Original Paper

Quality Assessment of Large Language Model–Generated Medical Dialogue for Clinical Vignettes: Evaluation Study

Yasutaka Yanagita, MD, PhD; Daiki Yokokawa, MD, PhD; Shiichi Ihara, MD; Ryo Yoshida, MD; Yoshihide Okano, MD; Takanori Uehara, MD, PhD

Department of General Medicine, Chiba University Hospital, Chiba, Japan, Chiba, Japan

Corresponding Author:
Yasutaka Yanagita, MD, PhD
Department of General Medicine
Chiba University Hospital, Chiba, Japan
1-8-1, Inohana, Chuo-ku
Chiba, 260-8677
Japan
Phone: 81 43 222 7171 ext 6438
Fax: 81 43 224 4758
Email: y.yanagita@gmail.com

Abstract

Background: Traditional clinical vignettes, though widely used in medical education, often focus on prototypical presentations; require substantial time and effort to develop; and fail to represent patient diversity, the complexity of clinical conditions, patients' perspectives, and the dynamic nature of physician-patient interactions.

Objective: This study aimed to evaluate the quality of Japanese-language physician-patient dialogues produced by generative artificial intelligence (AI), focusing on their medical accuracy and overall appropriateness as medical interviews.

Methods: We created an AI prompt that included a specific clinical history and instructed the model to simulate a cooperative patient responding to the physician's questions to generate a physician-patient dialogue. The target diseases were those covered by the Japanese National Medical Licensing Examination. Each dialogue consisted of 25 turns by the physician and 25 by the patient, reflecting the typical volume of conversation in Japanese outpatient settings. Three internists independently evaluated each generated dialogue using a 7-point Likert scale across 6 criteria: coherence of the conversation, medical accuracy of the patient's responses, medical accuracy of the physician's responses, content of the medical history, communication skills, and professionalism. In addition, a composite score for each dialogue was calculated as the overall mean of these 6 criteria. Each dialogue was also examined for the presence of 5 essential clinical components commonly included in medical interviews: chief concern and clinical course since onset, physical findings, test results, diagnosis, and treatment course. A dialogue was considered to include a component only if all 3 evaluators independently confirmed its presence.

Results: The mean composite score was 5.7 (SD 1.0), indicating high overall quality. Mean scores for each criterion were as follows: coherence of the conversation, 5.9 (SD 0.9); medical accuracy of the patient's responses, 6.0 (SD 0.9); medical accuracy of the physician's responses, 5.6 (SD 1.1); content of medical history taking, 5.9 (SD 0.9); communication skills, 5.6 (SD 0.9); and professionalism, 5.5 (SD 1.1). Among the 5 clinical components assessed in each dialogue across 47 clinical cases, chief concern and clinical course were included in all 47 (100%) cases, physical findings in 15 (32%) cases, test results in 27 (57%) cases, diagnosis in 45 (96%) cases, and treatment course in 0 (0%) cases.

Conclusions: While physician oversight remains essential, it is feasible to efficiently create AI-generated educational materials for medical education that overcome the limitations of traditional clinical vignettes. This approach may reduce time and financial burdens, enhancing opportunities to practice clinical interviewing in settings that closely mirror real-world encounters.

(JMIR Form Res 2025;9:e80752) doi: 10.2196/80752

KEYWORDS

ChatGPT; clinical vignettes; artificial intelligence; generative AI; medical education; physician-patient dialogue

Introduction

Artificial intelligence (AI) is critical for the advancement of medicine. In particular, generative AI, exemplified by large language models (LLMs), has the potential to fundamentally transform health care. Tools such as ChatGPT, developed by OpenAI, possess advanced conversational capabilities and broad applicability [1], with an increasing number of applications in the medical field. From the perspective of conversational capabilities, generative AI–based chatbots can provide rehabilitation guidance and mental health support to patients and assist health care professionals in patient management [2]. Generative AI is versatile enough to have passed the Japanese National Medical Licensing Examination [3]. Moreover, LLMS are now used to enter patients' medical histories and verify diagnostic accuracy in English and Japanese [4,5], indicating that the scope of generative AI applications is expected to expand further in the future.

The application of generative AI in medical education has attracted increasing interest. The use of generative AI to produce physician-patient dialogues has the potential to replicate disease-specific communication encountered in clinical practice and serve as a valuable educational resource for medical students.

A comparable and widely used educational tool in medical education research is the *clinical vignette* [6-8]. Vignettes are paper-based instructional resources that present a patient's medical history and key clinical information in an organized written format. They are considered highly effective in helping medical students understand disease concepts, symptomatology, and treatment pathways [9].

