Journal of Multidisciplinary Healthcare

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

AI-Driven Large Language Models in Health Consultations for HIV Patients

Chun-Yan Zhao[1,2,*], Chang Song[1,2,*], Tong Yang[3,*], Ai-Chun Huang[1], Hang-Biao Qiang[1], Chun-Ming Gong[1], Jing-Song Chen[4], Qing-Dong Zhu[1]

[1]Department of Tuberculosis, The Fourth People's Hospital of Nanning, Nanning, Guangxi, People's Republic of China; [2]Clinical Medical School, Guangxi Medical University, Nanning, Guangxi, People's Republic of China; [3]Department of Rehabilitation, Hepu County People's Hospital, Beihai, Guangxi, People's Republic of China; [4]Department of Gastroenterology, Hepu County People's Hospital, Beihai, Guangxi, People's Republic of China

*These authors contributed equally to this work

Correspondence: Qing-Dong Zhu, Department of Tuberculosis, The Fourth People's Hospital of Nanning, No. 1 Changgang Two-Li, Xingning District, Nanning, Guangxi, 530023, People's Republic of China, Tel +86 0771-5636973, Email zhuqingdong2003@163.com; Jing-Song Chen, Department of Gastroenterology, Hepu County People's Hospital, No. 95, Dinghai North Road, Hepu County, Beihai, Guangxi, 536100, People's Republic of China, Tel +86 0779-7106010, Email 410155791@qq.com

Purpose: This study endeavors to conduct a comprehensive assessment on the performance of large language models (LLMs) in health consultation for individuals living with HIV, delve into their applicability across a diverse array of dimensions, and provide evidence-based support for clinical deployment.

Patients and Methods: A 23-question multi-dimensional HIV-specific question bank was developed, covering fundamental knowledge, diagnosis, treatment, prognosis, and case analysis. Four advanced LLMs—ChatGPT-4o, Copilot, Gemini, and Claude—were tested using a multi-dimensional evaluation system assessing medical accuracy, comprehensiveness, understandability, reliability, and humanistic care (which encompasses elements such as individual needs attention, emotional support, and ethical considerations). A five-point Likert scale was employed, with three experts independently scoring. Statistical metrics (mean, standard deviation, standard error) were calculated, followed by consistency analysis, difference analysis, and post-hoc testing.

Results: Claude obtained the most outstanding performance with regard to information comprehensiveness (mean score 4.333), understandability (mean score 3.797), and humanistic care (mean score 2.855); Copilot demonstrated proficiency in diagnostic questions (mean score 3.880); Gemini illustrated exceptional performance in case analysis (mean score 4.111). Based on the post-hoc analysis, Claude outperformed other models in thoroughness and humanistic care ($P < 0.05$). Copilot showed better performance than ChatGPT in understandability ($P = 0.045$), while Gemini performed significantly better than ChatGPT in case analysis ($P < 0.001$). It is important to note that performance varied across tasks, and humanistic care remained a consistent weak point across all models.

Conclusion: The superiority of diverse models in specific tasks suggest that LLMs hold extensive application potential in the management of HIV patients. Nevertheless, their efficacy in the realm of humanistic care still needs improvement.

Keywords: artificial intelligence, large language model, HIV, health consultation, performance analysis

Introduction

Human immunodeficiency virus (HIV) infection and the subsequent acquired immunodeficiency syndrome (AIDS) continue to pose substantial challenges in the global public health arena.[1] As relevant studies illustrate, since 1981, approximately 80 million cases of AIDS infection have been documented worldwide, and more than 35 million people have succumbed to the disease.[2] Concurrently, it is projected that over 40 million people globally will need to receive continuous lifelong antiretroviral therapy (ART) to sustain their lives and health in the next few decades.[3] Notwithstanding the fact that ART has been instrumental in significantly prolonging the life expectancy of those infected, HIV patients still face a myriad of intricate and multi-faceted health management needs, encompassing guidance on medication adherence, prevention of opportunistic infections, mental health support, and navigating social

Received: 29 April 2025
Accepted: 6 August 2025
Published: 25 August 2025

discrimination, among other aspects. Nevertheless, ascribable to the constraints imposed of real-world factors such as a scarcity of skilled professionals, disparities in service accessibility, and the sensitivity of consultation scenarios, conventional health consultation models find it challenging to adequately address patients' needs for tailored, prompt, and confidential services.[4,5]

Against this backdrop, the leap-forward development of natural language processing (NLP) technology, at a technological level, has laid a robust foundation for the intelligentization of medical consultations.[6,7] Large language models (LLMs) grounded in deep learning in the field of artificial intelligence have brought revolutionary solutions to break through the efficiency bottleneck of existing health consultation systems. Early medical dialogue robots on the basis of rule-based systems could only handle structured clinical data on account of their limited knowledge-based scale and insufficient reasoning ability, making it difficult to meet the complex and ever-changing medical consultation needs. With the proposal of Transformer architecture and the rise of pre-trained language models, LLMs represented by GPT have demonstrated human-like language understanding and generation capabilities, enabling them to handle a wide spectrum of language interaction tasks more naturally and flexibly. As already demonstrated by the existing research, LLMs can pass the United States Medical Licensing Examination (USMLE) and generate diagnosis and treatment recommendations in line with evidence-based medicine.[8,9]

