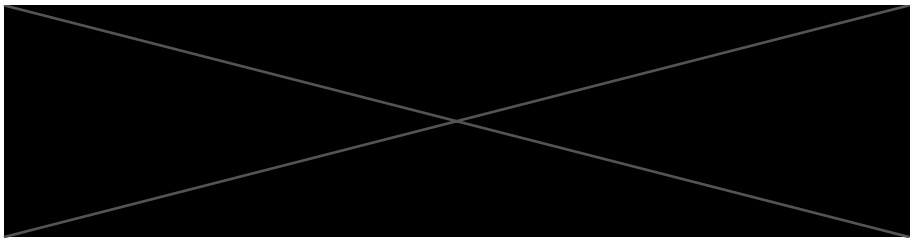
YEAR 2 STUDY
Agentic AI-Powered CliniKiosk: Expanding Healthcare Access with Explainable and Empathetic AI



Abstract

Background: Healthcare systems worldwide face widening gaps in access due to workforce shortages, facility closures, and prolonged wait times, disproportionately affecting rural and marginalized populations. Digital health innovations, particularly AI-powered tools, offer opportunities to expand access; however, existing health chatbots are often constrained by rule-based design, limited multilingual support, minimal personalization, inadequate empathic communication, and insufficient transparency and expressiveness in their responses. In a cross-sectional survey, both healthcare professionals and non–healthcare professionals expressed generally positive views of AI in healthcare while also highlighting significant concerns related to ethics, bias, and cultural or social barriers, with low overall trust in AI's ability to make fair and ethical healthcare decisions.

To address identified trust and ethical concerns, we designed CliniKiosk – a multilingual, agentic, retrieval-augmented digital health kiosk intended to deliver evidence-based, explainable,

and empathetic health recommendations. The objective of this study was to evaluate whether a multilingual, agentic, retrieval-augmented digital CliniKiosk can deliver accurate, empathetic, explainable, and linguistically rich health recommendations.

Methods: CliniKiosk was engineered as a modular platform integrating a React front-end, a Supabase-managed backend with Row Level Security, and a four-component agentic conversational architecture (Medical Router, Planner, Executor, and Manager). Audio input was implemented through Whisper-1, and multilingual output was supported via i18next. A Retrieval-Augmented Generation (RAG) pipeline connected to PubMed, CrossRef, ClinicalTrials.gov, AHA, ACP, and NICE guidelines was developed. We generated a clinician-verified dataset of 500 simulated patient cases representing diverse demographics, health-literacy levels, organ systems, conditions, and clinical complexities. CliniKiosk was benchmarked against GPT-4o, Claude 3 Opus, Mixtral 8×7B, and Gemma 3 12B using standardized prompts. Accuracy metrics were compared via χ^2 tests and pairwise two-proportion z-tests; empathy, explainability, and lexical diversity were assessed using ANOVA and Welch t-tests with Holm correction.

Results: Triage accuracy was high across all models (88.6% for CliniKiosk) with only modest differences. Differential-diagnosis accuracy exceeded 92% for all models with no significant differences ($p=0.865$). Final-diagnosis accuracy was lower for CliniKiosk (76.0%) compared with other LLMs. CliniKiosk achieved comparable treatment-recommendation accuracy (84.5%) to Mixtral and Gemma, though GPT-4o and Claude Opus performed significantly better ($p<0.0001$). Notably, CliniKiosk achieved the highest empathy score (3.98/4) and strong

explainability, outperforming most comparator models. Lexical diversity was moderate and appropriate for users with lower health literacy.

Conclusions: Our findings demonstrate that a multilingual, RAG-based, agentic CliniKiosk can deliver empathetic, evidence-based health recommendations and triage decisions with accuracy comparable to leading LLMs. CliniKiosk demonstrated superior empathy and explainability while using more accessible language, aligning human-centered communication with trust and transparency.

Keywords: Clinikiosk; multilingual; healthbot; rural; healthcare; digital; artificial intelligence

Introduction

Healthcare access is a pressing global problem affecting both developed and developing countries. In 2021, according to the World Health Organization (WHO), approximately 4.5 billion people lacked access to essential healthcare services.[1] Economically disadvantaged communities face the greatest challenges in accessing necessary care, resulting in millions of deaths annually.[2,3] Likewise, over the past two decades in the United States (US), nearly 200 rural hospitals have closed, leaving millions of Americans without local emergency services. Rural communities face significantly higher early death rates from major diseases, and healthcare access is forecasted to worsen in such areas–more than 30% of rural hospitals are at risk of closure. [4-10] Throughout the US, it also takes a long time to get a medical appointment (the average wait time to see a primary care physician is more than 30 days and almost 7 weeks for a behavioral health provider). Without new ways to address the growing demands on already

overwhelmed healthcare systems, wait times will get longer and emergency departments will become even more overcrowded.[13] A workforce shortage is also expected to further worsen access to healthcare. WHO estimates a critical shortage of 10 million healthcare professionals by 2030, especially in low- to middle-income countries.[14] According to the American Hospital Association, the US in particular is projected to face a critical shortage of up to 3.2 million healthcare workers by 2026.[15,16] This will result in a significant mismatch between supply and demand, disproportionately affecting under-resourced communities that are already burdened by poverty, longer travel distances, and limited transportation. This could potentially lead to higher rates of mortality, disability, and shorter life expectancy. Given these dire realities, we must look to the all-encompassing "digital revolution" for solutions. Indeed, younger generations already expect on-demand, digital-first healthcare, much like they get from banking and shopping. Digital health technologies can mitigate some of the challenges associated with healthcare access, especially the shortage of healthcare workers,[17] including addressing the mismatch between supply and demand in under-resourced communities.[18] In recent years, for instance, health chatbots with text or voice-based interfaces have multiplied and become increasingly popular.[19] However, current health chatbots are mostly rule-based, typically depend on pre-scripted responses, seldom incorporate empathy, and lack personalized recommendations and multilingual capabilities.[20] This limits their ability to effectively serve diverse populations, particularly among individuals with low health literacy.[21] Furthermore, such health chatbots lack intelligent, up-to-date, evidence-based recommendation capabilities, real-time adaptability to the user, and other capabilities such as diagnostic testing and medication dispensing. Cultural differences can also prevent effective interaction with health chatbots and reduce utilization.[22]