Exposure to various vignettes can deepen students' clinical understanding. However, traditional vignettes often lack diversity in case scenarios and tend to focus narrowly on diagnostic information, thereby failing to capture the complexity, ambiguity, and uncertainty that characterize real-world clinical practice [10]. In addition, the development of vignettes requires substantial time and effort. As clinical vignettes are generally written from the health care provider's perspective, they inherently fail to capture the patient's viewpoint and the dynamics of physician-patient interaction. Using generative AI such as ChatGPT to create dialogue-based educational materials that incorporate these missing elements is considered a promising approach. These educational materials offer several advantages. First, ChatGPT can generate a wide variety of dialogues tailored to different diseases and clinical scenarios, enabling repeated practice. Its use can reduce both the time and financial costs associated with producing educational content. Moreover, medical students can engage in clinical reasoning by organizing information while considering multiple differential diagnoses and can practice effective physician-patient communication, as would be required in actual clinical settings.

AI-generated dialogues incorporate the medical information necessary for diagnostic decision-making along with a variety of interactions that closely resemble those encountered in real clinical settings. This enables medical students to develop information-processing skills in authentic, context-rich scenarios. Through this process, students are expected to acquire the ability to identify clinically relevant information and develop a broader range of clinical competencies, including medical interviewing techniques, patient management skills, and effective communication strategies.

This study examined the feasibility of generating physician-patient dialogues using generative AI and evaluated the quality of the generated dialogues. Generative AI has already demonstrated the ability to generate differential diagnosis lists, illness scripts, and clinical vignettes based on medical information [11,12]. Accordingly, it is considered feasible to generate physician-patient dialogues with a high level of accuracy using generative AI.

In this study, we assessed the quality of dialogues generated in Japanese medical interviews, including their medical accuracy, and investigated their potential utility as educational materials in the context of medical education. By demonstrating the feasibility of easily generating diverse physician-patient dialogues using generative AI, it becomes possible to efficiently produce a large volume of educational content for use in medical education under the supervision of physicians with medical expertise. The appropriate application of AI technology is expected to increase the opportunities for medical students to engage with more practical and interactive content, thereby enhancing the effectiveness of medical education.

Methods

Overview

Physician-patient dialogues were generated using an LLM and were subjected to cross-sectional evaluation. To ensure relevance to medical education, 47 target diseases were selected to represent diseases that medical students were expected to learn. The selection was based on the content of the Japanese National Medical Licensing Examination [13] and finalized through discussion between a board-certified internist (YY) and a board-certified family physician (DY; Textbox 1). To evaluate the generated dialogues, we recruited 3 Japanese internists (SI, RY, and YO) who are actively involved in diagnostic practice at a university hospital and regularly manage a wide range of diagnostically challenging cases. These internists assessed the quality of the dialogues.

Textbox 1. List of 47 diseases used to generate and evaluate artificial intelligence–generated physician-patient dialogues.

Diseases

1. Gastroesophageal reflux disease
2. Functional dyspepsia
3. Esophageal achalasia
4. Crohn disease
5. Appendicitis
6. Pheochromocytoma
7. Primary aldosteronism
8. Cushing syndrome
9. Hashimoto disease
10. Subacute thyroiditis
11. Graves disease
12. Depression
13. Panic disorder
14. Gout
15. Pneumothorax
16. Myasthenia gravis
17. Trigeminal neuralgia
18. Parkinson disease
19. Alzheimer disease
20. Benign paroxysmal positional vertigo
21. Migraine
22. Cluster headache
23. Neurocardiogenic syncope
24. Intervertebral disk herniation
25. Insomnia
26. Bronchial asthma
27. Acute eosinophilic pneumonia
28. Rheumatoid arthritis
29. Systemic lupus erythematosus
30. Sjögren syndrome
31. Measles
32. Fibromyalgia
33. Gallstone attack
34. Acute cholecystitis
35. Unstable angina
36. COVID-19
37. Lumbar spinal stenosis
38. Thromboangiitis obliterans
39. Infectious mononucleosis
40. Streptococcal pharyngitis
41. Hepatitis
42. Transient ischemic attack

43. Iron deficiency anemia
44. Heatstroke
45. Acute pericarditis
46. Chronic obstructive pulmonary disease
47. Lung cancer

Production Environment

The analysis was performed on a terminal running Ubuntu (version 20.04.2 Long-Term Support; Canonical). The central processing unit was equipped with an AMD EPYC 7402*P* 24-core processor (Advanced Micro Devices, Inc), with 256 GB of main memory. An NVIDIA A100 graphics processing unit (NVIDIA Corp) with 40 GB of RAM was used for the computations. Python (version 3.6.9) and the OpenAI Python library (version 0.27.8; OpenAI) were used. The application programming interface used to generate the dialogues was the latest one available at the start of the study, gpt-4o-2024-11-20 [14], and the dialogues were generated on November 22, 2024.