AI-driven LLMs are gradually being integrated into the health management system for HIV patients, demonstrating significant potential in providing consultation support and assisting in treatment decision-making.[10,11] In the field of health consultation, leveraging their powerful natural language processing capabilities, LLMs can provide patients with accurate answers regarding key information such as HIV transmission routes, medication adherence, side effect management, and pre-exposure prophylaxis (PrEP) and post-exposure prophylaxis (PEP) around the clock. This effectively alleviates the shortage of professional resources and safeguards patients' privacy. Although preliminary studies have attempted to explore the application effectiveness of LLMs in specific HIV scenarios—for example, scholars such as Vito evaluated the potential of ChatGPT in answering questions related to HIV prevention and examined its accuracy, completeness, and inclusiveness.[12] Beegle et al compared the performance of four AI platforms (ChatGPT 3.5/4.0, Google Bard, and HIV.gov Chatbot) in providing information on HIV medications. All platforms emphasized consulting medical professionals and affirmed the safety of the medications. However, there were instances of "AI hallucinations" where treatment drugs were misjudged as preventive drugs.[10] However, the evaluation systems of such studies overly focus on medical accuracy indicators and generally neglect the crucial value of the humanistic care dimension in the consultation process (such as the quantitative analysis of patients' emotional needs and psychological experiences). Meanwhile, there is a lack of specialized tests targeting core HIV scenarios (such as window-period testing consultations and post-infection medication guidance). Moreover, no horizontal performance comparison of mainstream LLMs have been included. These deficiencies make it difficult for existing studies to comprehensively and objectively reflect the actual application potential of LLMs in the field of HIV health consultations.

Building on the above-mentioned research status, this study constructs 23 specific question banks that cover the full-cycle management of HIV. It integrates a multi-dimensional evaluation system encompassing humanistic care, information comprehensiveness, accuracy, safety, and understandability. Moreover, through a horizontal comparison of four mainstream models, namely ChatGPT-4o, Copilot, Gemini, and Claude, in the process of consultation interaction, this study aims to overcome the limitations of existing research in terms of scenario adaptability, quantification of the humanistic dimension, and clinical translation. It will provide evidence-based support for the clinical deployment of LLMs in the management of HIV patients.

Materials and Methods

Question Design and Sources

This study adopted a multifaceted approach to guarantee the exhaustiveness and representativeness of the question bank. First, we conducted in-depth interviews with clinical practitioners, HIV patients, public health experts, and community workers who have extensive experience in HIV health education to ensure a diverse professional background and enhance representativeness. The interviews focused on core issues in HIV health education, collecting frequent questions

encountered in actual medical scenarios, and were conducted based on a structured interview guide. All interviews were carried out either via online video conferencing or face-to-face, with each session lasting 45–60 minutes. The entire process was recorded with the participants' informed consent. The transcribed texts were analyzed using thematic analysis to systematically categorize patterns and identify themes. Additionally, this study extracted and summarized questions from multiple authoritative HIV-related forums and websites, including but not limited to the Frequently Asked Questions page of the Joint United Nations Programme on HIV/AIDS (UNAIDS) (https://www.unaids.org/en/frequently-asked-questions-about-hiv-and-aids), the HIV Health section of Everyday Health (https://www.everydayhealth.com/hs/hiv-health/top-questions-hiv/), and the China AIDS Health Education Database (https://www.chinaaids.cn/zlk/cllb/xcy/xccts/). To comprehensively capture patients' concerns, we used the Scrapy framework in Python to scrape relevant posts from well-known Chinese social media platforms such as "Weibo" and "DXY". Data collection was conducted up to March 2025 to ensure the inclusion of the latest research findings and public discussions. During the data preprocessing stage, we implemented a multi-step data cleaning process: first, duplicate content was removed using deduplication techniques; second, irrelevant noise data was filtered out based on keyword filtering and manual review; finally, the remaining data were standardized, including unifying text formats, standardizing terminology, and annotating data sources to ensure data quality and consistency. Moreover, we strictly adhered to data usage guidelines, collecting only publicly accessible data, and anonymized all personal information to protect user privacy. During the technical implementation, we complied with the terms of use of each platform, reasonably controlled the scraping frequency to avoid additional server burden, and ensured the compliance and ethical integrity of the research. Finally, after multiple rounds of team discussions and expert consultations, and through strict screening, based on the scale of similar studies, 23 questions were finally determined to balance research depth and assessment efficiency, spanning a multitude of dimensions from foundational disease knowledge, diagnostic procedures, treatment option selection, long-term management and prognosis assessment to clinical case analysis (Table 1). Given that the data utilized in this study were publicly accessible and the research did not involve any direct engagement with human participants, the need for formal ethics committee approval was deemed unnecessary.

