Evaluating the Quality of Health Information Generated by Generative AI: The Case of Chronic Disease Management in China

Yibo Meng, Zhiming Liu, Yongtuo Zhang, Bingyi Liu, Jingruo Chen, Ziyuan Qu, Zehao Li, Zhihao Lei, Zhefang Hu

Submitted to: JMIR Medical Informatics
on: August 15, 2025

Table of Contents

Evaluating the Quality of Health Information Generated by Generative AI: The Case of Chronic Disease Management in China

Yibo Meng[1*]; Zhiming Liu[2*]; Yongtuo Zhang[3*]; Bingyi Liu[4*]; Jingruo Chen[5*]; Ziyuan Qu[6*]; Zehao Li[7]; Zhihao Lei[5]; Zhefang Hu[8]

[1]Tsinghua University Tsinghua University Beijing CN
[2] University of Shanghai For Science and Technology Shanghai CN
[3] University of Shanghai For Science and Technology Shanghai CN
[4] University of Michigan Ann Arbor US
[5] Cornell University Ithaca US
[6] RMIT UNIVERSITY Melbourne AU
[7] NYU NYC US
[8] Imperial College London London GB
[*]these authors contributed equally

Corresponding Author:
Zhefang Hu

Imperial College London
South Kensington Campus, London, SW7 2AZ, United Kingdom
London
GB

Abstract

Background: In recent years, the rapid development of generative artificial intelligence (AI) in China has led to a growing number of AI platforms being applied in healthcare contexts, particularly in assisting communication with patients managing chronic conditions. Tools such as DeepSeek, Kimi, ChatGPT, and Wenxin Yiyan have demonstrated significant potential in supporting patient education, disease explanation, and decision-making. However, the quality of their generated content remains uneven, and inconsistencies in medical accuracy, clarity, and relevance may pose risks to patient health. A systematic evaluation of these tools is urgently needed to inform safe and effective use.

Objective: This study aims to assess the quality of health information generated by four commonly used generative AI tools in China: DeepSeek, Kimi, ChatGPT, and Wenxin Yiyan, in the context of chronic disease communication. The focus is on evaluating their ability to provide accurate, clear, and empathetic responses across a range of content types, including explanations of medical terminology, disease conditions, etiologies, treatment options, and medical costs.

Methods: We conducted a cross-sectional study consisting of two parts. First, we evaluated each AI system's performance on multiple-choice knowledge questions derived from validated instruments for ten chronic diseases. Each item was manually input into the AI systems, and responses were scored for accuracy. As a benchmark, domain-specific physicians completed the same questionnaires. Second, we conducted semi-structured interviews with 50 patients across the ten disease categories to collect a total of 108 real-world, patient-centered questions. These were submitted to the AI systems, and the resulting responses were independently evaluated by ten physicians based on three criteria: medical accuracy, linguistic clarity, and emotional empathy. Physicians also participated in follow-up interviews to compare overall strengths and limitations across platforms.

Results: ChatGPT achieved the highest accuracy in the objective knowledge assessment, correctly answering 216 of 230 questions. Wenxin Yiyan followed with 203 correct responses, Kimi with 201, and DeepSeek with 196. All AI systems scored below the physician baseline (228/230). In the subjective evaluation, ChatGPT again received the most favorable ratings, particularly for completeness and structure. Kimi and Wenxin Yiyan showed mixed results, often performing well on symptom descriptions but less reliably on treatment plans or rare conditions. DeepSeek produced the most inconsistent and verbose responses. Across categories, common weaknesses included a lack of personalized guidance, limited empathetic tone, and outdated or imprecise cost estimates.

Conclusions: Generative AI tools show promise in supporting health communication for chronic disease management in China, but their outputs remain inconsistent in accuracy, clarity, and emotional appropriateness. While ChatGPT and Wenxin Yiyan were generally more reliable, no system consistently matched physician-level performance, especially in nuanced or emotionally sensitive domains. This study highlights the importance of continued monitoring, culturally relevant evaluation, and iterative improvement to ensure that generative AI can serve as a trustworthy complement to professional healthcare services.

(JMIR Preprints 15/08/2025:82469)
DOI: https://doi.org/10.2196/preprints.82469

Preprint Settings

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

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

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

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Evaluating the Quality of Health Information Generated by Generative AI: The Case of Chronic Disease Management in China