Thus, there is a critical need for an innovative approach to address these limitations and improve healthcare access for diverse populations. The solution lies in scalable, cost-effective digital healthcare advances that ensure quality healthcare access for everyone without adding burden to existing healthcare systems. We have to reimagine the future of healthcare not just in hospitals, but in our homes, pharmacies, workplaces, and everyday spaces connected by artificial intelligence (AI).

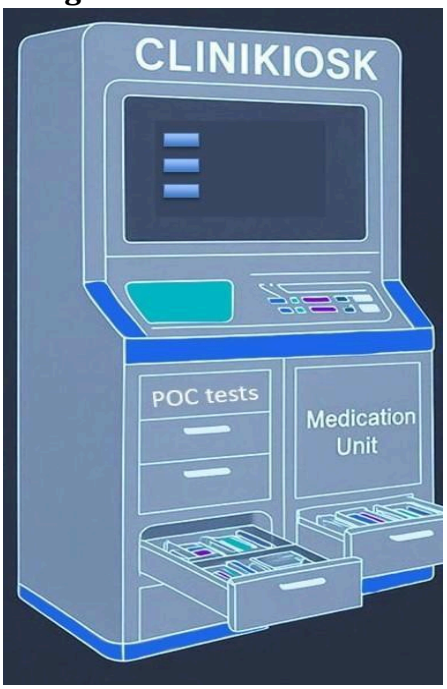
We propose to develop CliniKiosk (**Figure 1**), an AI-powered digital health kiosk designed to deliver real-time, evidence-based, multilingual, empathetic, and culturally sensitive health assessments. The system builds on recent advances in retrieval-augmented generation and large language models, combining structured data from authoritative medical sources with conversational AI. CliniKiosk dynamically adapts to a user's demographic characteristics, medical history and reported symptoms to produce personalized recommendations. In future versions, the kiosk will integrate Food and Drug Administration (FDA)–approved point-of-care diagnostic tests and dispense common medications directly at the kiosk. The proposed health-kiosk technology represents a transformative model for healthcare delivery, combining accessibility, efficiency, and patient empowerment. Strategically placed health kiosks can convert a four-hour trip to a healthcare facility into a 10-minute, on-site interaction, delivering timely access in rural clinics, pharmacies, and community centers.

In our recent cross-sectional survey of 297 respondents (63.6% healthcare professionals [HCPs] and 36.4% non–healthcare professionals [non-HCPs]), both groups reported broadly positive perceptions of AI in healthcare while also identifying substantial risks associated with its use.

Ethical concerns were reported by approximately 75% of HCPs and 65% of non-HCPs, while concerns related to bias and cultural or social barriers were reported by roughly half and more than 60% of respondents, respectively, across both groups. Notably, trust in AI's ability to make fair and ethical healthcare decisions was low in both populations, with 65% of HCPs and 60% of non-HCPs reporting little to no trust.

To address identified trust and ethical concerns, we designed CliniKiosk – a multilingual, agentic, retrieval-augmented digital health kiosk intended to deliver evidence-based, explainable, and empathetic health recommendations. We hypothesized that a stakeholder-informed, agentic, retrieval-augmented CliniKiosk would demonstrate improved trust-related performance, reflected in higher levels of empathetic communication, greater explainability of reasoning, and richer linguistic expression, while maintaining ethical integrity. The objective of this study was to evaluate whether a multilingual, agentic, retrieval-augmented digital CliniKiosk can deliver accurate, empathetic, explainable, and linguistically rich health recommendations.

Figure 1. CliniKiosk



Methods

System architecture

CliniKiosk consists of a React 18.3.1 frontend written in TypeScript and deployed via the Vite build system and hosted with real-time preview capabilities. A GitHub repository was created to contain each file of CliniKiosk. The backend infrastructure leverages Supabase's PostgreSQL database with Row Level Security (RLS) and edge functions to handle

serverless computation. A *four-component agentic architecture* manages the dialogue: 1. *Medical Router***:** uses GPT-4 to classify user intent and direct queries to appropriate modules. 2. *Medical Planner:* performs multi-step reasoning and selects the set of tools needed to answer each query. 3. *Medical Executor:* executes selected tools in parallel, including retrieval functions and API calls. 4. *Medical Manager:* synthesizes responses, validates safety, and constructs final output. Audio input is supported via the OpenAI Whisper-1 model. Speech is captured through the Web MediaRecorder API, encoded in WebM format, converted to base64, and streamed to the Whisper model. User authentication employs Supabase Auth with email/password flows, implementing OTP verification through a custom Edge Function that generates 6-digit codes, and email delivery utilizes SendGrid's REST API.