Model Selection

In this study, four representative and cutting-edge LLMs in the current field of natural language processing were meticulously chosen as test models, which principally encompass Chat GPT-4o, Copilot (previously known as New Bing), Gemini (formerly Bard), and Claude. This study employed the latest stable versions of several LLMs as of April 2025. All model calls were made through the official API interfaces. To ensure the consistency and controllability of the results, we adopted the default parameter configurations. From March to April 2025, the research team embarked on a comprehensive testing and evaluation. To guarantee the rigor and credibility of the evaluation, the researchers systematically fed a curated set of representative and targeted questions and scenarios, which had been compiled in advance, into each model sequentially. For each question of each model, independent simulation experiments were conducted. During the experiments, all operations were strictly carried out in accordance with the established standardized procedures. The researchers meticulously documented all the responses generated for each question in the experiments, covering the specific content, structural form of the responses, and any potential differences. The specific responses are shown in Supplementary Tables 1~4.

Content Evaluation

To comprehensively evaluate the performance of LLMs in HIV health consultations, this study constructs a multi-dimensional evaluation index system covering five dimensions: accuracy, humanistic care, comprehensiveness, reliability, and understandability. Among them, medical accuracy is the core indicator. By comparing the model's suggestions with the latest clinical guidelines and authoritative medical literature, it ensures that the model's suggestions comply with medical standards. Information comprehensiveness requires that the model's responses cover background information and alternative solutions. Reliability emphasizes avoiding risks to patients caused by inappropriate suggestions. Understandability requires that the model's responses be easy to understand. The scoring criteria for the dimension of "humanistic care" specifically include: evaluating whether the system respects users' cultural, age, and other individual

Table 1 Compilation of Questions

Category	Number	Question
Basic	1	What is HIV?
	2	How is HIV transmitted?
	3	What is the difference between HIV and AIDS?
	4	What are the symptoms of HIV infection?
	5	How can HIV infection be prevented?
Diagnosis	6	What is the "window period" of HIV testing?
	7	How long after HIV infection do symptoms appear?
	8	How soon after a high-risk behavior should one get tested?
	9	What are the criteria for diagnosing HIV?
	10	Does a positive HIV test mean a diagnosis of AIDS?
Treatment	11	What are the main drugs for HIV treatment?
	12	When is it necessary to start HIV treatment?
	13	Does ART require lifelong medication? What happens if you stop taking it?
	14	What are the potential chronic complications of long-term ART?
	15	Can HIV-positive individuals have healthy children?
Prognosis	16	What are the long-term monitoring indicators for HIV-positive individuals?
	17	How can the mental health of HIV-positive individuals be managed?
	18	How can HIV-positive individuals prevent opportunistic infections?
	19	How can HIV-positive individuals manage drug resistance in long-term management?
	20	How to scientifically assess HIV status after treatment?
Case	21	A 20-year-old male patient contracted HIV through male-male sexual contact. How should he be treated and educated about health after diagnosis?
	22	A 35-year-old male patient developed a widespread rash with high fever and facial swelling two weeks after starting ART. His ALT level rose to 200 U/L, and eosinophils were 1.5×10^9/L. What is the next step in treatment?
	23	A 24-year-old female patient was diagnosed with HIV but had not received treatment. She has been experiencing fever and dry cough for a week. Her temperature is 39.2°C, respiratory rate is 32 breaths per minute, and her oxygen saturation is 88%. A chest CT scan shows diffuse ground-glass opacities in both lungs. Her CD4 count is 85/μL. What is the next step in diagnosis and treatment?

differences and provides adaptive support (respect for individual differences); assessing its ability to effectively recognize users' emotions and offer corresponding encouragement or stress-relief strategies (effectiveness of emotional support); examining whether users' control over decision-making is fully safeguarded (autonomy protection); ensuring that the system strictly avoids stereotypical labeling behaviors and protects user privacy (dignity maintenance); and evaluating whether a comprehensive informed consent mechanism has been established and special protections are provided for vulnerable groups (ethical considerations). For each of the above criteria met, the dimension receives one point. This scoring mechanism effectively transforms the abstract concept of humanistic care into a quantifiable and operational evaluation tool. In this study, a five-point Likert scale (1 = strongly disagree, 5 = strongly agree) is used for

scoring. In the scoring stage of this study, to ensure the accuracy and consistency of the scores, we conducted systematic training and calibration for the raters. Before the scoring began, we organized a special training session and invited experts in the field to provide detailed training to the raters. The training content covered the interpretation of scoring criteria, demonstrations of scoring methods, and how to avoid scoring biases. Three experienced experts conduct independent evaluations. During the scoring process, anonymity is maintained, and the order of responses is randomly arranged to reduce bias. The final score is determined by calculating the average of the three independent evaluation results to ensure the objectivity and reliability of the evaluation.