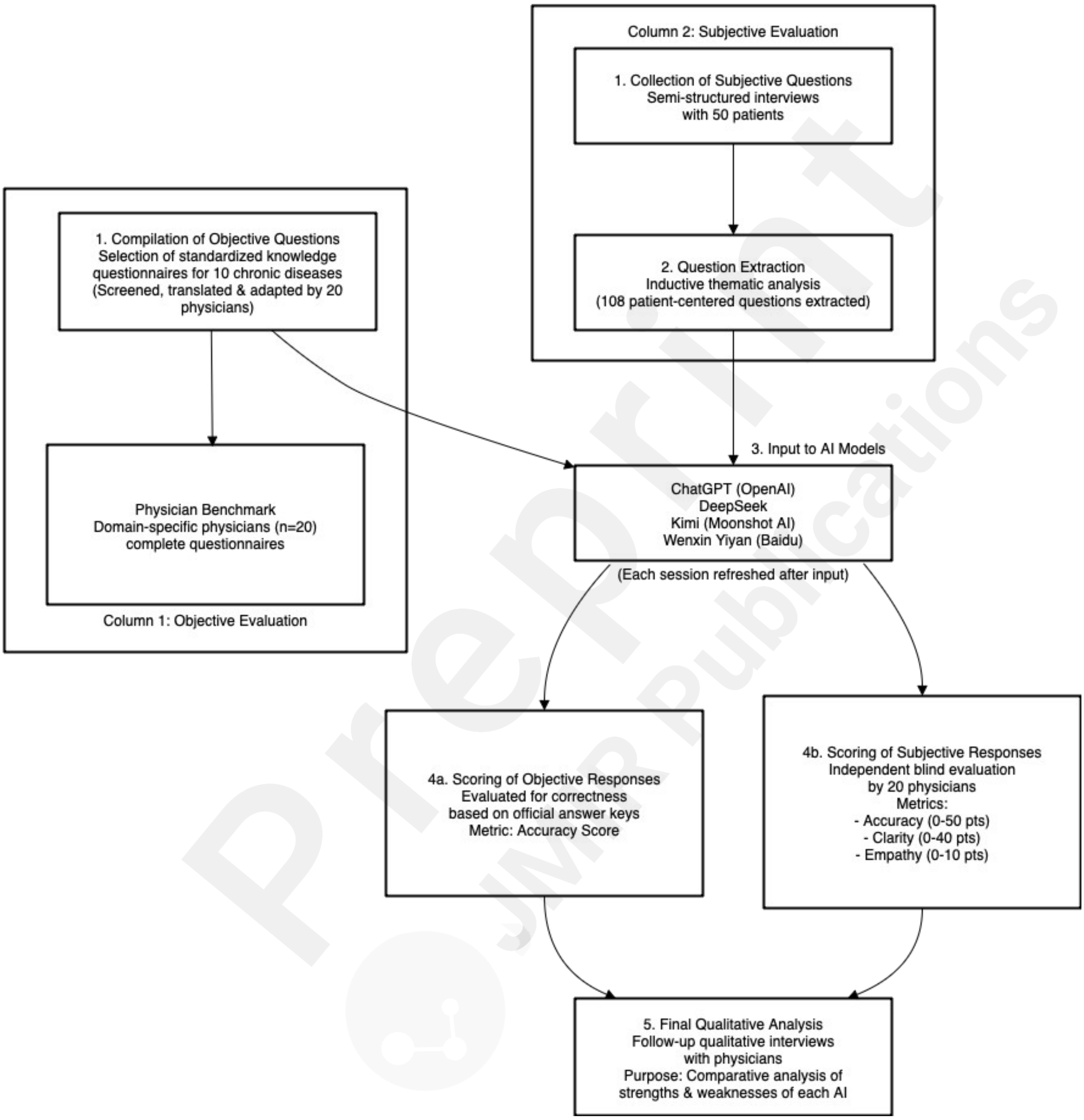

Column 2: Subjective Evaluation
1. Collection of Subjective Questions
Semi-structured interviews
with 50 patients
2. Question Extraction
Inductive thematic analysis
(108 patient-centered questions extracted)
1. Compilation of Objective Questions
Selection of standardized knowledge
questionnaires for 10 chronic diseases
(Screened, translated & adapted by 20
physicians)
Physician Benchmark
Domain-specific physicians (n=20)
complete questionnaires
Column 1: Objective Evaluation
3. Input to AI Models
ChatGPT (OpenAI)
DeepSeek
Kimi (Moonshot AI)
Wenxin Yiyan (Baidu)
(Each session refreshed after input)
4a. Scoring of Objective Responses
Evaluated for correctness
based on official answer keys
Metric: Accuracy Score
4b. Scoring of Subjective Responses
Independent blind evaluation
by 20 physicians
Metrics:
- Accuracy (0-50 pts)
- Clarity (0-40 pts)
- Empathy (0-10 pts)
5. Final Qualitative Analysis
Follow-up qualitative interviews
with physicians
Purpose: Comparative analysis of
strengths & weaknesses of each AI

Figure 1: Visualization of research methods

Abstract

Background: In recent years, the rapid development of generative artificial intelligence (AI) in China has led to a growing number of AI platforms being applied in healthcare contexts, particularly in assisting communication with patients managing chronic conditions. Tools such as DeepSeek, Kimi, ChatGPT, and Wenxin Yiyan have demonstrated significant potential in supporting patient education, disease explanation, and decision-making. However, the quality of their generated content remains uneven, and inconsistencies in medical accuracy, clarity, and relevance may pose risks to patient health. A systematic evaluation of these tools is urgently needed to inform safe and effective use.

Objective: This study aims to assess the quality of health information generated by four commonly used generative AI tools in China: DeepSeek, Kimi, ChatGPT, and Wenxin Yiyan, in the context of chronic disease communication. The focus is on evaluating their ability to provide accurate, clear, and empathetic responses across a range of content types, including explanations of medical terminology, disease conditions, etiologies, treatment options, and medical costs.

Methods: We conducted a cross-sectional study consisting of two parts. First, we evaluated each AI system's performance on multiple-choice knowledge questions derived from validated instruments for ten chronic diseases. Each item was manually input into the AI systems, and responses were scored for accuracy. As a benchmark, domain-specific physicians completed the same questionnaires. Second, we conducted semi-structured interviews with 50 patients across the ten disease categories to collect a total of 108 real-world, patient-centered questions. These were submitted to the AI systems, and the resulting responses were independently evaluated by ten physicians based on three criteria: medical accuracy, linguistic clarity, and emotional empathy. Physicians also participated in follow-up interviews to compare overall strengths and limitations across platforms.

Results: ChatGPT achieved the highest accuracy in the objective knowledge assessment, correctly answering 216 of 230 questions. Wenxin Yiyan followed with 203 correct responses, Kimi with 201, and DeepSeek with 196. All AI systems scored below the physician baseline (228/230). In the subjective evaluation, ChatGPT again received the most favorable ratings, particularly for completeness and structure. Kimi and Wenxin Yiyan showed mixed results, often performing well on symptom descriptions but less reliably on treatment plans or rare conditions. DeepSeek produced the most inconsistent and verbose responses. Across categories, common weaknesses included a lack of personalized guidance, limited empathetic tone, and outdated or imprecise cost estimates.

Conclusions: Generative AI tools show promise in supporting health communication for chronic disease management in China, but their outputs remain inconsistent in accuracy, clarity, and emotional appropriateness. While ChatGPT and Wenxin Yiyan were generally more reliable, no system consistently matched physician-level performance, especially in nuanced or emotionally sensitive domains. This study highlights the importance of continued monitoring, culturally relevant evaluation, and iterative improvement to ensure that generative AI can serve as a trustworthy complement to professional healthcare services.

Keywords: Generative artificial intelligence□chronic disease□patient communication□health information□medical AI

Introduction

In recent years, the incidence of multiple chronic diseases in China, including hypertension, hyperlipidemia, and type 2 diabetes, has been steadily increasing [1, 2]. These conditions now represent one of the most pressing public health challenges facing the country. Chronic diseases are typically characterized by long durations, complex clinical profiles, and the need for sustained self-management [3]. Effective health communication is critical in this context: it supports patient adherence to treatment, enhances quality of life, and improves long-term clinical outcomes [4,5].

However, as a developing country, China continues to face significant limitations in both the total supply and per capita availability of healthcare resources [1]. This resource strain is especially pronounced in rural and underdeveloped areas, where physicians often lack the capacity to provide comprehensive information during each patient interaction [6, 7]. In such settings, patients may struggle to obtain timely, accurate, and comprehensible health information to support their decision-making.

Against this backdrop, generative artificial intelligence (AI) has emerged as a promising new tool. Due to their efficiency, accessibility, and low cost, generative AI platforms have become increasingly integrated into everyday health information-seeking behaviors, particularly among chronic disease patients. In China, models such as DeepSeek, Kimi, ChatGPT, and Wenxin Yiyan have seen rapid adoption in healthcare [8]. These tools provide rapid textual responses, helping patients understand medical concepts, treatment procedures, and precautionary measures. In doing so, AI may serve as an informal extension of clinical communication, partially alleviating the burden on the formal healthcare system [9, 10].

Yet despite their widespread use, the quality of information generated by these AI systems remains inconsistent and has not been systematically assessed. Poorly formulated or inaccurate responses may mislead patients and interfere with treatment decisions. While prior studies have highlighted the potential of generative AI to improve health education and patient engagement [11,12], they have also documented important limitations, including factual inaccuracies, lack of contextual awareness, and the absence of emotional sensitivity in responses [13,14].