Prompts to Be Entered Into the LLM

Referring to previous studies that used AI to generate vignettes and illness scripts [12], concise instructions specifying the desired output conditions were provided to the LLM as prompts (Figure 1). To enhance the completeness and coherence of the generated dialogues, the prompts included conditions such as the user supplying a specific vignette before the generation, refraining from using medical terminology, and responding cooperatively to the physician's questions. The dialogue length was determined with reference to the average volume of conversation in typical outpatient consultations (approximately 5-10 min) [15], and the LLM was instructed to generate 25 turns each from the physician and the patient.

Figure 1. A portion of the prompt used to generate physician-patient dialogues in Japanese using generative artificial intelligence (Top: Japanese, Bottom: English).

dialogue_1 = 状況：医療面接の文章化。ユーザが指示した症例（Vignette）を元に、医師と患者の会話を ChatGPT が生成する。
ユーザの役割：Vignette を提示する。
ChatGPT の役割：会話文の生成を行う。
ChatGPT の役割の詳細：以下の条件に則って生成しなさい。
目的：この会話を客観的に解説し、問診の仕方から診断に至るプロセスを学ぶ。
条件：
1. Vignette に沿った内容で会話を生成すること。
2. 会話はなるべく短い一文にすること
3. 医師の会話は医師：で書き出すこと。「」や…はつけない。
4. 患者の会話は患者：で書き出すこと。「」や"はつけない。
5. 患者の客観的な症状を述べさせるために家族を登場させても良い。特に患者が話す内容に信頼性がないときは、患者家族の発言を医師が聞き出すようにしなさい。患者家族：で書き出すこと。
6. すべての会話は 1 から順番に番号をふること。少なくとも 30 番までは患者が経験した症状やその特徴についての内容にしなさい。
7. 最低でも 50 番まで会話を行いなさい。
8. 患者は医師の質問に Vignette に沿った内容で患者として振る舞い、回答する。
9. 患者は医師に協力的に振る舞う。
10. 患者は、医師に病名を伝えられるまで、自分の診断名は知らない。よって、医師に診断名を伝えることはできない。
11. 患者は、医師に診察所見や検査結果を伝えることはできない。
12. 日本語で生成する。
13. 医師は同じ症状を呈する他の鑑別診断に関する質問も行う。それは結果的に陰性かもしれない。
14. 医師も診断を知らないため、質問内容が診断名と関係ないこともある。
15. 医師は患者の心理・社会に関わることも、必要があれば積極的に質問する。
16. すこし会話に雑談を加えなさい。症状を広げて聞き、関係ない話題が出ても許容しなさい。必ずしも Vignette に含まれる情報だけでなくても良い。
17. Vignette のうち、病歴、症状、診察、検査、診断を用いて生成しなさい。
18. 病歴と症状の文章量を最も増やしなさい。
19. 診断を伝えたら終了しなさい。
20. 方針の説明・経過は不要であり生成してはいけない。
21. 会話以外の文章を出力しない。
22. 治療薬は一般名を日本語で記載する。
23. 治療薬は日本で承認されているもののみ。
24. 患者は医学用語を使わない。一般的な用語を使用する。
25. 内容は教育的なものである必要がある。

dialogue_1 = Situation: Medical Interview Text. The user presents a specified case (Vignette), and ChatGPT generates a conversation between a physician and a patient based on it.
User role: Present the vignette.
ChatGPT role: Generate the dialogue.
ChatGPT should strictly follow the conditions below and not output anything else.
Objective: To learn the reasoning process—from questioning to diagnosis—by observing the physician's conversational approach and flow.
Conditions
1. Generate a conversation that follows the content of the vignette.
2. Keep each dialogue line as short and concise as possible.
3. Prefix physician's utterances with "Doctor:" and do not add quotation marks.
4. Prefix patient's utterances with "Patient:" and do not add quotation marks.
5. It is acceptable to include a family member in the dialogue to describe the patient's objective symptoms. When the patient's statements are unreliable, the physician should elicit information from the family member. Prefix such utterances with "Family:".
6. Start all conversations from "line 1" and continue sequentially. For at least the first 30 turns, the content should focus on the patient's experienced symptoms and their characteristics.
7. Continue the dialogue for at least 50 turns.
8. The physician's questions and the patient's answers must be consistent with the vignette.
9. The patient behaves cooperatively toward the physician.
10. The patient does not know their diagnosis until the physician informs them. Therefore, the patient cannot tell the physician the name of the disease.
11. The physician cannot convey the diagnosis to the patient.
12. The dialogue must be generated in Japanese.
13. The physician should also ask questions related to other possible differential diagnoses presenting with similar symptoms, even if those findings ultimately turn out to be negative.
14. Because the physician does not know the diagnosis, some of their questions may be unrelated to the actual diagnosis.
15. The physician will also ask questions related to the patient's psychological and social aspects when necessary.
16. Add a small amount of casual conversation. Broaden the discussion of symptoms and allow unrelated topics to appear naturally. The content does not have to be limited strictly to what is included in the vignette.
17. Use the vignette's information on medical history, symptoms, physical findings, tests, and diagnosis to generate the dialogue.
18. Give the greatest amount of detail and volume to the patient's history and symptoms.
19. End the dialogue once the diagnosis has been communicated.
20. Explanations of treatment plans or follow-up courses are unnecessary and must not be generated.
21. Do not output any text other than the dialogue itself.
22. Drug names should be written in Japanese using their International Nonproprietary Names.
23. Only use drugs that are approved for use in Japan.
24. The patient should not use medical terminology.
25. The content must have an educational purpose.