Statistical Analysis

In this investigation, statistical analysis was performed using SPSS 21.0 software. For each question, the average scores of the three raters were calculated respectively. A wide spectrum of statistical metrics, including the mean, minimum value, maximum value, standard deviation, and standard error, were employed to provide a comprehensive description of the scoring outcomes for each question. Apart from that, Cronbach's α was computed to access the internal consistency of the scale. Subsequent to the above steps, the intra-class correlation coefficient (ICC) was calculated to evaluate the consistency among raters. ICC single was used to evaluate the reliability of the scores given by a single rater, reflecting the consistency level of a single scoring result; while ICC average was used to evaluate the reliability of the average scores of multiple raters, reflecting the average consistency of the scores given by the rater group. To assess the differences in scores across various response conditions, a one-way ANOVA analysis was employed to compare the scores based on the specific research design. During the analysis, the following hypotheses were set: H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ (the overall means of all groups are equal), H_1: At least one pair of $\mu_i \neq \mu_j$ (at least two group means are not equal). In the post-hoc analysis, if the data met the assumption of homogeneity of variances, the Least Significant Difference (LSD) post-hoc test was used, with the p-values adjusted using the Bonferroni method to strictly control the overall Type I error rate. If the data did not meet the assumption of homogeneity of variances, Tamhane's T2 post-hoc test was employed. For the post-hoc comparisons, the following hypotheses were set for pairwise comparisons: H_0: $\mu_i = \mu_j$ (no difference between Group i and Group j. H_1: $\mu_i \neq \mu_j$(two-sided). $P<0.05$ was considered statistically significant.

Results

Consistency Analysis of Question Scoring

Table 2 outlines the findings from the consistency analysis of the scores given by the three experts. The table structured into two distinct sections: question categories and diverse dimensions. Collectively, for the 23 questions, the ICC single value is 0.790, and the ICC average stands at 0.919 ($P < 0.001$), and the Cronbach's α value is 0.919.

When examining the diverse facts, for the basic, diagnosis, treatment, prognosis, and case aspects, both the single-score and average-score ICC values are higher than 0.750, the P-values are all less than 0.001, and the Cronbach's α coefficients are all higher than 0.9. With respect to dimensions, for accuracy, comprehensiveness, understandability, reliability, and humanistic care, both the single-score and average-score ICC values reach a significant level, with P-values less than 0.001, and the Cronbach's α values are all higher than 0.8. To conclude, the experts exhibit a high degree of consistency in scoring across all question categories, thereby ensuring the reliability of the scoring outcomes.

Scoring in Different Dimensions

The scores of four models across dissimilar dimensions are visually depicted in Figure 1. Meanwhile, Table 3 offers an in-depth examination of the scoring outcomes of four chatbots, categorized by different dimensions. For each dimension, the minimum score, maximum score, average score, standard deviation (SD), standard error (SE), F-value, and P-value of each chatbot are listed. Notably, the P-values of exhaustiveness, understandability, and humanistic care dimensions are unexceptionally less than 0.05. We conducted a post-hoc analysis on thoroughness, understandability, and humanistic care dimensions. Table 4 shows the pairwise comparison results with significant differences in these dimensions. As our

Table 2 Consistency Analysis of the Scores Given by the Three Experts

	Number of Questions	ICC Single	P value	ICC Average	P value	Cronbach's α
All questions	23	0.790	<0.001	0.919	<0.001	0.919
Question categories						
Basic	5	0.790	<0.001	0.919	<0.001	0.918
Diagnosis	5	0.775	<0.001	0.912	<0.001	0.911
Treatment	5	0.756	<0.001	0.903	<0.001	0.906
Prognosis	5	0.782	<0.001	0.915	<0.001	0.917
Case	3	0.847	<0.001	0.943	<0.001	0.943
Different dimensions						
Accuracy	23	0.699	<0.001	0.875	<0.001	0.878
Comprehensiveness	23	0.812	<0.001	0.928	<0.001	0.928
Understandability	23	0.686	<0.001	0.868	<0.001	0.869
Reliability	23	0.697	<0.001	0.873	<0.001	0.874
Humanistic care	23	0.721	<0.001	0.886	<0.001	0.885

research findings demonstrate evidently, the results showed that Claude scored the highest in terms of comprehensiveness and humanistic care ($P < 0.05$). In summary, Claude performed best in comprehensiveness and humanistic care, indicating its strong ability in providing detailed and empathetic responses.