Moreover, most existing research focuses on widely used English-language AI tools such as ChatGPT, with limited attention to Chinese-language platforms that are more commonly used by patients in China, such as DeepSeek, Kimi, and Wenxin Yiyan. These tools differ in their training data, architectural design, and target user base, which may significantly affect the quality and cultural relevance of their outputs. In addition, previous studies often rely on narrowly defined quantitative tests to assess AI accuracy. While such evaluations provide useful benchmarks, they do not reflect the range of real-world questions that patients are likely to pose: questions that often demand not only factual correctness but also interpretive explanation, reassurance, or personal relevance.

To address these gaps, this study systematically evaluates the quality of responses produced by four generative AI platforms across ten chronic diseases common in China. We focus on five key types of content relevant to patient-AI communication: medical terminology, disease conditions, causes, treatment options, and medical costs. Our evaluation combines objective knowledge testing using validated medical questionnaires and subjective assessment of AI responses to real patient questions collected through interviews. By incorporating expert evaluation from physicians, we offer a comprehensive analysis of the strengths and weaknesses of these platforms in the context of chronic disease management.

Method

Overview

This study adopts a cross-sectional observational design with two components: (1) evaluating the accuracy of generative AI responses to multiple-choice knowledge questions across ten prevalent chronic diseases in China; and (2) assessing the quality of AI-generated responses to open-ended, subjective questions frequently posed by patients with chronic conditions.

Procedure

We selected ten common chronic diseases in China: hypertension, hyperlipidemia, diabetes, heart disease, chronic kidney disease, idiopathic pulmonary fibrosis, depression, obesity, stroke, and gout, as the primary conditions for investigation. For each disease, we identified a corresponding validated knowledge assessment instrument widely used in clinical or public health contexts (see Table 1). To ensure compatibility with AI input, we recruited ten physicians (one per disease category) to help screen, translate, and adapt the questionnaires. Duplicate items were removed, and questions involving images were excluded to avoid errors related to the AI models' lack of visual interpretation capabilities.

Table 1. Chronic Diseases and Knowledge Assessment Instruments

Disease	Questionnaire	Source	Item Type	Number of Items	MCQ
Hypertension	Hypertension Knowledge Test (HKT)	AHA, WHO	Multiple choice	20	Yes
Hyperlipidemia	Lipid Knowledge Questionnaire (LKQ)	NCEP	Multiple choice	20	Yes
Diabetes	Diabetes Knowledge Test (DKT)	University of Michigan Diabetes Research Center	Multiple choice	24	Yes
Chronic Kidney Disease	CKD Knowledge Questionnaire (CKD-KQ)	National Kidney Foundation (NKF)	Multiple choice	24	Yes
Heart Disease	Heart Disease Knowledge Questionnaire (HDKQ)	AHA	Multiple choice	22	Yes

Pulmonary Fibrosis	IPF Knowledge Questionnaire (IPF-KQ)	ATS / ERS	Multiple choice	26	Yes
Obesity	Obesity Knowledge Scale (OKS)	World Obesity Federation	Multiple choice	16	Yes
Stroke	Stroke Knowledge Questionnaire (SKQ)	AHA / ASA	Multiple choice	24	Yes
Gout	Gout Knowledge Questionnaire (GKQ)	EULAR	Multiple choice	30	Yes
Depression	Depression Literacy Questionnaire (D-Lit)	Beyond Blue (Australia)	Multiple choice	25	Yes

Each question was manually entered into the four widely used generative AI platforms in China: ChatGPT (OpenAI), DeepSeek, Kimi (Moonshot AI), and Wenxin Yiyan (Baidu). Responses were recorded and evaluated as correct or incorrect based on the official answer keys accompanying each knowledge questionnaire. To minimize contextual contamination and ensure independent outputs, each AI session was refreshed after every input, with memory settings disabled where applicable. For baseline comparison, each physician completed the knowledge questionnaire corresponding to their area of clinical specialization, and their response accuracy was recorded using the same scoring protocol.

In the second part of the study, we investigated the quality of AI-generated responses to subjective, patient-driven queries. We conducted semi-structured interviews with 50 individuals living with chronic conditions, recruiting five participants for each disease. Participants were recruited through a combination of online platforms (e.g., Rednote, WeChat, Bilibili) and offline outreach. Interviews were conducted via Tencent Meeting or Zoom, with informed consent obtained prior to participation. Audio recordings were transcribed and anonymized.

From these transcripts, we identified 108 frequently asked patient-centered questions through inductive thematic analysis. A team of two researchers independently coded the transcripts, iteratively generating candidate question pools. Discrepancies in selection were resolved through discussion until consensus was reached. The final set of questions was manually input into each AI platform, and the resulting responses were independently reviewed by the ten physicians, each assessing responses related to their area of expertise. All participating physicians were trained in scoring consistency using 10 standardized example questions before the formal evaluation, and the formal evaluation began only after reaching Cohen's Kappa >0.8.

AI-generated responses were evaluated on three dimensions: 1. Accuracy (scored 0–50): Evaluated based on alignment with current clinical guidelines and medical knowledge. Physicians rated each response on a five-level rubric ranging from complete accuracy and high clinical applicability (40–50) to significant factual inaccuracies or misleading content (0–10). 2. Clarity (scored 0–40): Assessed by the readability and accessibility of the response, including the avoidance of complex medical jargon and clarity of structure. Responses were rated across three tiers, from fluent and easy-to-understand (30–40) to difficult to follow or confusing (0–10). 3. Empathy (scored 0–10): Assessed whether the AI response acknowledged the patient's emotional or psychological state and used a kind, supportive, and reassuring tone.

To reduce potential bias, physicians were blinded to the source AI system of each response. Finally, all ten physicians participated in follow-up qualitative interviews to discuss the comparative strengths and weaknesses of the AI responses across conditions and content types. Interview data were analyzed using reflexive thematic analysis to extract recurring patterns in physician evaluations and identify key areas for improvement in AI-generated health communication.

Results

Quantitative Result

Across the ten chronic disease categories, ChatGPT demonstrated the highest overall accuracy in responding to multiple-choice knowledge questions, correctly answering 216 out of 230 items. Wenxin Yiyan followed with 203 correct responses, Kimi with 201, and DeepSeek with 196. In comparison, the baseline established by physicians yielded 228 correct answers out of 230. Although all AI models performed reasonably well, none matched the accuracy of human medical professionals. A breakdown of model performance by disease category is provided in Table 2.