Evaluation of Dialogue

The presence of five key components commonly included in medical interviews was assessed: (1) chief concern and clinical course since onset, (2) physical findings, (3) test results, (4) diagnosis, and (5) treatment course. Three physicians independently reviewed each dialogue, and a dialogue was considered to contain all 5 components only if all 3 physicians confirmed their inclusion.

Next, criteria for assessing the quality of the medical interviews were developed. Six evaluation criteria were selected through discussions between a board-certified internist (YY) and a board-certified family physician (DY), with reference to the evaluation domains of the Mini-Clinical Evaluation Exercise (Mini-CEX) [16]: (1) coherence of the conversation, (2) medical accuracy of the patient's statements, (3) medical accuracy of the physician's statements, (4) quality of the physician's history taking, (5) communication skills, and (6) professionalism.

Criterion 1, coherence of the conversation, was evaluated as a linguistic criterion, focusing on the smoothness of the interaction between the physician and patient, the presence of inconsistencies, grammatical or typographical errors, and the logical relationship of the dialogue. Criteria 2 and 3, medical accuracy of the patient's and physician's statements, respectively, were assessed as clinical criteria by evaluating their consistency with the known clinical features of the respective diseases. Criterion 4, history taking, was assessed based on whether the physician elicited information regarding symptom characteristics and exacerbating or relieving factors. Criteria 5 and 6, communication skills and professionalism, respectively, were evaluated by assessing whether the physician

explored the patient's explanatory model and demonstrated respect, compassion, and empathy toward the patient (Table 1).

Following prior studies [12], we specifically employed 3 Japanese physicians affiliated with a university hospital, each of whom was involved in the supervision and education of medical students and residents, to evaluate the generated dialogues. Each of the 6 evaluation criteria was rated on a 7-point Likert scale, based on the perceived educational usefulness of the medical students. The scale was defined as follows: 1=not applicable at all or not useful—major overall revision required, 2=low usefulness—multiple major revisions needed, 3=limited usefulness—some valuable content but substantial revisions necessary, 4=moderate usefulness—both strengths and several areas for improvement, 5=generally useful—some revisions or adjustments desirable, 6=high usefulness—only minor adjustments possibly needed, and 7=extremely useful and complete—no further revisions required.

For each dialogue, the score of each evaluation criterion was calculated as the average of the ratings of the 3 evaluators. The composite score for the dialogue was then derived by averaging the scores across all 6 evaluation criteria and interpreted using the same 7-point Likert scale.

Table 1. Evaluation criteria and definitions used to assess artificial intelligence–generated physician-patient dialogues in medical interviews.

Evaluation criteria	Evaluation details
Coherence of the conversation	Assessed whether the dialogue between the physician and patient proceeded smoothly, with accurate grammar, no typographical or spelling errors, and overall linguistic clarity
Medical accuracy of the patient's statements	Evaluated whether the patient's utterances accurately reflected the typical onset patterns, symptoms, and clinical course associated with the relevant disease
Medical accuracy of the physician's statements	Assessed whether the physician's explanations and other statements were medically accurate and aligned with established clinical knowledge
Quality of the physician's history taking	Evaluated whether the physician asked about essential elements of the current illness, including symptom location, characteristics, severity, temporal course, contextual factors, aggravating and relieving factors, associated symptoms, and the patient's response to the symptoms
Communication skills	Assessed whether the physician conducted the interview in a way that facilitated open communication, explored the patient's explanatory model and psychosocial context, and confirmed the patient's understanding of the information discussed
Professionalism	Evaluated whether the physician demonstrated respect, compassion, and empathy toward the patient and whether a trusting therapeutic relationship was established

Ethical Considerations

This study did not involve human or animal participants, and therefore, ethics approval was not required.