Retrieval-augmented generation (RAG) pipeline

The Retrieval-Augmented Generation (RAG) system queries six knowledge sources: PubMed, CrossRef, ClinicalTrials.gov, AHA, ACP, and NICE guidelines. To build the core Retrieval-Augmented Generation (RAG) pipeline, we developed modular API connectors for all six key medical knowledge sources to accept user queries, send authenticated requests, and normalize returned data. This is fed into an LLM (e.g., GPT-4o) using prompts that enforce evidence grounding and citation of sources.

Frontend and conversation manager

The chat interface presents conversational messages with associated citations, supports free-text input and optional voice input, and provides quick-action buttons for clarifying questions (e.g., asking for more details about symptoms or pain severity). To build the front end and conversation manager, we implemented a chat interface using React, supporting text input, optional voice input via Whisper, and quick-action buttons for clarifying questions. The front end

communicates with back-end endpoints (e.g., /api/chat), streams responses and handles network errors. Multilanguage support is added using i18next to translate UI elements, system prompts, and model outputs, with terminology normalization ensuring accurate clinical terms across languages. The system supports internationalization through react-i18next with browser language detection and maintains user conversation history with vector embeddings for personalized medical memory. The platform was deployed via Netlify after domain setup on GoDaddy.

Simulated cases

A clinician-verified dataset of 500 simulated patient cases was constructed. Cases reflected diverse demographics (age, gender, ethnicity, language, and health-literacy level), risk factors (comorbidities and lifestyle factors), acuity levels (chronic, acute, and emergent), and covered a wide variety of conditions across all organ systems with varying levels of clinical complexity. Each case included a chief complaint, history of present illness, past medical history, medications, allergies, vital signs, review of systems and emotional tone. An experienced board-certified physician generated the gold-standard differential diagnosis, final diagnosis, triage categories, and recommended treatments. To evaluate empathy, the ECCS (Empathic Communication Coding System) Framework was applied. Explainability was assessed using the System Causality Scale, and lexical diversity was quantified using the Measure of Textual Lexical Diversity (MTLD)

To build the evaluation and benchmarking suite, we organized the 500 simulated clinical cases in a structured format with full demographics, clinical details, and expert-adjudicated gold standards. A test harness iterates through these cases, sends each one to every model (e.g., GPT-4o, Claude 3 Opus, Mixtral, and Gemma), standardizes prompts, randomizes case order, and records outputs and timestamps. Scoring scripts then compare model responses to gold

standards to calculate accuracy and empathy/explainability metrics. CliniKiosk was compared against GPT-4o, Claude 3 Opus, Mixtral 8x7B (via OpenRouter), and Gemma 3 12B (via OpenRouter).

Evaluation metrics

The following metrics were computed for each of the 500 cases:

- Triage accuracy: proportion of cases with correct urgency classification (self-care, outpatient, urgent referral) compared with the gold standard.
- Differential-diagnoses accuracy: proportion of cases in which the set of differential diagnoses matched the physician-generated differential diagnoses.
- Final diagnosis accuracy: proportion of cases in which the final diagnosis matched the gold standard final diagnosis.
- Treatment-recommendation accuracy: proportion of cases in which the recommended treatment matched the physician-defined gold standard.
- Empathy: average score on the Empathic Communication Coding System (ECCS),[23] which assesses acknowledgement and validation of emotional or psychosocial concerns.
- Explainability: average score on the System Causality Scale,[24] measuring clarity and completeness of reasoning.
- Lexical diversity: measured using the Measure of Textual Lexical Diversity (MTLD),[25] quantifying vocabulary variety independent of response length.

Sample Size:

To justify the sample size, we assessed the precision of estimating a binary accuracy proportion. Using the conservative assumption of an expected accuracy of 50% and a 95% confidence level,

a sample of 500 cases yields a margin of error of approximately ±4.4%. This level of precision is appropriate for comparing model accuracy estimates, and the dataset provides sufficient power to detect moderate differences between models in pairwise comparisons.

Statistical analysis

Accuracy metrics (triage, differential diagnoses, final diagnosis, and treatment recommendations) were compared across models using Pearson χ^2 tests for overall differences. Pairwise comparisons were conducted with two-proportion z-tests, and p-values were adjusted using Holm's method to control for multiple comparisons. For empathy, explainability and lexical diversity reported as model-level means one-way analysis of variance (ANOVA) assessed overall differences, followed by Welch t-tests with Holm correction. Statistical significance was defined as $p < 0.05$. Effect sizes were calculated using Cohen's *d* for continuous outcomes.

Results

Triage accuracy

Across 500 triage scenarios, overall triage accuracy differed ($p = 0.043$) (Figure 2). CliniKiosk's triage accuracy was 88.6 % (443/500), statistically comparable to GPT-4o (89.4 %, $p = 0.67$), Claude 3 Opus (85.6 %, $p = 0.13$) and Gemma 3 12B (90.4 %, $p = 0.36$). CliniKiosk significantly outperformed Mixtral 8×7B (84.2 %, $p = 0.023$). GPT-4o and Gemma both outperformed Mixtral ($p = 0.0076$ and $p = 0.001$, respectively), and Gemma also outperformed Claude Opus ($p = 0.0117$).

Differential-diagnoses accuracy

Differential-diagnoses accuracy was uniformly high as shown in Figure 3. Across 500 cases per model, accuracy ranged from 92.2 % to 93.8 %, with no statistically significant differences

across the five systems (p = 0.865). Pairwise comparisons showed no significant differences between any model pairs (all p > 0.32), indicating that CliniKiosk, GPT-4o, Claude 3 Opus, Mixtral 8×7B and Gemma 3 12B performed equivalently in generating differential diagnoses.