Table 2. Accuracy of AI Responses to Objective Knowledge Questions by Disease

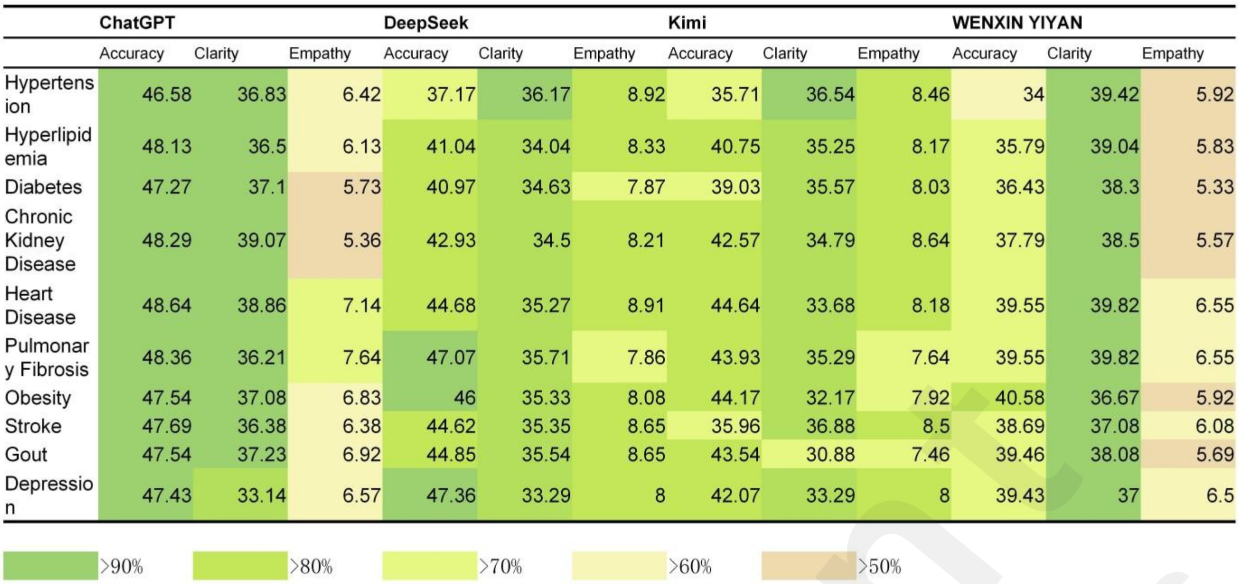
Disease	Total Questions	ChatGPT	DeepSeek	Kimi	Wenxin Yiyan	Physician
Hypertension	20	19	17	17	15	20
Hyperlipidemia	20	19	17	18	16	20
Diabetes	24	23	19	18	23	23
Chronic Kidney Disease	24	24	17	18	22	24
Heart Disease	22	20	20	21	20	22

Pulmonary Fibrosis	26	24	23	24	23	26
Obesity	16	16	13	15	15	16
Stroke	23	20	21	20	20	22
Gout	30	27	25	27	26	30
Depression	25	24	24	23	23	25
Total	230	216	196	201	203	228

As shown in Table 3, we coded and analyzed the results of the qualitative interviews and collected a total of 108 questions, each of which appeared more than or equal to 3 times.

Table 3. Number of Subjective Questions per Disease

Disease	Hypertension	Hyperlipidemia	Diabetes	Chronic Kidney Disease	Heart Disease	Pulmonary Fibrosis	Obesity	Stroke	Gout	Depression	Total
Number of Questions	12	12	15	7	11	6	12	13	13	7	108



	ChatGPT			DeepSeek			Kimi			WENXIN YIYAN		
	Accuracy	Clarity	Empathy	Accuracy	Clarity	Empathy	Accuracy	Clarity	Empathy	Accuracy	Clarity	Empathy
Hypertension	46.58	36.83	6.42	37.17	36.17	8.92	35.71	36.54	8.46	34	39.42	5.92
Hyperlipidemia	48.13	36.5	6.13	41.04	34.04	8.33	40.75	35.25	8.17	35.79	39.04	5.83
Diabetes	47.27	37.1	5.73	40.97	34.63	7.87	39.03	35.57	8.03	36.43	38.3	5.33
Chronic Kidney Disease	48.29	39.07	5.36	42.93	34.5	8.21	42.57	34.79	8.64	37.79	38.5	5.57
Heart Disease	48.64	38.86	7.14	44.68	35.27	8.91	44.64	33.68	8.18	39.55	39.82	6.55
Pulmonary Fibrosis	48.36	36.21	7.64	47.07	35.71	7.86	43.93	35.29	7.64	39.55	39.82	6.55
Obesity	47.54	37.08	6.83	46	35.33	8.08	44.17	32.17	7.92	40.58	36.67	5.92
Stroke	47.69	36.38	6.38	44.62	35.35	8.65	35.96	36.88	8.5	38.69	37.08	6.08
Gout	47.54	37.23	6.92	44.85	35.54	8.65	43.54	30.88	7.46	39.46	38.08	5.69
Depression	47.43	33.14	6.57	47.36	33.29	8	42.07	33.29	8	39.43	37	6.5

>90% >80% >70% >60% >50%

Figure 2: Scores of four AIs on 10 diseases (accuracy, clarity, and empathy)