Results

Using the gpt-4o-2024-11-20 model, physician-patient dialogues were generated for 47 clinical cases (Table 2). Among the 47 generated dialogues, clinical component 1, chief concern and clinical course since onset, was present in all 47 (100%) cases; clinical component 4, diagnosis, was included in 45 (96%) cases, and in each of these cases, the model accurately outputted the specified disease name, as instructed. In contrast, clinical component 2, physical findings, was included in 15 (32%) cases; clinical component 3, test results, was included in 27 (58%) cases; and clinical component 5, treatment course, was not included in any of the cases (0%). Regarding the quality of the medical interviews, the average score was 5.9 (SD 0.9) for coherence of the conversation, 6.0 (SD 0.9) for medical accuracy of the patient's statements, and 5.6 (SD 1.1) for medical accuracy of the physician's statements. The average score was 5.9 (SD 0.9) for quality of the physician's history taking, 5.6 (SD 0.9) for communication skills, and 5.5 (SD 1.1) for professionalism. The overall composite score, calculated as the mean of the 6 evaluation criteria, was 5.7 (SD 1.0).

A focused discussion was conducted among 5 physicians, 2 specialists (YY and DY), and 3 evaluators (SI, RY, and YO), centered on dialogues that were subject to point deductions to identify and clarify the specific issues present in the lower-rated dialogues. The results of this analysis are summarized in Table 3. For reference, one dialogue with a perfect average score of 7 and another that received a lower average score of 4 are provided in Multimedia Appendix 1 and Multimedia Appendix 2, respectively.

Table 2. Average scores for each of the 6 evaluation criteria used to assess artificial intelligence–generated physician-patient dialogues.

Evaluation criteria	Average score (SD)
Coherence of the conversation	5.9 (0.9)
Medical accuracy of the patient's statements	6.0 (0.9)
Medical accuracy of the physician's statements	5.6 (1.1)
Quality of the physician's history taking	5.9 (0.9)
Communication skills	5.6 (0.9)
Professionalism	5.5 (1.1)

Table 3. Problems identified for each evaluation criterion based on expert review of lower-rated artificial intelligence–generated physician-patient dialogues.

Evaluation criteria	Problems
Coherence of the conversation	• Responses to patient questions were omitted. • Although some expressions were unnatural, overall coherence was maintained.
Medical accuracy of the patient's statements	• Typical responses regarding aggravating and relieving factors were not provided.
Medical accuracy of the physician's statements	• A diagnosis was rendered, despite the interview being insufficient. • Diagnoses were finalized based on test results that were never mentioned as having been performed, indicating inappropriate or unjustified diagnostic reasoning. • The response is that it cures a disease that cannot be cured.
Quality of the physician's history taking	• Redundant questions were asked regarding information that had already been provided. • When multiple symptoms were present, it was unclear which symptom's aggravating or relieving factors were being discussed. • The structure and content of the medical interview were inadequate. • Additional questioning was conducted about symptoms not initially reported by the patient. • Unnecessary blood tests were included and described without justification.
Communication skills	• No attempts were made to confirm the patient's understanding of the information provided. • The patient's explanatory model was rarely elicited. • The patient's response lacked logical progression or flow.
Professionalism	• No appropriate responses were provided following expressions of anxiety. • The overall focus of the dialogue was limited to diagnostic questioning, with few utterances directed toward building rapport or fostering a therapeutic relationship.

Discussion

Principal Findings

This study aimed to generate physician-patient dialogue using generative AI and evaluate the prompt designs, the resulting outputs, and the overall quality of the dialogues. We created 47 dialogues based on diseases from the Japanese National Medical Licensing Examination and evaluated them using 6 criteria related to clinical communication. The dialogues consistently included the chief concern and clinical course and demonstrated high medical accuracy and coherence, while areas such as professionalism and inclusion of treatment information were less consistently addressed. The overall composite score was 5.7 (SD 1.0), indicating general usefulness with minor revisions required.

