jiacheng Zhou. Supervision: Yunyun Zhang and Zhiyong ZhangAll authors have given approval to the final version of the manuscript and agree to be accountable for all aspects of the work.

Funding
This study was funded by the "Top Talent Support Program for Young and Middle-aged people of Wuxi Health Committee (BJRC-8 and HB2023015, Dr Zhang); Taihu Light Basic Research Project, Wuxi Municipal Science and Technology Bureau (K20221023, Dr Zhang); Emerging Discipline Leader Program (2024-YZ-HBDTR-ZYY-2024, Dr Zhang); Wuxi Association for Science and Technology (KX-24-C154; Jiacheng Zhou).

Data availability
The data collected for this study were obtained from four different hospitals and were managed by the educational department. We are able to provide the original data upon request. However, a formal application via email is required, which must be signed or stamped by the requesting individual or institution. Specification the exact data items are needed. Once approved, we will translate the original data into English and provide it accordingly.

Declarations

Ethical approval
Ethical approval has been obtained for this study by Research Ethics Committee of Wuxi People's Hospital, and all participants have provided informed consent.

Consent for publication
The authors provide their consent for the publication of this article.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of General Practice, The Affiliated Wuxi People's Hospital of Nanjing Medical University, Wuxi, Jiangsu, China
[2]Wuxi Medical Center, Nanjing Medical University, Wuxi People's Hospital, Wuxi, Jiangsu, China
[3]Department of Postgraduate Education, The First Affiliated Hospital of Jiamusi University, Heilongjiang, China
[4]Department of General Practice, The Second Affiliated Hospital of Harbin Medical University, Heilongjiang, China
[5]Residency Training Center, The Second Affiliated Hospital of Qiqihar Medical University, Heilongjiang, China
[6]Education Department, The Affiliated Wuxi People's Hospital of Nanjing Medical University, Wuxi Medical Center, Wuxi People's Hospital, Qingyang road 299, Wuxi, China

Received: 25 October 2024 / Accepted: 24 February 2025
Published online: 19 March 2025

References

1. Mekki YM, Zughaier SM. Teaching artificial intelligence in medicine. Nat Rev Bioeng. 2024;2:450–1.
2. Yan M, Cerri GG, Moraes FY. ChatGPT and medicine: how AI Language models are shaping the future and health related careers. Nat Biotechnol. 2023;41:1657–8.
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large Language models in medicine. Nat Med. 2023;29:1930–40.
4. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language models in medicine:the potentials and pitfalls. Ann Intern Med. 2024;177:210–20.

5. Cheong RCT, Pang KP, Unadkat S, Mcneillis V, Williamson A, Joseph J, et al. Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT versus Google Bard. European archives of oto-rhino-laryngology: official journal of the European federation of Oto-Rhino-Laryngological societies (EUFOS): affiliated with the German society for Oto-Rhino-Laryngology -. Head Neck Surg. 2024;281:2137–43.
6. Tripathi S, Patel J, Mutter L, Dorfner FJ, Bridge CP, Daye D. Large Language models as an academic resource for radiologists stepping into artificial intelligence research. Curr Probl Diagn Radiol. 2024;S0363–0188(24):00232–9.
7. Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng P-C, et al. ChatGPT and large Language models in academia: opportunities and challenges. BioData Min. 2023;16:20.
8. Pfohl SR, Cole-Lewis H, Sayres R, Neal D, Asiedu M, Dieng A, et al. A toolbox for surfacing health equity harms and biases in large Language models. Nat Med. 2024;30:3590–600.
9. Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR et al. Socio-Demographic Biases in Medical Decision-Making by Large Language Models:A Large-Scale Multi-Model Analysis. 2024;2024. 10. 29. 24316368.
10. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large Language models. PLOS Digit Health. 2023;2:e0000198.
11. Tran CG, Chang J, Sherman SK, De Andrade JP. Performance of ChatGPT on American board of surgery In-Training examination Preparation questions. J Surg Res. 2024;299:329–35.
12. Herrmann-Werner A, Festl-Wietek T, Holderried F, Herschbach L, Griewatz J, Masters K, et al. Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: Mixed-Methods study. J Med Internet Res. 2024;26:e52113.
13. Jussupow E, Spohrer K, Heinzl A, Gawlitza J. Augmenting medical diagnosis decisions?? An investigation into physicians' decisions?-Making process with artificial intelligence. Inform Syst Res. 2021;32:713–35.
14. Summerton N, Cansdale M. Artificial intelligence and diagnosis in general practice. Br J Gen Practice: J Royal Coll Gen Practitioners. 2019;69 684:324–5.
15. Everson J, Hendrix N, Phillips RL, Adler-Milstein J, Bazemore A, Patel V. Primary care physicians' satisfaction with interoperable health information technology. JAMA Netw Open. 2024;7:e243793.
16. Buck C, Doctor E, Hennrich J, Jöhnk J, Eymann T. General practitioners' attitudes toward artificial Intelligence–Enabled systems: interview study. J Med Internet Res. 2022;24:e28916.
17. Tong L, Wang J, Rapaka S, Garg PS. Can ChatGPT generate practice question explanations for medical students, a new faculty teaching tool?Med. Teach. 2024;1–5.
18. Liu Z, Zhang W. A qualitative analysis of Chinese higher education students' intentions and influencing factors in using ChatGPT: a grounded theory approach. Sci Rep. 2024;14:1–11.
19. Gruda D. Three ways ChatGPT helps me in my academic writing. Nature. 2024. https://www.nature.com/articles/d41586-024-01042-3. Accessed 11 Jan 2025.
20. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J et al. Coding inequity: assessing GPT-4's potential for perpetuating Racial and gender biases in healthcare. 2023;2023.07.13.23292577.
21. Du QF, Wang JJ. 2024 General Medicine Practice Mock Exam. People's Medical Publishing House; 2023:3–39. ISBN:9787117355421.
22. Ten Cate O, Carraccio C, Damodaran A, Gofton W, Hamstra SJ, Hart DE, et al. Entrustment decision making: extending Miller's pyramid. Acad Med. 2021;96:199–204.
23. Thampy H, Willert E, Ramani S. Assessing clinical reasoning:targeting the higher levels of the pyramid. J Gen Intern Med. 2019;34:1631–6.
24. Hasani H, Khoshnoodifar M, Khavandegar A, Ahmadi S, Alijani S, Mobedi A, et al. Comparison of electronic versus conventional assessment methods in ophthalmology residents; a learner assessment scholarship study. BMC Med Educ. 2021;21:342.
25. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA et al. An evaluation framework for clinical use of large Language models in patient interaction tasks. Nat Med. 2025;1–10.
26. Meskó B. Prompt engineering as an important emerging skill for medical professionals:tutorial. J Med Internet Res. 2023;25:e50638.
27. Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med. 2024;7:41.
28. Aujla H. d[Formula:see text]:sensitivity at the optimal criterion location. Behav Res Methods. 2023;55:2532–58.
29. Wang S, Shi Y, Sui M, Shen J, Chen C, Zhang L, et al. Telephone follow-up based on artificial intelligence technology among hypertension patients: reliability study. J Clin Hypertens (Greenwich). 2024;26:656–64.
30. Li J, Guan Z, Wang J, Cheung CY, Zheng Y, Lim L-L, et al. Integrated image-based deep learning and Language models for primary diabetes care. Nat Med. 2024. https://doi.org/10.1038/s41591-024-03139-8.
31. Tung JYM, Gill SR, Sng GGR, Lim DYZ, Ke Y, Tan TF, et al. Comparison of the quality of discharge letters written by large Language models and junior Clinicians:Single-Blinded study. J Med Internet Res. 2024;26:e57721.
32. Zaretsky J, Kim JM, Baskharoun S, Zhao Y, Austrian J, Aphinyanaphongs Y, et al. Generative artificial intelligence to transform inpatient discharge summaries to Patient-Friendly Language and format. JAMA Netw Open. 2024;7:e240357.
33. Aljamaan F, Temsah M-H, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. JMIR Med Inf. 2024;12:e54345.
34. Huang Y, Gomaa A, Semrau S, Haderlein M, Lettmaier S, Weissmann T et al. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and red journal Gray zone cases: potentials and challenges for ai-assisted medical education and decision making in radiation oncology. Front Oncol. 2023;13.
35. Goddard J. Hallucinations in ChatGPT: A cautionary Tale for biomedical researchers. Am J Med. 2023;136:1059–60.
36. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Medicine: J Association Am Med Colleges. 2024;99:22–7.
37. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare:past, present and future. Stroke Vasc Neurol. 2017;2:230–43.
38. Fitzek S, Choi K-EA. Shaping future practices:German-speaking medical and dental students' perceptions of artificial intelligence in healthcare. BMC Med Educ. 2024;24:844.
39. Micocci M, Borsci S, Thakerar V, Walne S, Manshadi Y, Edridge F, et al. Attitudes towards trusting artificial intelligence insights and factors to prevent the passive adherence of GPs: A pilot study. J Clin Med. 2021;10:3101.
40. Shang L, Li R, Xue M, Guo Q, Hou Y. Evaluating the application of ChatGPT in China's residency training education: an exploratory study. Med Teach. 2024;1–7.
41. Li J, Zhou L, Zhan Y, Xu H, Zhang C, Shan F, et al. How does the artificial intelligence-based image-assisted technique help physicians in diagnosis of pulmonary adenocarcinoma?A randomized controlled experiment of multicenter physicians in China. J Am Med Inf Assoc. 2022;29:2041–9.
42. Wang W, Gao G (Gordon), Agarwal R, editors. Friend or Foe? Teaming Between Artificial Intelligence and Workers with Variation in Experience. Management Science. 2024;70:5753–75.
43. Larson BZ, Moser C, Caza A, Muehlfeld K, Colombo LA. Critical thinking in the age of generative AI. AMLE. 2024;23:373–8.
44. Moulin TC. Learning with AI Language models: guidelines for the development and scoring of medical questions for higher education. J Med Syst. 2024;48:45.
45. Student interaction with. ChatGPT can promote complex critical thinking skills. Learn Instruction. 2025;95:102011.

Publisher's note