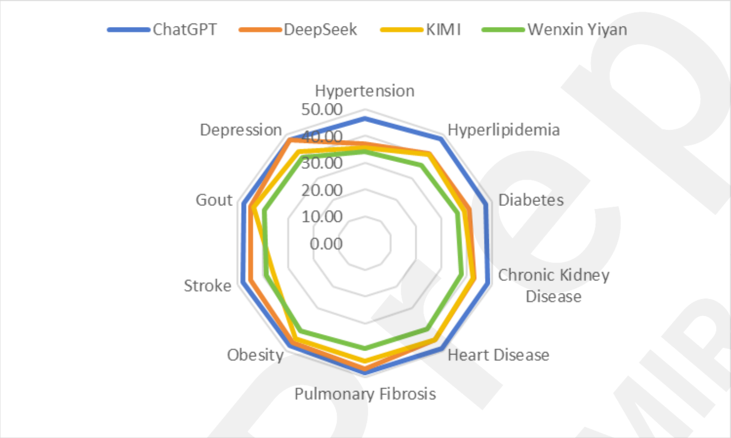


Figure 3: Average accuracy scores of the four AIs

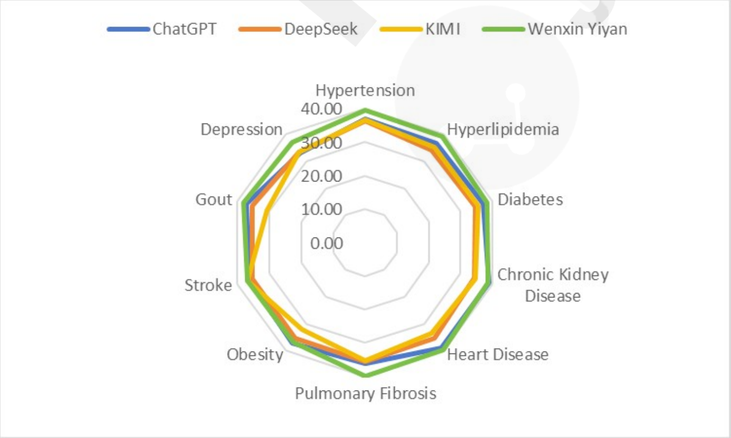


Figure 4: Average clarity scores of the four AIs

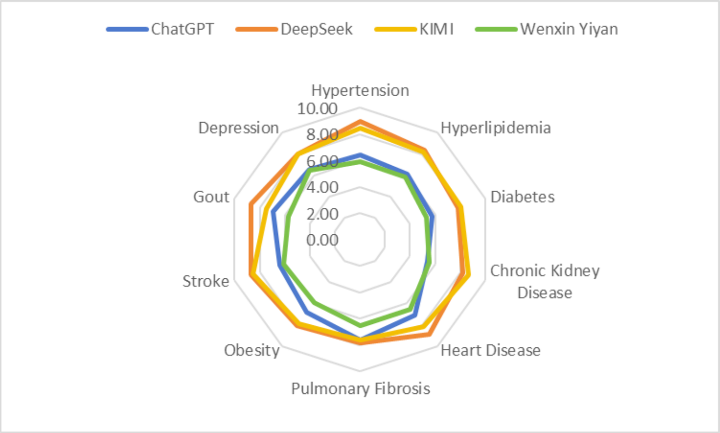


Figure 5: Average empathy scores of the four AIs

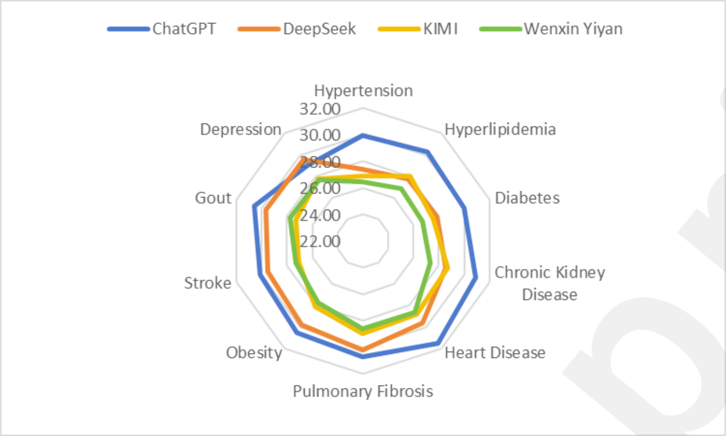


Figure 6: Average scores of four AIs on 10 diseases

As shown in Figure 2, four mainstream AI assistants (ChatGPT, DeepSeek, Kimi, and Wenxin yiyan) exhibit significant differences in their performance in answering subjective medical questions related to ten common diseases (hypertension, hyperlipidemia, diabetes, chronic kidney disease, heart disease, pulmonary fibrosis, obesity, stroke, gout, and depression). We conducted a quantitative evaluation based on three dimensions: accuracy, clarity, and empathy.

Overall, ChatGPT holds a clear advantage in accuracy, maintaining scores above 90% across all disease categories, with particularly strong performance in heart disease (48.64) and chronic kidney disease (48.29). DeepSeek follows closely behind, with accuracy scores mostly concentrated in the 40-47 range. Its performance is very close to ChatGPT in pulmonary fibrosis (47.07) and depression (47.36). The accuracy of Kimi and Wenxinyiyan declined significantly. KIMI's scores for most disease categories were between 80% and 90%, while Wenxin yiyan performed the worst, with most scores below 80%. Figure 3 visualizes the differences in accuracy among the four AI platforms. Overall, ChatGPT performed the best, followed closely by Deepseek and KIMI, with WENXIN YIYAN having the smallest score range.

In terms of information clarity, the differences between platforms were even more pronounced.Wenxin yiyan performed the best, scoring above 90% for all 10 diseases and maintaining the highest score for all disease categories except obesity. ChatGPT performed consistently, with only depression scoring below 90%. DeepSeek and KIMI performed slightly worse than ChatGPT. Kimi's clarity of expression declined significantly on topics such as gout (30.88) and obesity (32.17), revealing weaknesses in specific areas. This is confirmed in the visualization in Figure 4.

As shown in Figures 2 and 5, the four AIs exhibit distinct patterns in empathy. DeepSeek, with a consistently high score of 7.87-8.92, is the most humanistic AI, performing particularly well in categories requiring psychological support, such as heart disease (8.91) and gout (8.65). Kimi also demonstrates strong emotional expression in critical illnesses, such as chronic kidney disease (8.64) and stroke (8.5). In contrast, ChatGPT (5.36-7.14) and Wenxin yiyan (5.33-6.55) have relatively low empathy scores overall, particularly in chronic disease management scenarios like chronic kidney disease (5.36) and diabetes (5.33), where their emotional support capabilities need improvement.
Finally, as shown in Figure 6, ChatGPT's overall score is higher than DeepSeek's, and higher than Kimi's, with Wenxin Yiyan performing the lowest.

A horizontal comparison of specific disease performance reveals that in the highly technical diagnosis of pulmonary fibrosis, the accuracy gap among the four platforms is minimal (47.07-48.36 points), indicating relatively balanced coverage of knowledge across rare diseases. However, in depression, a field involving complex psychological factors, the empathy capabilities of the platforms vary the most (6.5-8 points), highlighting the difficulty of developing AI services for mental health. Longitudinal data reveals that AI services for chronic disease management (such as diabetes and hypertension) are generally more mature than those for mental illnesses, likely due to the more standardized diagnosis and treatment pathways of the former. This result suggests that current technology excels in objective clinical knowledge transfer (accuracy), but still has significant shortcomings in doctor-patient interactions requiring personalized communication (empathy). Future optimization may require differentiated improvements tailored to disease characteristics—strengthening follow-up management features for chronic diseases like diabetes and enhancing emotional interaction design for mental health issues like depression. WENXIN YY's local advantages in certain scenarios suggest that the development of localized knowledge graphs may be a key breakthrough in improving the effectiveness of medical AI.

Qualitative feedback

Medical Terminology Explanation

Physicians noted that both ChatGPT and Wenxin Yiyan generally explained medical terms with clarity, often using contextual examples to help patients grasp abstract or technical concepts. For instance, ChatGPT successfully simplified many terms but failed to fully elaborate on specialized ones. When explaining the Wagner classification system for diabetic foot, it acknowledged the grading system’s existence but did not provide a breakdown of each grade's clinical features (Physician 2).

Kimi was described as concise and direct. While this brevity was appreciated in certain contexts, its responses frequently lacked elaboration or illustrative examples. For example, its explanation of glycated hemoglobin merely identified it as a measure of blood sugar control, omitting normal reference ranges or its clinical implications (Physician 3). Moreover, Kimi occasionally introduced foundational inaccuracies, which contributed to lower ratings on simpler questions (Physicians 1 and 3).

DeepSeek’s explanations were characterized as verbose and structurally disorganized. Rather than providing precise definitions, it often listed a series of loosely related concepts. This approach led several physicians to question its suitability for patients with limited health literacy, especially in rural or underserved populations (Physicians 1, 3, 5, 6, 7).

Disease Condition Explanation

ChatGPT was consistently recognized for its detailed and structured responses when explaining disease progression and symptomatology. For example, in describing hypertension, it accurately delineated the differences among stage I, II, and III hypertension and outlined associated target organ complications. Physicians considered this approach systematic and educational.

In contrast, DeepSeek often failed to maintain coherence in its responses. Its explanations frequently lacked logical progression, resulting in fragmented and confusing narratives (Physicians 2, 4, 5, 6). Kimi and Wenxin Yiyan generally emphasized common symptoms and typical presentations, which sufficed for foundational patient education but offered little depth on severity gradations or atypical manifestations (Physician 8).

Disease Cause Explanation

ChatGPT's answers were generally accurate and up-to-date, often incorporating recent research findings and acknowledging multifactorial causes. However, its use of technical language occasionally diminished patient accessibility (Physicians 1, 2, 3, 4, 5, 6, 8, 9).

Wenxin Yiyan and Kimi provided logically structured and readable explanations of common etiologies. For example, in addressing the causes of diabetes, both models clearly explained the interaction between genetic predisposition, insulin resistance, and lifestyle factors (Physicians 4, 5).

DeepSeek, though more likely to reference rare or less-known causes, did not sufficiently address the more common and clinically significant factors. While its breadth was acknowledged, physicians criticized its lack of depth and prioritization in explanation (Physicians 3, 4, 7).

Treatment Plan Explanation

None of the AI systems was able to offer treatment recommendations that accounted for individual patient characteristics such as age, comorbid conditions, or disease stage. Most responses were generic and lacked contextual sensitivity.

DeepSeek was singled out for particularly poor performance in this category. Physicians noted that it often misunderstood the question's intent and resorted to stringing together unrelated medical terms. For instance, when prompted about treatment options for coronary artery disease, DeepSeek failed to differentiate between medication, interventional therapy, or surgery, providing no meaningful guidance based on disease severity (Physician 3).

Kimi tended to emphasize risks associated with treatment, which helped raise awareness of potential complications but did little to clarify actionable options. Its responses lacked discussion of clinical indications, efficacy, or expected outcomes (Physicians 3, 5, 6).

Wenxin Yiyan offered more coherent summaries of standard treatment protocols but did not adequately cite evidence-based guidelines or personalize recommendations. While its tone and structure were often clear, its content lacked clinical specificity (Physicians 2, 3).

Medical Cost Explanation

ChatGPT and Wenxin Yiyan generally provided accurate and comprehensive information regarding treatment costs. They successfully explained how costs vary by treatment modality, medical institution, and geographic region. For example, when discussing the cost of diabetes care, both systems listed typical price ranges for common medications and insulin regimens, and noted how hospital-grade and insurance coverage affect expenses (Physicians 2, 3, 5, 6, 7).

Kimi's responses went further in distinguishing between institutional and regional differences, helping patients form realistic expectations about financial burdens. However, it lacked recommendations for cost mitigation or affordability strategies (Physicians 3, 4, 10).

DeepSeek's responses in this area were described as unreliable. Physicians highlighted its failure to provide up-to-date pricing information and its vague references to cost structures, rendering it an untrustworthy source for patients seeking financial clarity (Physicians 2, 3, 5, 9).

Discussion

Our evaluation reveals substantial disparities in the quality of information produced by different generative AI platforms when engaging with chronic disease patients. ChatGPT and Wenxin Yiyan generally offered balanced performance across most content types, providing responses that met baseline expectations for accuracy, clarity, and relevance. However, both models demonstrated notable shortcomings in conveying empathy. At times, their replies felt overly mechanical, lacking the warmth and emotional sensitivity that patients often expect, especially when discussing conditions that are both chronic and emotionally taxing. This observation is consistent with prior systematic reviews noting that empathy remains a uniquely human capacity, and while large language models can exhibit elements of cognitive empathy, they often lack affective empathy [15]. Similarly, prior studies have found that "empathetic" responses from LLM-based chatbots are typically the product of linguistic mimicry rather than genuine emotional understanding [16]. Such limitations may hinder rapport-building and patient reassurance, which are critical in long-term disease management.

DeepSeek, although less consistent in factual accuracy, exhibited strengths in providing detailed and personalized explanations, particularly when outlining treatment plans. These advantages, however, were often offset by verbose and loosely structured language that reduced accessibility, especially for individuals with lower health literacy. Consistent with prior findings, DeepSeek-R1 also displayed notable limitations, including the use of overly specialized terminology and a lack of concise expression [17-19]. In particular, when responding to straightforward medication-related questions, its answers tended to be unnecessarily complex and lengthy, making it more suitable as a reference tool for healthcare professionals rather than for general patient use. Kimi, in contrast, stood out for its brevity and its ability to highlight potential treatment risks. However, it frequently struggled to deliver comprehensive, context-rich information, especially in cases involving complex disease presentations or rare etiologies. Prior studies have likewise reported inconsistent performance by Kimi in addressing clinical pharmacy questions [17].

Despite the convenience and efficiency that generative AI tools offer, our findings raise critical concerns about the reliability and patient-centeredness of AI-generated medical content. The first issue is inconsistency in content quality [20]. The platforms varied considerably in terms of factual correctness, clarity, and empathetic tone. These inconsistencies pose a risk of misinformation, particularly in scenarios where patients are making treatment decisions based on AI-provided guidance [21-23].

Second, most platforms lacked deeper expressions of empathy. Their responses seldom conveyed emotional attunement or concern, elements that are essential to fostering trust, reducing anxiety, and supporting adherence in chronic disease care [24,25]. This deficit reflects a broader challenge in current AI design: models remain limited in their ability to recognize and appropriately respond to the emotional dimensions of patient communication [15].

Third, the issue of outdated knowledge remains pressing. In the rapidly evolving field of medicine, even slight lags in knowledge base updates can result in the dissemination of outdated or inaccurate recommendations [26, 27]. This issue was especially apparent in platforms that failed to reflect current clinical guidelines or the latest evidence-based practices [17].

Finally, several AI models were prone to factual errors or inappropriate generalizations, sometimes offering vague, off-topic, or misleading responses [28-30]. In a healthcare context, such inaccuracies can have significant consequences, not only confusing patients but also potentially delaying appropriate medical treatment or undermining trust in professional care.

In light of these findings, future development of AI tools for health communication must address not only content accuracy but also emotional intelligence, cultural relevance, and mechanisms for continual knowledge base updating. However, prior work has cautioned that fine-tuning chatbot outputs to prioritize empathy may inadvertently compromise medical accuracy [16, 31]. This underscores the need for balanced approaches that maintain clinical validity while also enhancing warmth, clarity, and personalization. The ability of AI to function as a trustworthy and supportive assistant in chronic disease management will ultimately depend on its capacity to integrate up-to-date, evidence-based guidance with empathetic and accessible communication.

Conclusion

This study provides one of the first systematic evaluations of generative AI tools, specifically ChatGPT, Kimi, DeepSeek, and Wenxin Yiyan, within the context of chronic disease communication in China. Our findings underscore both the promise and limitations of current AI systems in supporting patient understanding and health information access. While platforms like ChatGPT and Wenxin Yiyan performed comparatively well in terms of accuracy and clarity, none of the AI models consistently delivered responses that met professional standards across all dimensions, especially in scenarios requiring empathy, individualized guidance, or up-to-date medical insights.

Our approach combines objective assessments using standardized medical knowledge questionnaires with subjective evaluations grounded in real patient concerns and expert judgment. By incorporating both quantitative and qualitative data, the study offers a more holistic view of AI performance in a high-stakes, real-world application domain. Moreover, by focusing on chronic diseases and widely used Chinese-language AI platforms, this study fills a gap in the literature, which has predominantly focused on Western-language models and formal medical exams.

Nonetheless, several limitations must be acknowledged. Our sample size was constrained, and the analysis did not include all chronic disease types. Additionally, while we included multiple evaluators for subjective assessment, inter-rater reliability was not formally tested. Future research should aim to expand the participant pool, explore a wider range of conditions, and refine evaluation frameworks to capture more subtle differences in AI-human interaction.

In sum, generative AI tools are poised to become important actors in the health information ecosystem. However, their current capabilities remain uneven, and their safe and effective deployment in clinical-adjacent settings requires continuous monitoring, improvement, and a careful balance between automation and human oversight.

References:

1. Wang Y, Mi J, Shan XY, Wang QJ, Ge KY. Is China facing an obesity epidemic and the consequences? The trends in obesity and chronic disease in China. Int J Obes (Lond). 2007;31(1):177-188. PMID:16652128 doi:10.1038/sj.ijo.0803354

2. Wang Z, Wang S, Lin H, Wang C, Gao D. Prevalence of hypertension and related risk factors in older Chinese population: a meta-analysis. Front Public Health. 2024;12:1320295. PMID:38686031 doi:10.3389/fpubh.2024.1320295

3. Fuller J. What are chronic diseases? Synthese. 2018;195(7):3197-3220. doi:10.1007/s11229-017-1368-1

4. Casey LM, Clough BA, Mihuta ME, Green H, Usher W, James DA, et al. Computer-based interactive health communications for people with chronic disease. Smart Homecare Technol TeleHealth. 2014;2:29-38. doi:10.2147/SHTT.S42684

5. Murray E, Burns J, Tai SS, Lai R, Nazareth I. Interactive Health Communication Applications for people with chronic disease. Cochrane Database Syst Rev. 2005;(4):CD004274. PMID:16235356 doi:10.1002/14651858.CD004274.pub4

6. Tian M, Chen Y, Zhao R, Chen L, Chen X, Feng D, Feng Z. Chronic disease knowledge and its determinants among chronically ill adults in rural areas of Shanxi Province in China: a cross-sectional study. BMC Public Health. 2011;11:948. PMID:22192681 doi:10.1186/1471-2458-11-948

7. Gao Q, Liu M, Peng L, Zhang Y, Shi Y, Teuwen DE, Yi H. Patient satisfaction and its health provider-related determinants in primary health facilities in rural China. BMC Health Serv Res. 2022;22(1):946. PMID:35883080 doi:10.1186/s12913-022-08349-9

8. Zhi Z, Zhao J, Li Q, Li Q, Xu M, Zuo Y, et al. Evolving perceptions and attitudes to adopting generative AI in professional settings: multicenter longitudinal qualitative study of

senior Chinese hospital leaders. J Med Internet Res. 2025 Jun 27;27:e75531. PMID:40577803 doi:10.2196/75531

9. Zeng L, Li Q, Zuo Y, Zhang Y, Li Z. Perceptions and attitudes of Chinese oncologists toward endorsing AI-driven chatbots for health information seeking among patients with cancer: phenomenological qualitative study. J Med Internet Res. 2025 Jul 23;27:e71418. PMID:40699917 doi:10.2196/71418

10. Feng G, Weng F, Lu W, Xu L, Zhu W, Tan M, Weng P. Artificial intelligence in chronic disease management for aging populations: a systematic review of machine learning and NLP applications. Int J Gen Med. 2025;18:3105-3115. PMID:40529344 doi:10.2147/IJGM.S516247

11. Almansour M, Alfhaid FM. Generative artificial intelligence and the personalization of health professional education: a narrative review. Medicine (Baltimore). 2024;103(31):e38955. PMID:39093806 doi:10.1097/MD.0000000000038955

12. Conrad EJ, Hall KC. Leveraging generative AI to elevate curriculum design and pedagogy in public health and health promotion. Pedagogy Health Promot. 2024;10(3):178-186. doi:10.1177/23733799241232641

13. Taylor-Drigo CA, Kumar A. Strengths and limitations of using ChatGPT: a preliminary examination of generative AI in medical education. medRxiv. Preprint posted online March 13, 2025. doi:10.1101/2025.03.12.25323842

14. Sai S, Gaur A, Sai R, Chamola V, Guizani M, Rodrigues JJ. Generative AI for transformative healthcare: a comprehensive study of emerging models, applications, case studies, and limitations. IEEE Access. 2024;12:31078-31106. doi:10.1109/ACCESS.2024.3367715

15. Sorin V, Brin D, Barash Y, Konen E, Charney A, Nadkarni G, Klang E. Large language models and empathy: systematic review. J Med Internet Res. 2024 Dec 11;26:e52597. PMID:39661968 doi:10.2196/52597

16. Chen D, Chauhan K, Parsa R, Liu ZA, Liu FF, Mak E, et al. Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer. NPJ Digit Med. 2025;8(1):275. PMID:40360673 doi:10.1038/s41746-025-01671-6

17. Li L, Du P, Huang X, Zhao H, Ni M, Yan M, Wang A. Comparative analysis of generative artificial intelligence systems in solving clinical pharmacy problems: mixed methods study. JMIR Med Inform. 2025 Jul 24;13:e76128. PMID:40705654 doi:10.2196/76128

18. Jido JT, Al-Wizni A, Le Aung S. Readability of AI-generated patient information leaflets on Alzheimer’s, vascular dementia, and delirium. Cureus. 2025 Jun 6;17(6):e85463. PMID:40621318 doi:10.7759/cureus.85463

19. Taloni A, Sangregorio AC, Alessio G, Romeo MA, Coco G, Busin LM, Sollazzo A, Scorcia V, Giannaccare G. Large language models provide discordant information compared to ophthalmology guidelines. Sci Rep. 2025 Jul 1;15(1):20556. PMID:40596239

doi:10.1038/s41598-025-06404-z

20. Whiles BB, Bird VG, Canales BK, DiBianco JM, Terry RS. Caution! AI bot has entered the patient chat: ChatGPT has limitations in providing accurate urologic healthcare advice. Urology. 2023;180:278-284. PMID:37467806 doi:10.1016/j.urology.2023.07.010

21. Taloni A, Sangregorio AC, Alessio G, Romeo MA, Coco G, Busin LM, Sollazzo A, Scorcia V, Giannaccare G. Large language models provide discordant information compared to ophthalmology guidelines. Sci Rep. 2025 Jul 1;15(1):20556. PMID:40596239 doi:10.1038/s41598-025-06404-z

22. Sallam M. ChatGPT utility in healthcare education, research, practice: systematic review on the promising perspectives, valid concerns. Healthcare (Basel). 2023;11(6):887. doi:10.3390/healthcare11060887

23. Haupt CE, Marks M. AI-generated medical advice—GPT and beyond. JAMA. 2023;329(16):1349-1350. PMID:36972070 doi:10.1001/jama.2023.5321

24. Kourakos M, Vlachou ED, Kelesi MN. Empathy in the health professions: an ally in the care of patients with chronic diseases. Int J Health Sci Res. 2018;8(2):233-240.

25. Gertsman S, Ene IC, Palmert S, Liu A, Makkar M, Shao I, et al. Clinical empathy as perceived by patients with chronic illness in Canada: a qualitative focus group study. CMAJ Open. 2023;11(5):E859-E868. PMID:37751921 doi:10.9778/cmajo.20220211

26. Rouzrokh P, Khosravi B, Faghani S, Moassefi M, Shariatnia MM, Rouzrokh P, Erickson B. A current review of generative AI in medicine: core concepts, applications, and current limitations. Curr Rev Musculoskelet Med. 2025;18(7):246-266. PMID:40304941 doi:10.1007/s12178-025-09961-y

27. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. Interact J Med Res. 2025 Jan 28;14(1):e59823. PMID:39874574 doi:10.2196/59823

28. Zada T, Tam N, Barnard F, Van Sittert M, Bhat V, Rambhatla S. Medical misinformation in AI-assisted self-diagnosis: development of a method (EvalPrompt) for analyzing large language models. JMIR Form Res. 2025 Mar 10;9(1):e66207. PMID:40063849 doi:10.2196/66207

29. Chen C, Shu K. Combating misinformation in the age of LLMs: opportunities and challenges. AI Mag. 2024;45(3):354-368. doi:10.1002/aaai.12188

30. Moëll B, Sand Aronsson F. Harm reduction strategies for thoughtful use of large language models in the medical domain: perspectives for patients and clinicians. J Med Internet Res. 2025 Jul 25;27:e75849. PMID:40712151 doi:10.2196/75849

31. Ibrahim L, Hafner FS, Rocher L. Training language models to be warm and empathetic makes them less reliable and more sycophantic. arXiv. Preprint posted July 29, 2025. doi:10.48550/arXiv.2507.21919

Table 1. Physician Evaluation of AI Responses on Hypertension (12 Questions)

Question No.	1	2	3	4	5	6	7	8	9	10	11	12
ChatGPT	95	95	95	95	90	100	80	80	80	85	85	80
DeepSeek	87	85	85	90	91	91	87	85	87	83	83	85
Kimi	70	65	50	75	77	79	91	89	85	87	90	90
Wenxin Yiyan	75	75	75	80	80	85	85	84	84	84	83	90

Table 2. Physician Evaluation of AI Responses on Hyperlipidemia (12 Questions)

Question No.	1	2	3	4	5	6	7	8	9	10	11	12
ChatGPT	90	100	100	100	95	95	90	89	87	90	91	86
DeepSeek	86	85	85	87	87	87	88	90	90	91	89	95
Kimi	78	79	79	80	80	80	85	85	85	88	94	94
Wenxin Yiyan	82	78	78	80	82	85	86	86	88	90	92	94

Table 3. Physician Evaluation of AI Responses on Diabetes (15 Questions)

Qu	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

estion No.															
ChatGPT	88	89	89	96	95	96	95	89	87	88	89	90	91	78	82
DeepSeek	87	85	94	85	87	85	90	91	91	88	87	90	92	93	94
Kimi	79	80	81	82	82	84	84	83	85	85	90	90	92	93	95
Wenxin Yiyan	80	80	78	84	82	83	85	86	90	90	94	94	95	90	90

Table 4. Physician Evaluation of AI Responses on Chronic Kidney Disease (7 Questions)

Question No.	1	2	3	4	5	6	7
ChatGPT	95	96	89	95	87	85	84
DeepSeek	88	85	88	84	90	92	90
Kimi	82	83	83	85	85	90	89
Wenxin Yiyan	80	84	84	84	89	90	89

Table 5. Physician Evaluation of AI Responses on Heart Disease (11 Questions)

Question No.	1	2	3	4	5	6	7	8	9	10	11
ChatGPT	89	90	90	94	95	95	95	85	85	80	82
DeepSeek	85	84	85	86	86	89	91	91	93	90	92
Kimi	78	79	90	83	85	85	89	88	94	95	93
Wenx	83	82	80	84	82	85	87	89	90	93	94

in Yiyan											

Table 6. Physician Evaluation of AI Responses on Pulmonary Fibrosis (7 Questions)

Question No.	1	2	3	4	5	6	7
ChatGPT	90	98	98	97	100	84	83
DeepSeek	89	85	89	90	95	92	97
Kimi	83	84	85	85	87	93	92
Wenxin Yiyan	83	85	85	90	89	92	93

Table 7. Physician Evaluation of AI Responses on Obesity (12 Questions)

Question No.	1	2	3	4	5	6	7	8	9	10	11	12
ChatGPT	97	97	96	100	99	98	89	88	89	85	87	83
DeepSeek	85	84	83	87	88	88	89	96	94	90	90	99
Kimi	80	82	78	79	80	81	82	84	86	90	93	95
Wenxin Yiyan	79	79	78	82	80	84	83	84	86	88	83	92

Table 8. Physician Evaluation of AI Responses on Stroke (13 Questions)

Question No.	1	2	3	4	5	6	7	8	9	10	11	12	13
ChatGPT	87	88	88	88	94	93	93	95	99	95	78	77	82
DeepSeek	85	84	84	86	86	85	93	92	92	89	89	92	93

Kimi	78	77	77	76	75	78	80	83	83	82	87	89	93
Wenxin Yiyan	79	78	77	77	76	79	82	85	83	82	89	89	86

Table 9. Physician Evaluation of AI Responses on Gout (13 Questions)

Question No.	1	2	3	4	5	6	7	8	9	10	11	12	13
ChatGPT	99	97	95	95	94	95	95	89	89	88	86	85	85
DeepSeek	84	85	84	85	87	87	89	89	90	90	93	92	94
Kimi	79	77	78	78	80	80	81	83	84	83	85	88	89
Wenxin Yiyan	78	80	76	76	80	80	82	83	86	87	90	94	90

Table 10. Physician Evaluation of AI Responses on Depression (7 Questions)

Question No.	1	2	3	4	5	6	7
ChatGPT	89	89	90	92	87	83	80
DeepSeek	85	85	86	87	89	92	93
Kimi	85	83	78	78	83	89	92
Wenxin Yiyan	78	83	83	80	87	89	92