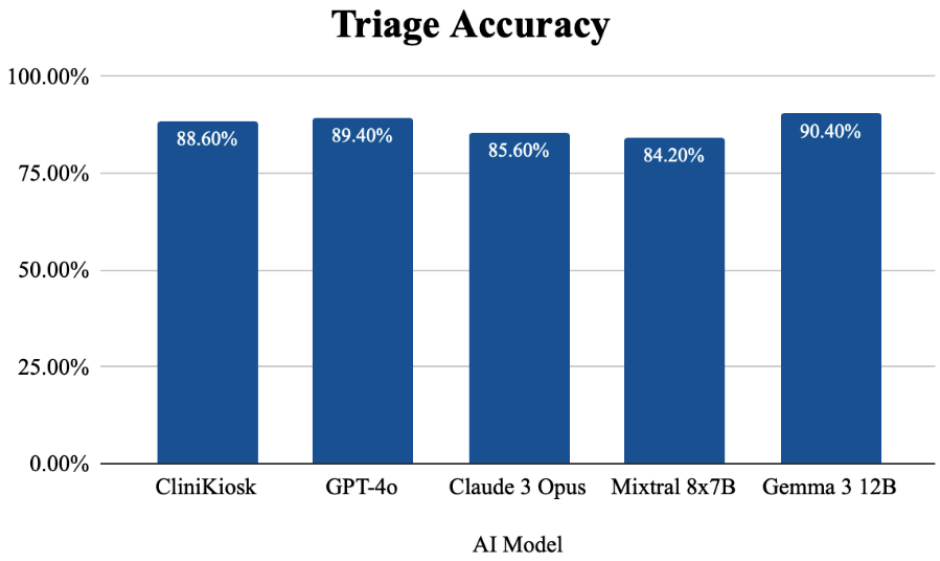


Figure 2. Triage Accuracy

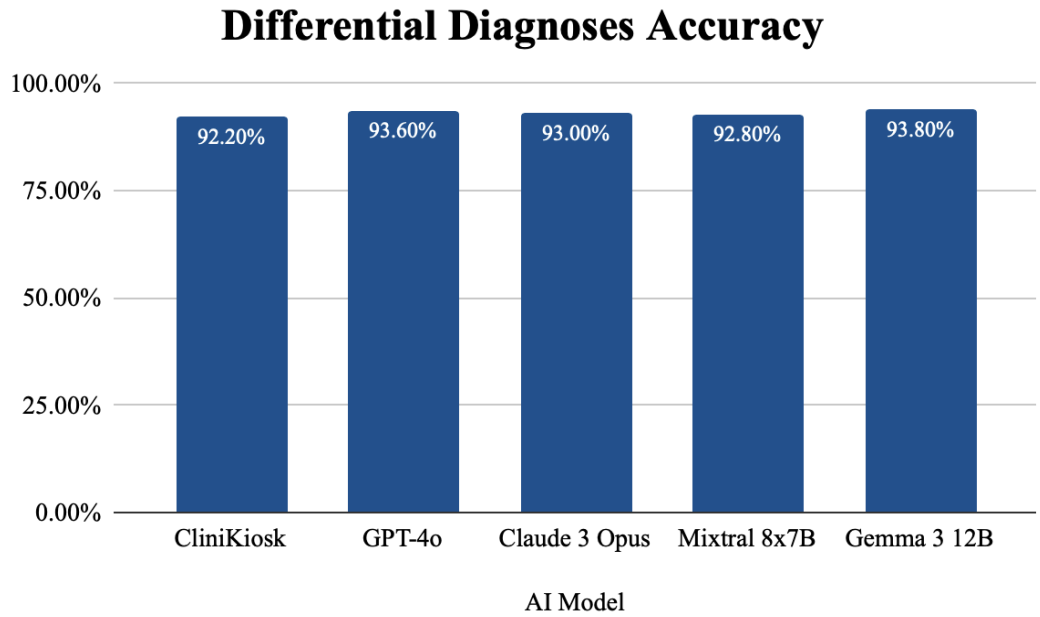


Figure 3. Differential Diagnoses Accuracy

Final-diagnosis accuracy

Final-diagnosis accuracy differed significantly across the five models ($p < 0.0001$) as shown in Figure 4. CliniKiosk achieved 76.0 % accuracy (380/500), significantly lower than GPT-4o (87.0 %, $p < 0.000001$), Claude 3 Opus (86.8 %, $p < 0.000001$), Mixtral 8×7B (81.0 %, $p = 0.019$) and Gemma 3 12B (84.8 %, $p < 0.0001$). GPT-4o and Claude Opus performed similarly ($p = 0.91$) and both significantly outperformed Mixtral ($p < 0.01$). Gemma was not significantly different from GPT-4o ($p = 0.27$) or Claude Opus ($p = 0.31$).

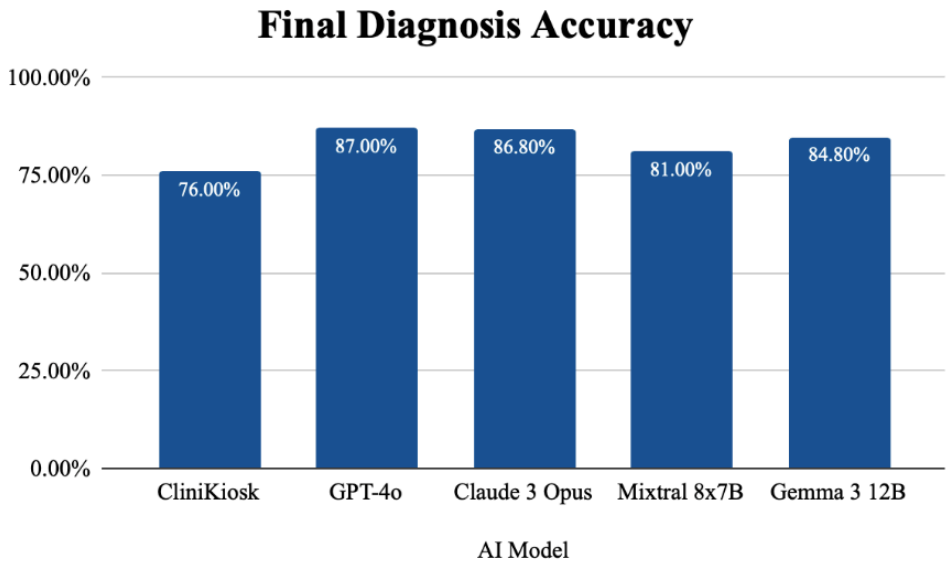


Figure 4. Final Diagnosis Accuracy

Treatment-recommendation accuracy

Among the 500 simulated patient cases, overall treatment-recommendation accuracy differed significantly between models ($p < 0.0001$) (Figure 5). CliniKiosk achieved 84.5 % accuracy (423/500), which was statistically similar to Mixtral 8×7B (86.4 %, $p = 0.33$) and Gemma 3 12B

(86.6 %, p = 0.29). In contrast, GPT‑4o (93.8 %, p < 0.00001) and Claude 3 Opus (94.2 %, p < 0.000005) performed significantly better than CliniKiosk. GPT‑4o and Claude Opus did not differ from one another (p = 0.82) but significantly outperformed Mixtral and Gemma (all p < 0.001).

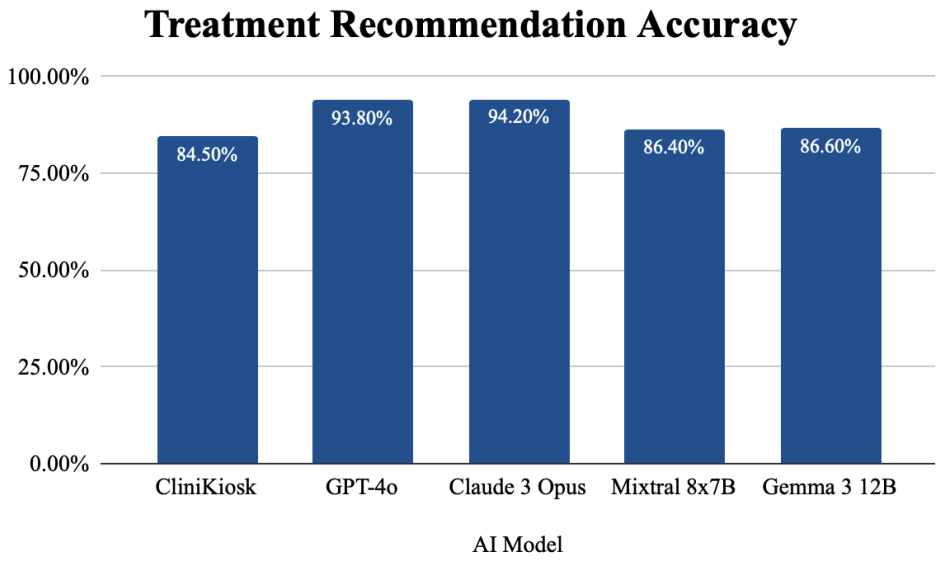


Figure 5. Treatment Recommendation Accuracy

Empathy

Empathy scores differed significantly across models (p < 0.0001) as shown in Figure 6. CliniKiosk achieved the highest mean ECCS score statistically comparable to Gemma 3 12B (p = 0.62) but significantly higher than GPT‑4o (p < 0.0001), Mixtral 8×7B (p < 0.0001) and Claude 3 Opus (p < 0.0001). Pairwise Welch t‑tests showed that GPT‑4o and Mixtral outperformed Claude Opus (p < 0.01), while Gemma showed significantly higher empathy than all models except CliniKiosk (p > 0.05). Effect sizes were large for comparisons between CliniKiosk and GPT‑4o (d = 0.82), Mixtral (d = 1.05) and Claude 3 (d = 1.35).

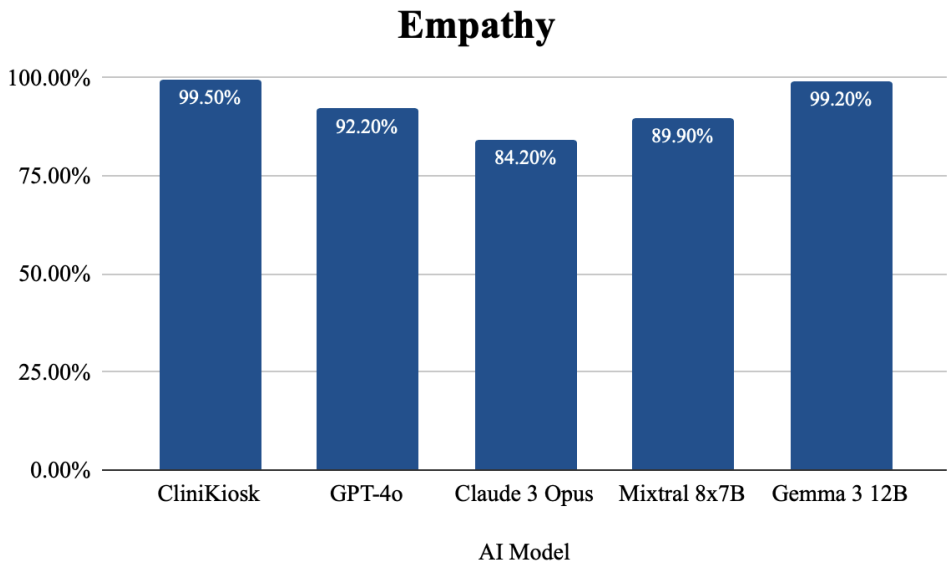


Figure 6. Empathy

Explainability

Explainability scores, measured using the SCS, differed significantly across models ($p = 0.0003$) (Figure 7). CliniKiosk achieved a mean score which is significantly higher than GPT-4o ($p = 0.013$), Claude 3 Opus ($p < 0.0001$) and Mixtral 8×7B ($p = 0.009$). The score was statistically comparable to Gemma 3 12B ($p = 0.20$). Among baseline models, Gemma outperformed GPT-4o ($p = 0.002$), Claude Opus ($p < 0.0001$) and Mixtral ($p = 0.0011$). Effect sizes were small to moderate.

Lexical diversity

Lexical diversity, quantified using MTLD, differed significantly across models ($p < 0.00001$) as shown in Figure 8. CliniKiosk demonstrated an MTLD score which was significantly lower than GPT-4o ($p < 0.0001$) and Claude 3 Opus ($p < 0.0001$), statistically comparable to Mixtral 8×7B ($p = 0.053$) and not significantly different from Gemma 3 12B ($p = 0.13$). Claude Opus exhibited

the highest lexical diversity and outperformed GPT-4o ($p = 0.026$), Mixtral ($p = 0.0004$) and Gemma ($p < 0.0001$). Effect sizes were small to moderate ($d = 0.28–0.52$).

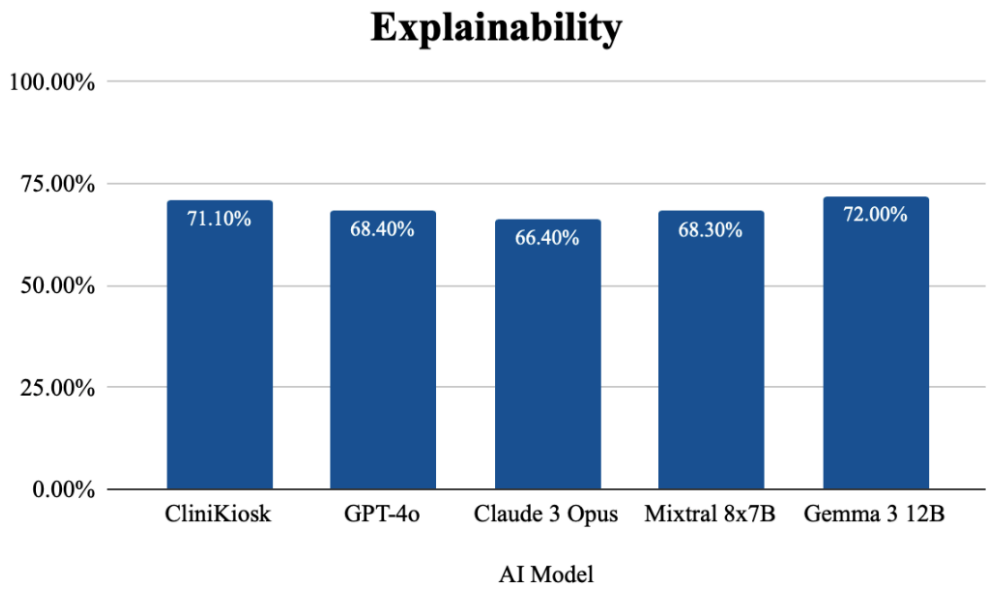


Figure 7. Explainability

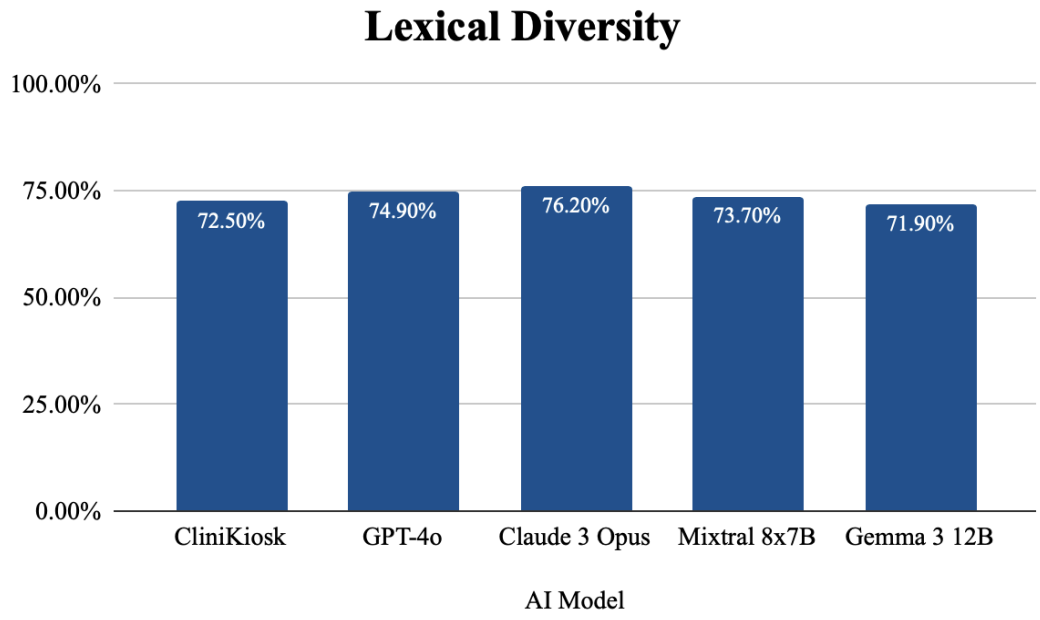


Figure 8. Lexical Diversity

Discussion

This study shows that a domain-specific, multilingual RAG-based kiosk can approach or surpass the performance of general-purpose LLMs in several important areas. CliniKiosk delivered triage and treatment recommendations on par with best performing open AI models and achieved the highest empathy scores among all systems evaluated, despite having far fewer parameters and more focused training scope than models such as GPT-4o and Claude 3 Opus. The most remarkable finding was CliniKiosk's high empathy score. Empathy in digital health tools is consistently identified as a key determinant of user trust, usability, and patient adherence especially among individuals with limited health literacy or prior adverse healthcare experiences.[26,27] Clinikiosk's strong empathy performance reflects the system's use of tailored prompting, culturally adaptive phrasing, and avoiding overly technical or emotionally detached language, features that are often missing from generic chatbots. General LLMs may provide

clinically accurate answers but often lack consistency in empathetic tone, as demonstrated in multiple prior assessments of AI patient-facing communication. Furthermore, the explainability advantage noted in CliniKiosk indicates that retrieval-grounded reasoning supports clearer, and more structured responses. In contrast, general LLMs may fabricate explanatory mechanisms, provide inconsistent causal reasoning, or deviate from evidence-based guidance.

The lower final-diagnosis accuracy observed in CliniKiosk (76%) underscores the challenge of selecting a single diagnosis based on text-only clinical vignettes. This may be attributable to its narrower training data or conservative decision thresholds designed to prioritize safety. These findings mirror prior evidence that general LLMs, particularly GPT-4–class systems outperform domain-specific models in precise diagnostic reasoning due to their broad training data and ability to integrate subtle contextual cues. However, because kiosks are not intended to definitively diagnose serious conditions, the capacity to list appropriate differentials and recommend safe triage pathways may be more clinically meaningful than single-label diagnostic accuracy. Differential-diagnosis accuracy exceeding 92 % across all models indicates that the major diagnostic possibilities are consistently captured, which is generally sufficient for triage and referral.

The finding that triage accuracy surpassed 88% and did not differ significantly from the highest-performing models in several comparisons is important as triage rather than final diagnosis represents the primary safety gate for autonomous or semi-autonomous digital health systems. High-acuity misclassification is a major concern in AI-supported care, and the comparable triage performance suggests that a focused, RAG-supported system can prioritize

safety without relying on the full generative capacity of state-of-the-art LLMs. Lexical diversity showed modest performance for CliniKiosk, which may actually enhance usability. Complex vocabulary can hinder comprehension for users with low health literacy, a group disproportionately represented in underserved communities. A slightly reduced MTLD score means CliniKiosk uses more consistent and accessible language, aligning with evidence that simplified phrasing improves patient comprehension without reducing clinical accuracy.

Our findings are consistent with prior published evidence showing that existing medical chatbots often lack personalization, context awareness, empathy and explainability. The strong empathy and explainability demonstrated by CliniKiosk reinforce arguments made by Laymouna et al. and Popa et al. advocating for patient-centric, culturally sensitive digital tools.[17,21] At the same time, the superior diagnostic accuracy of large LLMs such as GPT-4o and Claude 3 Opus is consistent with reports that general-purpose models excel in medical reasoning yet may generate hallucinations or insensitive responses. By integrating LLMs capabilities with curated medical knowledge and safety filters, CliniKiosk aims to balance accuracy with empathy and transparency.

The results indicate that patient-centered AI kiosks could help mitigate health-care access disparities by providing on-demand, culturally aligned guidance in multiple languages. Such kiosks could serve as initial points of contact in rural communities, retail pharmacies, public libraries, workplaces, and schools, and settings where traditional healthcare resources are limited. Strong performance in empathy and explainability may enhance trust and adherence, countering concerns that digital tools dehumanize care. With proper regulatory oversight,

integration of point-of-care diagnostics and medication dispensing could further reduce barriers for underserved populations. Real-world implementation studies, ongoing monitoring for safety and bias, and integration with telehealth and electronic health records are essential next steps. CliniKiosk is not designed to replace clinicians but to augment their capacity, triage low-acuity cases and provide health education allowing providers to focus on complex care.

Limitations

Several important limitations of this study should be noted. First, the evaluation used 500 simulated case scenarios, which, while diverse, may not represent the full complexity and variability of real-world patient presentations. Second, models were evaluated in a controlled environment; performance in busy clinical settings, where user input may be unclear and resources limited, is uncertain. Third, bias analyses were limited to basic demographic variables; disparities related to socioeconomic status, comorbidities or regional factors were not explored. Fourth, the study did not examine user trust, interpretability, or clinical workflow integration, all critical for successful deployment. Finally, the simulated scenarios may favor text-based models and may not reflect sensor data or lab results that real kiosks might use.

Conclusions

Health-care access gaps, driven by workforce shortages, facility closures and socioeconomic inequities, demand scalable innovative solutions. Our findings demonstrate that a multilingual, RAG-based kiosk can deliver empathetic, evidence-based recommendations and triage decisions with accuracy comparable to leading LLMs.

CliniKiosk demonstrated superior empathy and explainability while using more accessible language, aligning human-centered communication with trust and transparency. Scaling CliniKiosk and incorporating real-world data and diagnostics may help broaden access to high-quality health care particularly in underserved communities while complementing, rather than replacing healthcare providers.

References

1. World Health Organization; World Bank. Tracking Universal Health Coverage: 2023 Global Monitoring Report. Geneva, Switzerland: World Health Organization; 2023.

2. Akinyemi JO. Healthcare access disparities among marginalized communities. Glob Public Health Manag Netw. 2024;2(1):44-59.

3. Wen H, Sadeghi N. Racial disparities in access to health care infrastructure across US counties. Front Public Health. 2023;11:897007.

4. Rural Health Redesign Center. 20 States With the Most Rural Hospital Closures. Pennsylvania Office of Rural Health; 2024. Accessed November 29, 2025. https://www.porh.psu.edu/20-states-with-the-most-rural-hospital-closures/

5. 705 hospitals at risk of closure, state by state. Becker's Hospital CFO Report. Accessed November 29, 2025. https://www.beckershospitalreview.com/finance/705-hospitals-at-risk-of-closure-state-by-state-november-2024.html.

6. United States Department of Agriculture Economic Research Service. 146 Rural Hospitals Closed or Stopped Providing Inpatient Services Between 2005 and 2023. USDA ERS; 2025.

Accessed November 29, 2025.

http://ers.usda.gov/data-products/charts-of-note/chart-detail?chartId=110927

7. Chartis. 2025 Rural Health – State of the State: Rural Hospital Closures & Care-Access Crisis. Chartis; 2025. Accessed November 29, 2025.

https://www.chartis.com/insights/2025-rural-health-state-state

8. Garcia MC, Faul M, Massetti G, et al. Premature and preventable deaths in rural versus urban counties in the United States. MMWR Surveill Summ. 2024;73(SS-x):1-12.

9. Centers for Disease Control and Prevention. Rural Americans Are at Higher Risk of Early Death Than Urbanites. CDC; 2024. Accessed November 29, 2025.

https://abcnews.go.com/Health/rural-americans-higher-risk-early-death-urbanites-cdc/story?id=109742216

10. Rural Health Information Hub. Rural Health Disparities. Rural Health Information Hub; 2025. Accessed November 29, 2025.

https://www.ruralhealthinfo.org/topics/rural-health-disparities

11. AMN Healthcare. 2025 Survey of Physician Appointment Wait Times. AMN Healthcare; 2025. Accessed November 29, 2025.

12. Association of American Medical Colleges. Walk-in clinics tackle urgent mental health needs. AAMC News. August 21, 2024. Accessed November 29, 2025.

13. Derlet RW, Richards JR. Overcrowding in the nation's emergency departments: complex causes and disturbing effects. Ann Emerg Med. 2000;35(1):63-68.

14. World Health Organization. Health Workforce: Overview and Key Facts. Geneva, Switzerland: World Health Organization; 2020. Accessed November 29, 2025.

https://www.who.int/health-topics/health-workforce

15. Mercer. 2021 U.S. Healthcare Labor Market: The Future of Healthcare Labor. New York, NY: Mercer; 2021. Accessed November 29, 2025.

16. American Hospital Association. Data Brief: Health Care Workforce Challenges Threaten Hospitals' Ability to Care for Patients. American Hospital Association; October 2021. Accessed November 29, 2025. https://www.aha.org/fact-sheets/2021-11-01-data-brief-health-care-workforce-challenges-threaten-hospitals-ability-care

17. Popa V, Geissler J, Vermeulen R, Priest E, Capperella K, Susuzlu G, Terry SF, Brooke N. Delivering Digital Health Solutions that Patients Need: A Call to Action. Ther Innov Regul Sci. 2024 Mar;58(2):236-241.

18. Sun G, Zhou X. AI in healthcare: navigating opportunities and challenges with chatbots for digital health. Front Public Health. 2023;11:1307636.

19. Rebelo F, et al. Chatbots in healthcare: A systematic literature review and directions for future research. Comput Biol Med. 2022;150:106081.

20. Abd-Alrazaq AA, et al. Technical Aspects of Developing Chatbots for Medical Applications: Scoping Review. J Med Internet Res. 2020;22(12):e19127.

21. Laymouna M, et al. Roles, Users, Benefits, and Limitations of Chatbots in Healthcare: A Rapid Review. J Med Internet Res. 2024;26(7):e484931.

22. Jin E, Ryoo Y, Kim WJ, Song YG. Bridging the health literacy gap through AI chatbot design: the impact of design cues on trust and use intention for low health literacy users. Internet Res. 2024;35(3):1299-1329.

23. Koenig CJ, Young AJ, Stephenson J, et al. Empathic communication in dignity therapy: feasibility of the Empathic Communication Coding System (ECCS). Patient Educ Couns. 2022;105(10):2783-2791.

24. Holzinger A, Müller H, Kieseberg P, et al. The System Causability Scale (SCS): assessing the quality of machine explanations. Artif Intell Med. 2020;103:101792

25. McCarthy PM, Jarvis S. MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. Behav Res Methods. 2010;42(2):381-392.

26. Hojat M, Louis DZ, Markham FW, Wender R, Rabinowitz C, Gonnella JS. Physicians' empathy and clinical outcomes for diabetic patients. Bull Med Libr Assoc. 2011;89(3):291-298.

27. Topol EJ, et al. A systematic review and meta-analysis of empathy in patient care: human clinicians vs generative AI chatbots. Br Med Bull. 2025;156(1):ldaf017