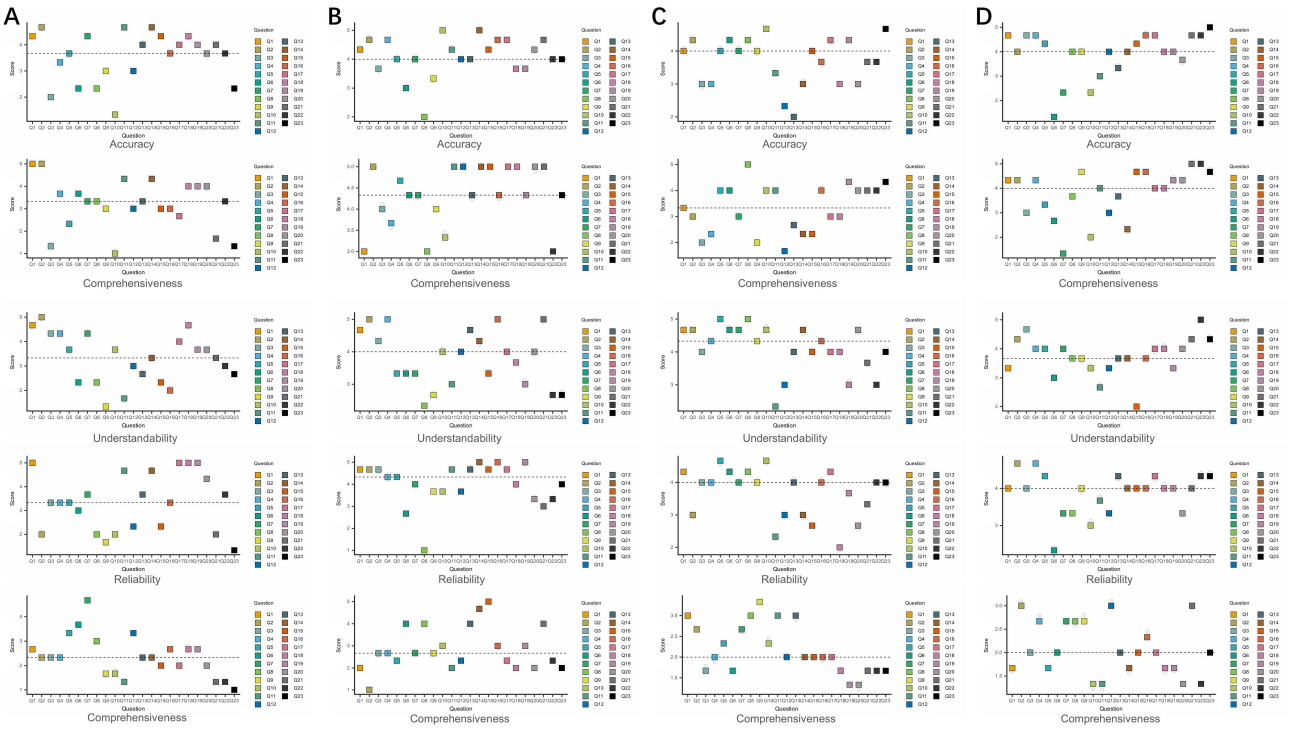

Figure 1 The specific response performances of the four large language models (LLMs) in different dimensions. (**A**) ChatGPT; (**B**) Claude; (**C**) Copilot; (**D**) Gemini.

Table 3 The Scores of the Four Chatbots in Different Dimensions

Different Dimensions	LLM	Minimum	Maximum	Mean	SD	SE	F	P value
Accuracy	ChatGPT	1.330	4.670	3.550	0.942	0.196	1.967	0.125
	Claude	2.000	5.000	4.087	0.684	0.143		
	Copilt	2.000	4.670	3.681	0.728	0.152		
	Gemini	1.330	5.000	3.928	0.917	0.191		
Comprehensiveness*	ChatGPT	1.000	5.000	3.202	1.104	0.230	6.877	<0.001
	Claude	3.000	5.000	4.333	0.711	0.148		
	Copilt	1.670	5.000	3.318	0.918	0.191		
	Gemini	1.330	5.000	3.797	0.994	0.207		
Understandability*	ChatGPT	1.330	5.000	3.304	1.020	0.213	3.801	0.013
	Claude	2.330	5.000	3.797	0.857	0.179		
	Copilt	2.330	5.000	4.117	0.709	0.148		
	Gemini	2.000	5.000	3.739	0.651	0.136		
Reliability	ChatGPT	1.330	5.000	3.333	1.207	0.252	2.619	0.056
	Claude	1.000	5.000	4.030	0.937	0.195		
	Copilt	2.000	4.670	3.667	0.759	0.158		
	Gemini	2.330	4.670	3.883	0.557	0.116		
Humanistic care*	ChatGPT	1.000	4.670	2.377	0.855	0.178	4.652	0.005
	Claude	1.000	5.000	2.855	0.989	0.206		
	Copilt	1.330	3.330	2.174	0.593	0.124		
	Gemini	1.330	3.000	2.073	0.569	0.119		

Notes: * indicates that a statistically significant difference exists among the three large language models according to the one-way ANOVA ($P < 0.05$), necessitating the use of post-hoc analysis.

Table 4 The Models with Significant Differences in the Post-Hoc Analysis of the Three Dimensions of Comprehensiveness, Understandability, and Humanistic Care

Different Dimensions	Difference in Means	SE	P value
Comprehensiveness			
ChatGPT VS Claude	−1.130	0.274	<0.001
Claude VS Copilt	1.014	0.242	<0.001
Understandability			
ChatGPT VS Claude	−0.493	0.242	0.045
ChatGPT VS Copilt	−0.812	0.242	0.001

(*Continued*)

https://doi.org/10.2147/JMDH.S533621

Table 4 (Continued).

Different Dimensions	Difference in Means	SE	P value
Humanistic care			
Claude VS Copilt	0.681	0.241	0.044
Claude VS Gemini	0.782	0.238	0.014

Scoring Across Question Categories

Table 5 provides a comprehensive breakdown of the scoring outcomes of four LLMs across various dimensions, namely basic knowledge, diagnosis, treatment, prognosis, and case analysis). As suggested by our research findings, in terms of basic knowledge, Claude has the highest average score (3.893), but the P-value is 0.554. In the diagnosis aspect, Copilot has the highest average score (3.880), and the P-value is less than 0.001. In the treatment aspect, Claude has the highest average score (4.240), and the P-value is less than 0.001. In the prognosis aspect, Claude has the highest average score (3.947), but the P-value is 0.119, which reveals that the difference is negligible. In the case analysis aspect, Gemini has the highest average score (4.111), and the P-value is less than 0.001. To further explore the specific manifestations of these conspicuous distinctions, we conducted a post-hoc analysis on the three question categories of diagnosis, treatment, and case analysis (Table 6). In summary, Claude excelled in treatment and prognosis, while Gemini performed best in case analysis, highlighting the need for task-specific model selection.