The analysis first focused on the outputs generated for 47 clinical cases and the corresponding prompt designs that produced them. Examination of the medical content essential for a clinical interview revealed that all dialogues included descriptions of the chief concern and clinical course since onset and that accurate diagnostic labels were presented in 45 (96%) of the 47 cases. In contrast, none of the dialogues included information regarding the treatment course. In designing the prompts, the number of dialogue turns was set at 25 for each participant (50 in total), considering the average consultation time in Japanese outpatient settings, which is approximately 5 minutes. As an exploratory extension, we tested longer dialogues (50 and 100 turns per speaker). Although brief mentions of treatment began to appear, these dialogues became increasingly verbose, with redundant differential diagnoses and questioning. This indicated a decline in dialogue quality and educational effectiveness, highlighting the need to identify an optimal dialogue length in prompt design. Although the generation of information related to the treatment course may be improved through more sophisticated prompt engineering [17,18] or future updates to generative AI models [19], prior studies have demonstrated that diagnostic reasoning and treatment management involve fundamentally distinct cognitive processes [20]. Therefore, separating these processes into distinct dialogue scenarios may be more educationally effective than integrating them into a single prompt.

The quality of the generated dialogues, produced in Japanese, was evaluated based on 6 criteria. Regarding criterion 1, coherence of the conversation, the dialogues were generally smooth and grammatically correct. This reflects the high-level natural language processing capabilities of generative AI and suggests its potential to replicate physician-patient interactions at a basic level of linguistic fidelity within the context of Japanese-language medical interviews. Regarding criterion 2, medical accuracy of the patient's statements, although there were instances in which the patient did not provide clear responses to the physician's questions, particularly in certain disease contexts, such ambiguity may reflect the nature of real-world clinical encounters. From an educational perspective, these instances may be valuable for simulating authentic dialogue dynamics. Criterion 3, medical accuracy of the physician's statements, received comparatively lower ratings than other criteria. This may be attributable to the perceived need for greater domain-specific precision, particularly when formulating clinical questions and providing medically accurate explanations to patients. This finding aligns with previous research on the limitations of generative AI in clinical reasoning [12]. Regarding criterion 4, quality of the physician's history taking, in actual clinical practice, once the probability of a particular disease increases based on the patient's narrative, physicians typically engage in further in-depth inquiry. Generative AI appears to struggle with appropriately weighing clinical information and identifying which elements warrant further exploration; thus, at present, AI may be limited in its capacity to conduct history taking guided by probabilistic diagnostic reasoning. Regarding criterion 5, communication skills, the dialogues showed limited use of verbal cues such as acknowledgments or responses that convey an understanding of the patient's statements, and expressions of empathy toward patients' concerns were insufficient. In addition, there were a few attempts to elicit a patient's explanatory model, which is a critical step toward building rapport. Prompt design that encourages empathic and interactive responses may help address these limitations in future development. Regarding criterion 6, professionalism, although no ethically inappropriate expressions were identified, the dialogues generally lacked explicit demonstrations of patient-centered attitudes, respect, or empathic concern.

It is important to note that, in face-to-face medical interviews, nonverbal cues, such as facial expressions and nodding, play a substantially role in promoting patient satisfaction and emotional attunement [21,22], and the absence of such elements constitutes a fundamental limitation of text-based dialogue evaluations. Moreover, a small number of dialogues contained clinically inappropriate content, such as failure to provide justification for a diagnosis or incorrect assertion that a typically incurable disease could be cured. However, no ethically problematic or professionally inappropriate statements were identified in the dataset. Taken together, these results, along with the overall composite score of 5.7 (SD 1.0), suggest that, while physician oversight and revision are essential, generative AI–based dialogues may serve as a valuable educational resource for teaching clinical communication skills to medical students and early-stage trainees.

Because of their underlying architecture, LLMs exhibit inherent output variability, and identical prompts do not consistently yield identical responses. Although this randomness poses challenges in terms of reproducibility and control, it offers opportunities for prompt-based modulation, whereby carefully designed prompts can elicit diverse and contextually appropriate outputs [23]. In this respect, the ability to generate a wide range of physician-patient dialogues represents a particularly compelling and pedagogically valuable feature.

In this study, to facilitate the acquisition of fundamental structures in history taking, the AI was instructed to assume the role of a cooperative patient who provided clear and responsive answers. However, the simulated patient did not need to be restricted to cooperative profiles. It is also feasible to generate dialogues featuring patients with diverse communicative behaviors, such as those who are angry, uncommunicative, or unable to articulate clearly because of their underlying health conditions.

On the basis of our findings, we suggest that physician-patient dialogues generated by generative AI can be developed into high-quality educational materials with relatively minimal effort, provided that particular attention is paid to the medical accuracy of the physician's utterances and that supervising physicians revise the content as necessary. Given the capability of generative AI to rapidly produce medically relevant content, its application in medical education and clinical practice is expected to expand further in the coming years [24,25]. As demonstrated in previous studies, when generative AI is used in medical education [26], the ability of instructors to review and modify the generated content enables its use as a supplementary instructional resource. Considering the capacity of the model to generate dialogues tailored to a wide range of clinical scenarios, this approach, when used with appropriate oversight, could offer a highly flexible and scalable educational tool.

Limitations

This study had 3 primary limitations. First, the version of the generative AI used in this study was gpt-4o-2024-11-20, and the evaluation was based on the outputs generated on November 13, 2024. As the performance of generative AI models may evolve with future updates, periodic reevaluation is necessary. Moreover, as multiple LLMs are available, the quality and characteristics of dialogues generated by other models may differ from those evaluated in this study.

Second, there is currently no standardized method for prompt construction when interacting with generative AI, and the content of the input prompt can significantly influence the quality and nature of the output. In this study, the extent to which variations in the input conditions affect the generated dialogues was not examined systematically. Further investigations are warranted to clarify the impact of prompt design on output quality.

Third, this study was conducted in Japanese, and all prompts and generated dialogues were also in Japanese. Although prior research suggests a strong potential for adaptation to other languages, further validation is warranted.

Conclusions

In this study, physician-patient dialogues were generated using an LLM, that is, generative AI. These findings indicate that, although physician supervision remains essential, such dialogues can be easily developed into materials suitable for use in medical education. In particular, dialogues that incorporate the patient's perspective and the interactive elements of physician-patient communication could provide practical learning experiences that are difficult to achieve with conventional educational resources, suggesting a wide range of possible applications in medical education.

Moreover, the use of generative AI enables the efficient and large-scale creation of diverse educational content. This represents a significant advantage in terms of reducing the substantial time and effort traditionally required to develop medical instructional materials. By using this approach, medical students are expected to have more opportunities to learn in contexts that closely resemble actual clinical practice, thereby facilitating the acquisition of more practical medical interview skills.

Acknowledgments

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this paper.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

YY, DY, and TU designed and coordinated the study. YY and DY carried out data analysis and interpretation. YY, DY, and TU drafted the manuscript. SI, RY, YO, and TU revised the manuscript for important intellectual content. All authors read and approved the final manuscript and take responsibility for all aspects of the work, ensuring that any questions regarding accuracy or integrity are appropriately investigated and resolved. This work was supported by JSPS KAKENHI (grant JP25K20499).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Dialogue with a perfect average score (case 21: migraine).
[DOCX File , 18 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Dialogue with a lower average score of 4 (case 27: acute eosinophilic pneumonia).
[DOCX File , 17 KB-Multimedia Appendix 2]

References

1. Mu Y, He D. The potential applications and challenges of ChatGPT in the medical field. Int J Gen Med. Mar 2024;Volume 17:817-826. [doi: 10.2147/ijgm.s456659]
2. Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B. Roles, users, benefits, and limitations of chatbots in health care: rapid review. J Med Internet Res. Jul 23, 2024;26:e56930. [FREE Full text] [doi: 10.2196/56930] [Medline: 39042446]
3. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan: evaluation study. JMIR Form Res. Oct 13, 2023;7:e48023. [FREE Full text] [doi: 10.2196/48023] [Medline: 37831496]
4. Fukuzawa F, Yanagita Y, Yokokawa D, Uchida S, Yamashita S, Li Y, et al. Importance of patient history in artificial intelligence-assisted medical diagnosis: comparison study. JMIR Med Educ. Apr 08, 2024;10:e52674. [FREE Full text] [doi: 10.2196/52674] [Medline: 38602313]
5. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. Jul 03, 2023;330(1):78-80. [FREE Full text] [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
6. Coşkun Ö, Kıyak YS, Budakoğlu I. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: a randomized controlled experiment. Med Teach. Feb 2025;47(2):268-274. [doi: 10.1080/0142159X.2024.2327477] [Medline: 38478902]