Table 5 The Scores of the Four Chatbots in Question Categories

Question Categories	LLM	Minimum	Maximum	Mean	SD	SE	F	p value
Basic	ChatGPT	1.330	5.000	3.667	1.106	0.221	0.701	0.554
	Claude	1.000	5.000	3.893	1.066	0.213		
	Copilt	1.670	5.000	3.493	0.968	0.194		
	Gemini	1.670	4.670	3.787	0.952	0.190		
Diagnosis*	ChatGPT	1.000	4.670	2.747	1.029	0.206	7.127	<0.001
	Claude	1.000	5.000	3.333	0.866	0.173		
	Copilt	1.670	5.000	3.880	0.932	0.186		
	Gemini	1.330	4.670	2.933	0.923	0.185		
Treatment*	ChatGPT	1.330	4.670	3.227	1.017	0.203	10.38	<0.001
	Claude	2.000	5.000	4.240	0.831	0.166		
	Copilt	1.670	4.670	2.907	0.814	0.163		
	Gemini	1.330	4.670	3.240	0.916	0.183		
Prognosis	ChatGPT	2.000	5.000	3.601	0.927	0.185	2	0.119
	Claude	2.000	5.000	3.947	0.961	0.192		
	Copilt	1.330	4.670	3.266	1.045	0.209		
	Gemini	1.330	4.670	3.600	0.991	0.198		

(*Continued*)

Table 5 (Continued).

Question Categories	LLM	Minimum	Maximum	Mean	SD	SE	F	p value
Case*	ChatGPT	1.000	4.000	2.399	1.041	0.269	7.149	<0.001
	Claude	2.000	5.000	3.600	0.961	0.248		
	Copilt	1.670	4.670	3.406	1.022	0.273		
	Gemini	1.330	5.000	4.111	1.125	0.291		

Notes: * indicates that a statistically significant difference exists among the three large language models according to the one-way ANOVA (P < 0.05), necessitating the use of post-hoc analysis.

Discussion

Nowadays, the application of LLMs has witnessed explosive growth and actualized substantial advancement in the medical field. As relevant studies suggest, cutting-edge models represented by GPT and Claude have performed excellently in a vast array of tasks such as providing breast cancer screening recommendations and explaining diabetes management guidelines.[13–15] However, in the management of infectious diseases such as HIV/AIDS, which are both medically complex and socially sensitive, the exploration of LLM is still in its infancy.

As evidently demonstrated by the aforementioned research findings, these models exhibit notable disparities across various dimensions and question categories. In a prior study evaluating the efficacy of diverse LLMs in urolithiasis health consultations and patient education, Claude similarly distinguished itself, underscoring its exceptional proficiency in medical knowledge dissemination and patient guidance.[16] For instance, when addressing sensitive questions such as "Can HIV-positive individuals have healthy children?", Claude not only furnishes detailed recommendations and strategies for both parents but also specifically adds psychological support suggestions. This capacity to seamlessly integrate medical guidelines with humanistic care epitomizes the unique demands inherent in HIV consultations, setting them apart from conventional medical interactions. From the multi-aspect evaluation results, Copilot demonstrates exceptional proficiency in addressing diagnostic queries, securing an average score of 3.880, which is likely intricately linked to its robust

Table 6 Models with Significant Differences in the Post-Hoc Analysis of the Three Question Categories of Diagnosis, Treatment, and Case Analysis

Question Categories	Difference in Means	SE	P value
Diagnosis			
ChatGPT VS Copilt	−1.133	0.278	0.001
Copilt VS Gemini	0.947	0.262	0.004
Treatment			
ChatGPT VS Claude	−1.013	0.263	0.002
Claude VS Copilt	1.333	0.233	<0.001
Claude VS Gemini	1.000	0.247	0.001
Case			
ChatGPT VS Claude	−1.201	0.366	0.017
ChatGPT VS Gemini	−1.711	0.396	0.001

language comprehension and generation capabilities. To be specific, it not only exhibits the capability to quickly and accurately understand the intricate inquiries posed by patients or doctors, but also formulates precise, well-structured responses by leveraging its extensive medical knowledge repository. Within the clinical case analysis context, Gemini significantly surpasses its counterparts, achieving an average score of 4.111 ($P < 0.001$), with its primary edge manifesting in its reasoning prowess within convoluted scenarios. As suggested by this study, all the assessed models exhibit unfavorable performance in the dimension of humanistic care (Claude has the highest score, but it is only 2.855). This phenomenon may be attributed to two key factors. First, the existing medical corpora predominantly consist of evidence-based guidelines, scientific research papers, and Internet information, lacking real doctor-patient dialogue records and patients' self-narration texts.[17] Second, there is a notable shortage of attention paid to soft skills such as emotional support and humanistic care during the current model training process. In the foreseeable future, LLMs are still unable to fully replace the core role of doctors in doctor-patient communication. The limitations of the current models are particularly prominent in medical scenarios that demand emotional resonance and personalized support. On this basis, we arrived at a pertinent conclusion that the models are inclined to revolve around pure medical explanations rather than providing emotional support. The emotion recognition of LLMs relies on surface-level lexical analysis (such as keywords like "depression" and "anxiety") and fails to capture the implicit emotions in patients' narratives. Across dissimilar dimensions, Copilot performs outstandingly in diagnosis, Claude excels in treatment, while Gemini exhibits exceptional performance in case analysis. Based on the research findings, we recommend that future studies should explore the potential applications of LLMs in clinical settings in a phased manner, while emphasizing the importance of rigorous validation in real-world clinical environments. In primary care settings, the potential of Claude as a consultation assistant warrants further investigation; however, its advantages in comprehensiveness and humanistic care need to be validated within real patient-provider workflows, with an assessment of error risks and legal liabilities. In specialized care settings, the hybrid architecture of Gemini and Copilot may provide technical support for complex case reasoning and real-time guideline updates, but its actual efficacy and safety must be confirmed through multicenter clinical trials. In patient self-service scenarios, the development of an anonymous consultation platform based on the Claude engine may help address privacy concerns, yet its design must fully incorporate patient feedback and ethical review. In summary, the clinical integration of LLMs should follow a step-by-step approach, with rigorous effectiveness evaluations at each stage to ensure safety, efficacy, and ethical compliance.

The application of artificial intelligence in the medical field is reshaping the traditional diagnosis and treatment models and the paradigms of doctor-patient interactions. However, it must be admitted that there are still certain acceptance barriers to AI-based medical applications among both doctors and patients at present. At the patient level, concerns are mainly manifested as psychological resistance to algorithm-based decision-making, fear of robot-assisted diagnosis and treatment, anxiety about the lack of humanistic care, obstacles in human-machine interaction, and doubts about the universality of AI. On the other hand, the concerns of doctors are concentrated on the fear of job replacement, the dilemma of defining medical liability, the pressure of work-flow transformation, and doubts about the reliability of the technology.[18] These acceptance barriers not only reflect the real challenges in the process of integrating technological innovation with medical practice but also highlight the key issues that urgently need to be improved in terms of technological ethics and humanistic care in AI-based medical applications. In the AI era, the core value of doctors lies not only in their ability to apply technology but also in maintaining the irreplaceable humanistic care in medical practice, fulfilling multiple functions such as technical explanation, ethical mediation, and decision-making supervision.[19] Doctors do not need to resist the development of AI or worry about being replaced. Reasonable use of AI can relieve the burden of procedural work, enabling doctors to focus more on creative diagnosis and treatment activities. LLMs can supplement the work of doctors by providing patients and healthcare providers with fast and accurate information.[20]

Although the research findings are encouraging, it is essential to acknowledge the certain limitations in this study. Firstly, the evaluation indicators of this study mainly focus on dimensions such as medical accuracy, information comprehensiveness, understandability, and humanistic care, yet there is a lack of evaluation of the model's application effectiveness in the actual clinical environment. Secondly, the sample size of this study is limited, covering only 23 HIV-specific consultation questions, which may not fully reflect the performance of the LLMs in a broader range of medical scenarios. In addition, although the expert evaluation method adopted in this study has significant advantages in

evaluating technical performance indicators (eg, accuracy, reliability), it has inherent limitations in capturing the subjective experiences of real patients regarding specific interaction dimensions. Expert evaluation can hardly fully replace the patients' personal feelings and evaluations of dimensions such as "humanistic care", "information clarity", and "trust-building". Future research should focus on breaking through in three major directions: First, conduct multi-center pragmatic clinical trials to evaluate the long-term impact of LLMs-based consultations on core HIV management indicators (such as the frequency of viral load testing, ART interruption rate, and the incidence of opportunistic infections) in a real clinical environment. Second, construct a cross-cultural evaluation matrix. By incorporating localized cases from high-HIV-burden regions such as Africa and Southeast Asia, test the response effectiveness of the model under different languages, medical resource levels, and socio-cultural backgrounds. Third, future research will incorporate direct feedback from patients or the general public and systematically integrate the patients' perspectives into the evaluation process. In addition, future research should also focus on collaborating with developers. On one hand, at the technical level, leveraging the developers' expertise to conduct in-depth optimization of the existing model structure, and improving the model's performance in data processing, feature extraction, and result prediction, so as to achieve more efficient and accurate health consultations and treatment decisions for HIV patients. On the other hand, efforts should be made to strengthen the integration of the model with real-world application scenarios. Developers can assist in developing applications suitable for different platforms, such as mobile medical apps and telemedicine systems, enabling the model's outcomes to serve medical workers and patients more conveniently and enhancing its practicality and generalizability. Meanwhile, continuous attention should be paid to the model's manifestation in multilingualism and humanistic care. By integrating the cultural and linguistic characteristics of different regions around the world, better support can be provided for people from diverse backgrounds. Furthermore, the research will be placed in a broader dynamic health context to explore the application potential of the model in other relevant health fields, thus continuously expanding the influence and application scope of the research.

To conclude, the application of LLMs in HIV health consultations has taken a crucial step. Only through sustained technological iterations, rigorous validation of effectiveness, and an inclusive governance framework can this innovative technology truly meet the health needs of people living with HIV globally and provide intelligent solutions for ending the AIDS epidemic.

Conclusion

In conclusion, the integration of LLMs into HIV health consultations marks a pivotal advancement. While demonstrating promise in providing personalized support and addressing complex medical queries, these models face limitations in humanistic care and real-world applicability. Future progress hinges on iterative technological refinement, rigorous clinical validation, and the development of inclusive governance frameworks. By addressing these challenges, LLMs hold the potential to significantly enhance HIV management and contribute to the global effort to combat HIV/AIDS.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This work was supported by the Guangxi Key Research and Development Plan Project (Guike AB25069097) and Guangxi Health Commission Self-funded Research Project (Z-A20231211).

Disclosure

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Wu M, Hong C, Dou Z. Joinpoint regression and age period cohort analysis of global and Chinese HIV incidence trends from 1990 to 2021. *Sci Rep*. 2025;15(1):8153. doi:10.1038/s41598-025-92882-0
2. Bekker LG. The HIV epidemic 40 years on. *Nat Rev Microbiol*. 2023;21(12):767–768. doi:10.1038/s41579-023-00979-y
3. Carter A, Zhang M, Tram KH, et al. Global, regional, and national burden of HIV/AIDS, 1990-2021, and forecasts to 2050, for 204 countries and territories: the global burden of disease study 2021. *Lancet HIV*. 2024;11(12):e807–e822.
4. van Velthoven MH, Tudor Car L, Car J, Atun R. Telephone consultation for improving health of people living with or at risk of HIV: a systematic review. *PLoS One*. 2012;7(5):e36105. doi:10.1371/journal.pone.0036105
5. Vermund SH, Mallalieu EC, Van Lith LM, Struthers HE. Health communication and the HIV continuum of care. *J Acquired Immune Deficiency Syndromes*. 2017;74 Suppl 1(Suppl 1):S1–s4. doi:10.1097/QAI.0000000000001211
6. Berger J, Packard G. Using natural language processing to understand people and culture. *Am Psychol*. 2022;77(4):525–537. doi:10.1037/amp0000882
7. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than english: opportunities and challenges. *J Biomed Semantics*. 2018;9(1):12.
8. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States medical licensing examination. *Med Teach*. 2024;46(3):366–372.
9. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312.
10. Beegle S, Gomez LA, Blackard JT, et al. HIV prevention and treatment information from four artificial intelligence platforms: a thematic analysis. *AIDS Behav*. 2025:1–10.
11. Wei W, Shao J, Lyu RQ, et al. Enhanced language models for predicting and understanding HIV care disengagement: a case study in Tanzania. *Res Square*. 2025:rs–3.
12. De Vito A, Colpani A, Moi G, et al. Assessing ChatGPT's potential in HIV prevention communication: a comprehensive evaluation of accuracy, completeness, and inclusivity. *AIDS Behav*. 2024;28(8):2746–2754. doi:10.1007/s10461-024-04391-2
13. Leung YW, So J, Sidhu A, et al. The extent to which artificial intelligence can help fulfill metastatic breast cancer patient healthcare needs: a mixed-methods study. *Curr Oncol*. 2025;32(3). doi:10.3390/curroncol32030145
14. Liu RJ, Forsythe A, Rege JM, Kaufman P. BIO25-024: real-time clinical trial data library in non-small cell lung (NSCLC), prostate (PC), and breast cancer (BC) to support informed treatment decisions: now a reality with a fine-tuned large language model (LLM). *J Nat Comprehensive Cancer Network*. 2025;23(3.5). doi:10.6004/jnccn.2024.7156
15. Abbasian M, Yang Z, Khatibi E, et al. Knowledge-infused LLM-powered conversational health agent: a case study for diabetes patients. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual International Conference*. 2024;2024:1–4. doi:10.1109/EMBC53108.2024.10781547
16. Song H, Xia Y, Luo Z, et al. Evaluating the performance of different large language models on health consultation and patient education in urolithiasis. *J Med Sys*. 2023;47(1):125.
17. Tan XW, Chen WF, Wang NN, et al. Efficiency of different large language models in China in response to consultations about PCa-related perioperative nursing and health education. *Zhonghua nan ke xue*. 2024;30(2):151–156.
18. Li W, Liu X. Anxiety about artificial intelligence from patient and doctor-physician. *Patient Educ Couns*. 2025;133:108619. doi:10.1016/j.pec.2024.108619
19. Jotterand F, Bosco C. Artificial intelligence in medicine: a sword of damocles? *J Med Syst*. 2021;46(1):9. doi:10.1007/s10916-021-01796-7
20. Cheng K, Li Z, He Y, et al. Potential use of artificial intelligence in infectious disease: take ChatGPT as an example. *Ann Biomed Eng*. 2023;51(6):1130–1135. doi:10.1007/s10439-023-03203-3