7. Tremblay D, Turcotte A, Touati N, Poder TG, Kilpatrick K, Bilodeau K, et al. Development and use of research vignettes to collect qualitative data from healthcare professionals: a scoping review. BMJ Open. Jan 31, 2022;12(1):e057095. [FREE Full text] [doi: 10.1136/bmjopen-2021-057095] [Medline: 35105654]
8. Wofford JL, Singh S. Exploring the educational value of clinical vignettes from the Society of General Internal Medicine national meeting in the internal medicine clerkship: a pilot study. J Gen Intern Med. Nov 2006;21(11):1195-1197. [FREE Full text] [doi: 10.1111/j.1525-1497.2006.00596.x] [Medline: 17026730]
9. Trullàs JC, Blay C, Sarri E, Pujol R. Effectiveness of problem-based learning methodology in undergraduate medical education: a scoping review. BMC Med Educ. Feb 17, 2022;22(1):104. [FREE Full text] [doi: 10.1186/s12909-022-03154-8] [Medline: 35177063]
10. Evans SC, Roberts MC, Keeley JW, Blossom JB, Amaro CM, Garcia AM, et al. Vignette methodologies for studying clinicians' decision-making: validity, utility, and application in ICD-11 field studies. Int J Clin Health Psychol. 2015;15(2):160-170. [FREE Full text] [doi: 10.1016/j.ijchp.2014.12.001] [Medline: 30487833]
11. Yanagita Y, Yokokawa D, Fukuzawa F, Uchida S, Uehara T, Ikusaka M. Expert assessment of ChatGPT's ability to generate illness scripts: an evaluative study. BMC Med Educ. May 15, 2024;24(1):536. [FREE Full text] [doi: 10.1186/s12909-024-05534-8] [Medline: 38750546]
12. Yanagita Y, Yokokawa D, Uchida S, Li Y, Uehara T, Ikusaka M. Can AI-generated clinical vignettes in Japanese be used medically and linguistically? J Gen Intern Med. Dec 2024;39(16):3282-3289. [doi: 10.1007/s11606-024-09031-y] [Medline: 39313665]
13. Questions and answers for the 118th National Medical Examination. Ministry of Health, Labour, and Welfare Japan. 2024. URL: https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp240424-01.html [accessed 2025-10-27]
14. Model OpenAI API. OpenAI Platform. URL: https://platform.openai.com/docs/models#gpt-4o [accessed 2024-12-05]
15. Silverman J, Kurtz S, Draper J. Skills for Communicating with Patients. Boca Raton, FL. CRC Press; 2013.
16. Martinsen SS, Espeland T, Berg EA, Samstad E, Lillebo B, Slørdahl TS. Examining the educational impact of the mini-CEX: a randomised controlled study. BMC Med Educ. Apr 21, 2021;21(1):228. [FREE Full text] [doi: 10.1186/s12909-021-02670-3] [Medline: 33882913]
17. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res. Oct 04, 2023;25:e50638. [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
18. Liu X, Wang J, Sun J, Yuan X, Dong G, Di P, et al. Prompting frameworks for large language models: a survey. ArXiv. Preprint posted online on November 21, 2023. [FREE Full text] [doi: 10.48550/arXiv.2311.12785]
19. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. Aug 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
20. Cook DA, Sherbino J, Durning SJ. Management reasoning: beyond the diagnosis. JAMA. Jun 12, 2018;319(22):2267-2268. [doi: 10.1001/jama.2018.4385] [Medline: 29800012]
21. Collins LG, Schrimmer A, Diamond J, Burke J. Evaluating verbal and non-verbal communication skills, in an ethnogeriatric OSCE. Patient Educ Couns. May 2011;83(2):158-162. [doi: 10.1016/j.pec.2010.05.012] [Medline: 20561763]
22. Little P, White P, Kelly J, Everitt H, Gashi S, Bikker A, et al. Verbal and non-verbal behaviour and patient perception of communication in primary care: an observational study. Br J Gen Pract. May 25, 2015;65(635):e357-e365. [doi: 10.3399/bjgp15x685249]
23. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art. May 18, 2023;6(1):9. [FREE Full text] [doi: 10.1186/s42492-023-00136-5] [Medline: 37198498]
24. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ. Jun 06, 2023;9:e48163. [FREE Full text] [doi: 10.2196/48163] [Medline: 37279048]
25. Li Q, Qin Y. AI in medical education: medical student perception, curriculum recommendations and design suggestions. BMC Med Educ. Nov 09, 2023;23(1):852. [doi: 10.1186/s12909-023-04700-8] [Medline: 37946176]
26. Han Z, Battaglia F, Udaiyar A, Fooks A, Terlecky SR. An explorative assessment of ChatGPT as an aid in medical education: use it with caution. Med Teach. May 2024;46(5):657-664. [doi: 10.1080/0142159X.2023.2271159] [Medline: 37862566]

Abbreviations

AI: artificial intelligence
LLM: large language model
Mini-CEX: Mini-Clinical Evaluation Exercise

Edited by A Mavragani; submitted 16.Jul.2025; peer-reviewed by M Al-Agil, MA Roshani, S Mohamed Shaffi, S Banerjee, X Liang; comments to author 08.Sep.2025; revised version received 18.Sep.2025; accepted 15.Oct.2025; published 03.Nov.2025

Please cite as:
Yanagita Y, Yokokawa D, Ihara S, Yoshida R, Okano Y, Uehara T
Quality Assessment of Large Language Model–Generated Medical Dialogue for Clinical Vignettes: Evaluation Study
JMIR Form Res 2025;9:e80752
URL: https://formative.jmir.org/2025/1/e80752
doi: 10.2196/80752
PMID: