Rx-AI: A Retrieval-Augmented Health Agent for Truly Personal Care

Himanshu Joshi
University of Texas, Austin
USA
himanshujoshi@utexas.edu

Sunita Kumari
University of Texas, Austin
USA
sunitakumari@utexas.edu

ABSTRACT

This paper presents the design and implementation of an AI-powered health agent leveraging Retrieval-Augmented Generation (RAG) for personalized health management. The system integrates large language models (LLMs) with curated medical knowledge bases and dynamic user health profiles (modeled using FHIR principles) to provide tailored, evidence-grounded health recommendations via conversational interaction. We detail the RAG-based architecture addressing challenges of knowledge grounding and personalization in healthcare AI. Key components include a vector database for semantic knowledge retrieval, a User Profile Manager incorporating longitudinal data, a Health Recommendation Engine employing context-aware prompt engineering, and robust safety mechanisms including emergency detection and explicit scope limitations. Preliminary evaluations demonstrate the agent's ability to generate contextually relevant, personalized guidance while appropriately deferring to professional medical advice, highlighting the potential of RAG-enhanced LLMs for responsible health support.

CCS CONCEPTS

• **Computing methodologies** → Discourse, dialogue and pragmatics; • **Applied computing** → Health informatics; • **Information systems** → Retrieval-augmented generation.

KEYWORDS

health agent, personalized healthcare, large language models, retrieval-augmented generation, conversational AI, health informatics, RAG, explainable AI, FHIR

1 INTRODUCTION

The demand for continuous, personalized health guidance is escalating, driven by the rising prevalence of chronic conditions [2] and the inherent limitations of traditional, episodic healthcare delivery. While digital health technologies offer potential solutions, generic advice often falls short, failing to account for individual complexities, histories, and contexts.[3] Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation, suggesting potential applications in health guidance.[4, 5] However, deploying standalone LLMs in the high-stakes healthcare domain presents significant challenges. These models often suffer from knowledge cutoffs (leading to outdated information), a propensity for hallucination (generating factually incorrect statements), and a lack of domain-specific depth [6, 7, 8, 9], all critical limitations when providing health advice.

Creating AI health agents that deliver trustworthy, personalized, and contextually relevant support necessitates overcoming these LLM limitations. This requires robust mechanisms for grounding responses in reliable, up-to-date medical knowledge and dynamically tailoring guidance based on comprehensive, longitudinal user health data, all while ensuring user safety and system transparency. Many existing health chatbots lack this deep integration, often relying on rule-based systems with limited conversational flexibility or failing to provide adequately personalized or evidence-based advice.[3, 10, 11, 12]

This paper details the design and implementation of an AI Health Agent employing Retrieval-Augmented Generation (RAG) [9, 13, 14, 15] to specifically address these challenges. RAG architecture allows the underlying LLM to access and incorporate information retrieved from external, verified knowledge sources during the response generation process.[6, 16] This approach directly mitigates the risks of hallucination and knowledge staleness by grounding the agent's responses in evidence. Furthermore, the RAG framework provides a natural mechanism for integrating dynamically retrieved user-specific context from a dedicated profile manager, enabling evidence-grounded personalization.[17, 18] Unlike general-purpose chatbots or earlier generations of health bots [5], our agent architecture focuses on the tight integration of RAG with structured user profiles, inspired by health data standards like HL7 FHIR [19, 20, 21, 22], and longitudinal data to achieve nuanced, context-aware personalization.[23, 24]

The key contributions of this work include: (1) The design of a RAG-based system architecture tailored for personalized conversational health management. (2) The integration of a dynamic User Profile Manager, conceptualized using FHIR principles, into the RAG pipeline. (3) The implementation of context-aware prompt engineering strategies designed to generate personalized, knowledge-grounded recommendations.[25, 26, 27, 28, 29, 30] (4) Detailed description of safety mechanisms, including logic for emergency situation detection [31, 32, 33] and the use of explicit scope limitation disclaimers.[34, 35, 36]

The adoption of RAG is not merely a technical enhancement but a fundamental architectural decision for building responsible AI health agents. It directly addresses the critical need for verifiable knowledge grounding [6], a necessity stemming from the limitations of standalone LLMs and the high-stakes nature of healthcare. Concurrently, RAG provides a flexible mechanism for incorporating the deep, dynamic personalization required for effective health management support [23, 37], moving beyond the capabilities of earlier chatbot generations that often struggled with accuracy and personalization depth.[5, 10]

2 RELATED WORK

Our approach builds upon several interconnected research areas:-

2.1 Health Chatbots and Virtual Assistants

Early explorations into health-focused conversational agents demonstrated potential but also significant limitations. Studies like Miner et al. [1] found that general-purpose voice assistants struggled with complex health queries. Subsequent chatbots often faced challenges in providing empathetic responses, ensuring consistent information accuracy, and achieving user engagement.[3, 11] Systematic reviews frequently report mixed results regarding efficacy and usability, often highlighting a lack of rigorous evaluation methodologies and insufficient technical detail in published studies.[2, 3, 5, 10, 38] Many early systems relied on predefined rules or scripts, limiting their flexibility and ability to handle diverse user inputs [12], or lacked the mechanisms for deep personalization.[3] Our work distinguishes itself by employing a RAG-enhanced LLM, aiming for greater conversational flexibility, verifiable knowledge grounding, and integration with structured personalization data.

2.2 Knowledge-Grounded LLMs (RAG) in Healthcare

The inherent limitations of standalone LLMs for medical applications—such as knowledge confined to training data cutoffs and the potential for generating plausible but incorrect information (hallucination)—are widely recognized.[6, 8] Retrieval-Augmented Generation has rapidly emerged as a primary technique to address these issues.[13, 14] RAG systems enhance LLMs by dynamically retrieving relevant information from external, often domain-specific, knowledge sources (e.g., PubMed databases, curated medical guidelines, textbooks) during the generation process.[6, 7, 16, 17, 39, 40, 41, 42, 43] Numerous studies have demonstrated that RAG significantly improves LLM performance on medical question-answering tasks [6, 16] and enhances the reliability of outputs for clinical assessment support.[7] Frameworks like GUIDE-RAG offer structured guidelines for developing RAG applications in clinical settings.[6] The demonstrated success and rapid adoption of RAG in biomedical NLP [6, 7, 16] indicate its crucial role in building credible, state-of-the-art healthcare AI. Our work applies RAG specifically within a conversational health management context, integrating its grounding capabilities with dynamic user profiles to move beyond static question-answering towards personalized dialogue.

2.3 Personalization in Digital Health

Effective digital health interventions hinge on personalization.[23, 37] Techniques vary widely, leveraging data from Electronic Health Records (EHRs) [44], wearable sensors (tracking activity, sleep, vitals) [24], genomic data [24], and patient-reported outcomes to tailor advice and support. AI and machine learning models are increasingly used to analyze these diverse data streams, identify individual patterns, predict health trajectories, and optimize interventions.[45, 46] Advanced concepts like digital twins aim to create comprehensive virtual patient models for highly individualized simulation and prediction.[24] Our agent employs a multi-faceted personalization strategy, combining explicitly stored user profile information (demographics, diagnosed conditions, medications—modeled using FHIR concepts [22]) with longitudinal data (tracked vitals, activity levels, sleep patterns) retrieved during the RAG process to inform both knowledge retrieval relevance and LLM prompt construction.[25]

2.4 Safety, Ethics, and Explainability (XAI) in Healthcare AI

The deployment of AI in healthcare necessitates stringent attention to safety, ethics, and transparency.[47, 48, 49, 50] Critical concerns include ensuring patient data privacy and security (e.g., compliance with HIPAA [51, 52]), mitigating algorithmic bias, establishing accountability for potential errors [47], and maintaining appropriate human oversight.[34] Explainable AI (XAI) techniques, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), aim to increase the transparency of complex "black-box" models by providing insights into how predictions are derived from input features.[53, 54, 55, 56] These methods have shown applicability in interpreting models trained on EHR data or medical text.[53, 57, 58, 59, 60] While our current implementation prioritizes safety through knowledge grounding (via RAG, which provides source traceability [6, 17, 30]) and explicit safety mechanisms (emergency detection, scope limitation), integrating more formal XAI methods represents an important direction for future work.[49]

3 METHODOLOGY

3.1 System Architecture

The AI Health Agent is designed as a modular system comprising four interconnected components, orchestrated to deliver personalized, knowledge-grounded health advice through a RAG-centric workflow (Figure 1). The core components are: (1) a Health Knowledge Base, (2) a User Profile Manager, (3) a Health Recommendation Engine, and (4) an Interface Layer.

The interaction flow begins at the **Interface Layer**, which captures the user's natural language query. This query is passed to the **Health Recommendation Engine**, the central coordinating component. The Engine first queries the **User Profile Manager** to retrieve relevant, up-to-date user context, including demographic details, known health conditions, medications, recent vital sign trends, and conversation history. Concurrently or subsequently, the Engine formulates a query, potentially refined using the retrieved user context [26], to search the **Health Knowledge Base**. The Engine receives retrieved knowledge snippets and combines them with the user's query and profile context within a structured prompt template. This augmented prompt is then processed by an LLM integrated within the Engine. The LLM generates a candidate response, which the Engine post-processes through safety filters (e.g., emergency detection, scope adherence checks) before delivering the final, grounded, and personalized response back to the user via the Interface Layer. This architecture explicitly implements the RAG pattern [13, 14, 61], ensuring that LLM generation is consistently informed by both external evidence and individual user data.

3.2 Knowledge Retrieval and Integration (RAG Implementation)

The RAG mechanism is central to ensuring the agent's responses are evidence-based and up-to-date.

Health Knowledge Base:- This component is implemented using a vector database (e.g., Chroma, Pinecone, FAISS, selected

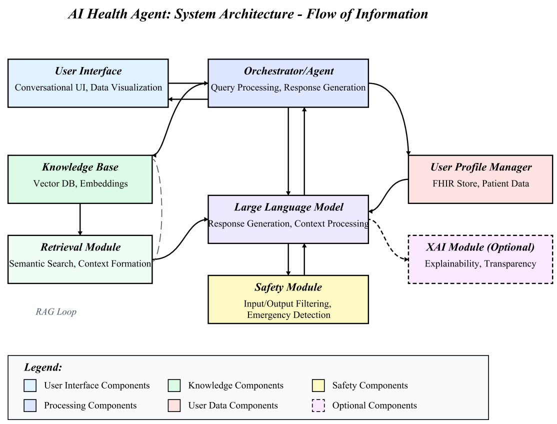

Figure 1: AI Health Agent RAG-based System Architecture.

based on scalability and latency requirements [62]). It stores curated, evidence-based health information derived from reputable sources (e.g., Mayo Clinic articles, American Heart Association guidelines, potentially indexed PubMed abstracts [6, 42, 43]). Source documents are pre-processed by segmenting them into smaller, semantically meaningful chunks (e.g., using recursive character splitting, with chunk size optimized for medical content retrieval [8, 25]). Each chunk is then converted into a high-dimensional vector embedding using a suitable pre-trained model optimized for biomedical text (e.g., BioBERT, Sentence-BERT variants [57]). These embeddings are stored and indexed in the vector database.[8, 9] Crucially, relevant metadata, such as the original source document and publication date, is stored alongside each chunk's embedding to enable source attribution.[8, 43]

Retrieval Process:- When a user query is received, it is first converted into a vector embedding using the same embedding model employed for the knowledge base. The Recommendation Engine then performs a semantic search (typically using cosine similarity) against the vector database to retrieve the top-k document chunks whose embeddings are most similar to the query embedding.[8, 9, 63, 64] While not implemented in the current version, future enhancements could incorporate hybrid search approaches (combining semantic similarity with keyword matching [63]) or re-rank retrieved chunks based on their relevance to the specific user's profile, adding another layer of personalization to the retrieval step itself.

Augmentation and Generation:- The retrieved knowledge chunks $\{d_i\}$ are formatted and integrated into a carefully designed prompt template. This template also incorporates the original user query Q and relevant contextual information U extracted from the User Profile Manager. The final augmented prompt is then passed to the core LLM (e.g., GPT-4, Llama 3 [7, 65]) within the Recommendation Engine. The LLM generates the response A based on this rich context: $A = \text{LLM}(\text{prompt}(Q, U, \{d_i\}))$.[8, 39, 41] This process ensures the LLM's generative capabilities are steered by both the user's specific situation and verified external knowledge.

3.3 Personalization Mechanism

Effective health guidance requires deep personalization, moving beyond generic advice.

User Profile Manager:-This component maintains persistent, structured information about each user. To ensure data consistency and potential future interoperability, the profile schema is inspired by HL7 FHIR (Fast Healthcare Interoperability Resources) standards.[19, 20, 21, 22, 44, 66, 67] Key FHIR resource concepts utilized include:-

- *Patient*:- Stores core demographic data (age, gender, unique identifier).
- *Condition*:- Records diagnosed health conditions (e.g., hypertension, diabetes).
- *MedicationStatement*:- Lists current and potentially past medications.
- *Observation*: Captures longitudinal, time-series data, crucial for identifying trends. This includes vital signs (blood pressure, heart rate), activity metrics (steps, active minutes), sleep duration, nutrition logs, and patient-reported symptoms or mood states. Standardized coding systems (like LOINC) are used where applicable for semantic consistency.

This structured approach [22] provides a robust foundation for incorporating diverse health data points relevant to personalization.

Personalization Techniques:- Personalization is achieved through multiple mechanisms integrated into the RAG workflow:-

- *Profile-Aware Retrieval (Future Work):* -As mentioned, user profile data (conditions, medications) could be used to filter or boost the relevance scores of retrieved knowledge chunks during the search phase, ensuring the information presented is highly pertinent to the individual's health status.
- *Contextual Prompt Engineering:*- This is the primary personalization method in the current implementation. Relevant elements from the user's profile (e.g., age, specific conditions like 'hypertension', current medications like 'Lisinopril', recent trends like 'average sleep 6.2 hours', known allergies) are dynamically selected and inserted into the LLM prompt template alongside the query and retrieved knowledge.[25, 29] This explicitly guides the LLM to consider the user's specific context when formulating its response. An example prompt structure might resemble: "User Profile: Age [age], Gender [gender], Conditions [list of conditions], Medications [list of medications], Recent Vitals [key recent observations/trends]. User Query: '[user query]'. Retrieved Knowledge: [formatted knowledge chunks]. Task: Based *only* on the provided profile and retrieved knowledge, generate a helpful, non-diagnostic health suggestion. Address the user's query considering their specific health context. Cite the source for any medical information provided. Include a disclaimer that this is not medical advice and advise professional consultation if symptoms persist or worsen.".[25, 27, 28, 30]
- *Longitudinal Awareness:*- The system can reference trends observed in the user's longitudinal data (stored as 'Observation' resources). For example, the prompt might include information like "Note: User's average sleep duration has

decreased by 1 hour over the past week," enabling the LLM to generate advice sensitive to recent changes.
- *Conversational Context:-* The Recommendation Engine maintains a short-term memory of the current conversation turn, allowing the agent to understand follow-up questions and maintain dialogue coherence.[8, 25]

Using FHIR concepts for modeling provides a standardized, extensible structure, facilitating the integration of diverse health data and enabling more robust, multi-faceted personalization compared to ad-hoc profile designs.[22]

3.4 Safety Mechanisms

Ensuring user safety is paramount given the healthcare context. A multi-layered approach is implemented:

Medical Emergency Detection:- This critical safeguard aims to identify situations requiring immediate professional medical attention. It combines keyword/phrase pattern matching with potential intent classification. A predefined list of keywords and phrases indicative of emergencies (e.g., "chest pain," "severe difficulty breathing," "sudden weakness," "uncontrolled bleeding," "suicidal thoughts," based on established guidelines like those from ACEP [32]) is scanned in the user input.[33, 68] If a high-risk pattern is detected, the system immediately halts the standard response generation process and outputs a clear, predefined directive urging the user to seek immediate help (e.g., "Based on your description, this could be a medical emergency. Please call your local emergency number or go to the nearest emergency room immediately.").[31, 32, 69, 70] Future work includes training a dedicated machine learning classifier for more nuanced emergency detection.[31, 69]

Scope Limitation and Disclaimers:- The agent explicitly communicates its limitations to manage user expectations and mitigate risks of misuse or over-reliance.[3] This includes:

- An initial disclaimer presented at the start of the first interaction.
- Clear statements embedded within responses, reinforcing that the agent is an AI, cannot provide diagnoses or medical advice, and is not a substitute for professional healthcare consultation.[34, 35, 36, 52, 71, 72, 73] Example: "As an AI assistant, I cannot provide medical diagnoses or replace the advice of a healthcare professional. This information is for general guidance only. Please consult your doctor for any health concerns.".[34, 36]

Knowledge Grounding (via RAG):-The RAG architecture itself acts as a safety feature by ensuring that the information provided is based on retrieved evidence from the curated knowledge base, rather than potentially inaccurate or fabricated information generated solely by the LLM.[6]

Source Attribution:- To enhance transparency and allow users to verify information, all health guidance derived from the knowledge base explicitly cites the source(s) (e.g., "Source: Mayo Clinic, 2024").[6, 34]

Content Filtering (Potential):- Integration with external content safety services (e.g., Azure Content Safety [34]) can provide an additional layer to filter potentially harmful, biased, or inappropriate LLM outputs.

This layered strategy, combining technical detection, architectural design for grounding, and clear communication of limitations, is essential for responsible AI deployment in healthcare.[3, 34]

4 RESULTS

4.1 System Capabilities

The implemented AI Health Agent demonstrates the core capabilities outlined in the methodology:-

- **Conversational Health Management:-** Engages in natural language dialogue on various health topics, maintaining context across conversational turns.
- **Knowledge-Grounded Responses:-** Leverages the RAG pipeline to retrieve relevant information from the Health Knowledge Base and integrates it into responses, providing evidence-based guidance with source attribution.
- **Personalized Recommendations:-** Tailors advice based on user-specific data retrieved from the User Profile Manager (incorporating demographics, conditions, medications, and longitudinal metrics).
- **Safety-Conscious Interaction:-** Implements emergency detection logic to identify critical situations and provides clear disclaimers regarding its scope and limitations.

4.2 Usage Example

The following dialogue excerpt illustrates the agent's ability to integrate user profile data and retrieved knowledge for personalized, safe guidance:-

---

User:- I've been feeling more tired than usual lately. Any suggestions?

Agent:- Based on your profile, I notice a few factors that might contribute to your fatigue. Your recent sleep logs show an average of 6.2 hours per night, which is below the recommended 7-9 hours for adults. Also, given your hypertension condition and current medication (Lisinopril), fatigue can sometimes be a side effect.

Here are some personalized suggestions grounded in general health guidelines:-

(1) Aim to gradually increase your sleep duration towards the recommended 7 hours per night. Consistent sleep is crucial for energy levels. (Source: National Sleep Foundation, 2023)
(2) Ensure adequate hydration throughout the day, as dehydration can contribute to feelings of tiredness. Your recent logs suggest intake could be improved. (Source: Mayo Clinic, 2024)
(3) Consider adjusting the timing of moderate exercise to earlier in the day, as late-day intense activity can sometimes interfere with sleep. (Source: American College of Sports Medicine, 2022)
(4) Continue to monitor your blood pressure regularly, as significant fluctuations can impact energy levels, especially with hypertension. (Source: American Heart Association Guidelines, 2023)

Remember, I am an AI assistant and cannot provide medical advice. If your fatigue persists for more than two weeks despite trying these general suggestions, or if it is severe, it is important to consult your healthcare provider. They can help rule out other potential causes and discuss whether your medication might be contributing.

This example demonstrates the integration of profile data (sleep logs, hypertension, Lisinopril), generation of personalized suggestions linked to this data, grounding in external sources with attribution, and appropriate safety messaging advising professional consultation.

4.3 Preliminary Evaluation and Limitations

The current system represents an initial implementation and has several limitations requiring further work:-

Knowledge Base Scope:- The curated Health Knowledge Base currently covers common health topics but lacks depth in specialized medical domains. Scaling the KB using larger corpora (e.g., full PubMed indexing [6, 43]) introduces challenges related to maintaining retrieval precision, managing computational costs, and ensuring the quality and currency of information.[6]

RAG Limitations:- The effectiveness of the agent is dependent on the RAG pipeline. Potential issues include retrieving irrelevant or incomplete information, difficulties in synthesizing conflicting information from multiple retrieved sources, and inherent LLM context window limitations restricting the amount of retrieved data that can be processed effectively.[6, 8, 64]

Personalization Depth:- While personalization incorporates profile data, it does not yet model complex interactions between various health factors (e.g., comorbidities, lifestyle, genetics [24]). The integration with FHIR is conceptual; real-time, secure integration with live EHR systems via standards like SMART on FHIR [21, 67] remains a significant future step.[6, 44]

Evaluation Rigor:- Current evaluation is preliminary, based primarily on qualitative assessment of example interactions and internal testing of component functions (e.g., safety trigger checks). A formal, rigorous evaluation is necessary, employing established frameworks and metrics.[38, 74, 75] This should include usability testing (e.g., System Usability Scale - SUS [75, 76]), task completion analysis, safety assessments in realistic scenarios, and eventually, clinical trials to measure impact on health behaviors and outcomes.[38]

Explainability:- The system currently lacks dedicated XAI capabilities beyond source attribution.[49, 53] Users and clinicians cannot easily understand the specific reasoning process connecting the user's profile, retrieved knowledge, and the final recommendation.

5 DISCUSSION

This work positions a RAG-based AI health agent as a promising approach to delivering personalized, evidence-grounded conversational health support. By integrating an LLM with external knowledge retrieval and dynamic user profiles, the system aims to overcome key limitations of both earlier rule-based chatbots [12] and standalone LLMs.[6]

Compared to traditional health chatbots that often lack flexibility and deep knowledge [5, 11], our RAG approach provides adaptability and grounding in potentially vast, up-to-date information sources. Unlike RAG systems focused solely on question-answering against static documents [6, 17], our agent incorporates dynamic personalization based on a structured user profile inspired by FHIR standards [22], allowing tailored responses that consider individual health contexts. Furthermore, it integrates explicit safety mechanisms, addressing a critical need often overlooked in general conversational AI.[34] The novelty lies in this specific synergy: using RAG not just for accuracy but as a conduit for personalization by augmenting prompts with both retrieved knowledge and retrieved user data, within a safety-conscious framework.

The strengths of this architecture include potentially enhanced trustworthiness due to verifiable knowledge grounding [6], improved relevance through multi-faceted personalization [23, 24], 24/7 availability for users [77], and proactive safety checks designed for the healthcare domain. The RAG approach offers potential scalability benefits compared to constant LLM retraining.[10, 13]

However, significant challenges remain. The quality of RAG output is highly dependent on the quality of the retrieval process; poor retrieval can lead to irrelevant or incorrect responses, even with a powerful LLM.[6] Managing and updating the knowledge base requires ongoing effort. Latency introduced by the retrieval step can impact user experience.[62] Personalization, while enhanced, is constrained by the available data and the sophistication of the models used to interpret it. Despite RAG providing source traceability, the internal reasoning of the LLM remains largely opaque, highlighting the need for future XAI integration.[49, 60] Perhaps the most significant hurdle is the need for rigorous clinical validation.[38, 74] Demonstrating safety, efficacy, and real-world utility requires extensive testing beyond technical benchmarks.

Ethically, the collection and use of detailed personal health data necessitate robust privacy and security measures.[47, 52] Algorithmic bias, potentially present in the LLM or the knowledge base, must be actively monitored and mitigated. A critical societal consideration is the risk of user over-reliance on the agent, potentially leading them to delay seeking necessary professional care.[3] Therefore, it is crucial to consistently position the agent as a supportive tool that complements, rather than replaces, the expertise and judgment of human healthcare professionals.[34, 36] While such tools hold potential to improve health literacy and access to information [3, 77, 78], ensuring equitable access and avoiding the exacerbation of health disparities due to digital divides is essential.[3]

6 CONCLUSION AND FUTURE WORK

We have presented the design and preliminary implementation of an AI Health Agent utilizing Retrieval-Augmented Generation (RAG) to provide personalized, evidence-grounded health management support. The architecture integrates an LLM with external knowledge retrieval, dynamic user profiling based on FHIR concepts, and explicit safety mechanisms. Our work demonstrates the feasibility of this approach and highlights its potential to address key limitations of previous conversational health systems by enhancing trustworthiness, personalization, and safety.

Future work will focus on addressing the identified limitations and advancing the capabilities of the agent through several key directions:-

Robust Evaluation:- The immediate priority is rigorous evaluation. This includes conducting formal usability studies using established metrics (e.g., SUS, task completion rates [74, 79]) and controlled trials to assess clinical efficacy, impact on health behaviors, safety in diverse scenarios, and user engagement across different demographic groups.[10, 38, 74] Applying comprehensive evaluation frameworks tailored for health chatbots is crucial.[75, 80]

Enhanced RAG Techniques:- We plan to explore more advanced RAG strategies to improve retrieval relevance and robustness. This includes investigating iterative retrieval methods where the system can ask clarifying questions or perform multi-step searches [39], implementing hybrid search combining semantic and keyword matching [63], and potentially exploring graph-based RAG for navigating structured knowledge.[41, 64] Optimizing document chunking and indexing strategies specifically for medical text is also needed.[6]

Deeper Personalization and Integration:- Future iterations will aim for deeper personalization by developing models that capture complex interactions between health factors.[46] A major goal is to enable secure, real-time integration with EHR systems using standards like SMART on FHIR [21, 67], allowing the agent to access a richer, clinically validated dataset.[6] Integration with data streams from wearable devices will also be explored.[24]

Explainability (XAI):- To improve transparency and build trust, we will investigate the integration of XAI techniques. This could involve using methods like LIME or SHAP to explain the influence of specific user profile features or retrieved knowledge on the generated recommendations, or prompting the LLM to generate natural language rationales for its advice.[49, 53, 57]

Longitudinal and Proactive Support:- We aim to extend the agent's capabilities beyond reactive QA to provide proactive support for long-term health management. This includes features for goal setting, progress tracking, personalized reminders, and adaptive behavior change interventions based on longitudinal data analysis.[10, 29]

Agentic Capabilities:- Exploring the integration of the RAG system within a broader LLM-powered agent framework could enable more complex reasoning, multi-step planning, and the ability to interact with other tools or APIs (e.g., scheduling systems, external health calculators) within the healthcare context.[4, 6, 26]

Safety Refinement:- Safety protocols, particularly emergency detection [31], will be continuously refined based on evaluation data and evolving best practices in AI safety for healthcare.

By pursuing these directions, we aim to develop AI health agents that can responsibly and effectively augment traditional healthcare, empowering individuals in their day-to-day health management.

REFERENCES

[1] Emily Alsentzer, John R. Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (2019), 72–78. https://aclanthology.org/W19-5009/

[2] David W. Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs* 33, 7 (2014), 1123–1131. https://doi.org/10.1377/hlthaff.2014.0041

[3] Jordan Braun, Jacques Botha, Brenda Hernandez, William R. Hogan, and Mathias Brochhausen. 2022. Data modeling for clinical research using the HL7 FHIR standard: A case report from the Mobilize Center. *Journal of Biomedical Informatics* 125 (2022), 103969. https://doi.org/10.1016/j.jbi.2021.103969

[4] John Brooke. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7. Classic reference on usability evaluation.

[5] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832. https://doi.org/10.3390/electronics8080832

[6] Houman Esmaeilzadeh. 2024. Graph RAG: A Survey. *arXiv preprint arXiv:2401.13108* (2024). https://arxiv.org/abs/2401.13108

[7] A. Fiske, B. Prainsack, and A. Buyx. 2020. Meeting the challenges of artificial intelligence in medicine. *Zeitschrift für Ethik und Moralphilosophie* 3, 1 (2020), 39–51. https://doi.org/10.1007/s42048-019-00052-2

[8] Valerio Izzo, Ana Popovic, Kaveh Hassani, and Boi Faltings. 2021. A Survey on Vector Databases. *arXiv preprint arXiv:2101.00410* (2021). https://arxiv.org/abs/2101.00410

[9] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

[10] T. Kowatsch, S. Harperink, U. Dittler, G. Xiao, C. Stettler, E. Fleisch, and F. von Wangenheim. 2021. Digital health interventions based on conversational agents or chatbots: A systematic literature review and recommendations for future research. *Journal of Medical Internet Research* 23, 9 (2021), e25513. https://doi.org/10.2196/25513

[11] S. M. Lauritsen, M. Kristensen, M. S. Lehmann, M. S. Larsen, K. M. Lauritsen, M. J. Jørgensen, J. Lange, and B. Thiesson. 2020. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications* 11, 1 (2020), 3852. https://doi.org/10.1038/s41467-020-17431-x

[12] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sun Kim, Chang Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[13] E. P. Lehman, S. Abbasi-Sureshjani, M. Luck, B. van Ginneken, and M. de Bruijne. 2020. Explainable AI for medical applications: Evaluating LIME and SHAP on chest X-rays. *Medical Imaging 2020: Computer-Aided Diagnosis* 11314 (2020), 113140H. https://doi.org/10.1117/12.2549537

[14] James R. Lewis. 1995. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction* 7, 1 (1995), 57–78. https://doi.org/10.1080/10447319509526110

[15] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* 30 (2017). https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[16] Y. J. Oh, J. Zhang, M. L. Fang, and Y. Fukuoka. 2017. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. *International Journal of Behavioral Nutrition and Physical Activity* 14, 1 (2017), 160. https://doi.org/10.1186/s12966-017-0614-1

[17] S. D. Persell, R. M. Cronin, K. A. Fitzner, and G. Elwyn. 2022. A framework for evaluating chatbots in health care. *Journal of Medical Internet Research* 24, 10 (2022), e38790. https://doi.org/10.2196/38790

[18] W. Nicholson Price and I. Glenn Cohen. 2019. Privacy in the age of medical big data. *Nature Medicine* 25, 1 (2019), 37–43. https://doi.org/10.1038/s41591-018-0272-7

[19] S. Provoost, H. M. Lau, J. Ruwaard, and H. Riper. 2017. Embodied Conversational Agents in Clinical Psychology: A Scoping Review. *Journal of Medical Internet Research* 19, 5 (2017), e151. https://doi.org/10.2196/jmir.6553

[20] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 1135–1144. https://doi.org/10.1145/2939672.2939778

[21] Stephen M. Schueller, Xiaoqian Luo, B. Kaveladze, M. L. Dobias, and J. Torous. 2021. Use of mental health apps and chatbots: Applying the technology acceptance model. *JMIR Mental Health* 8, 11 (2021), e27108. https://doi.org/10.2196/27108

[22] Mariya Toneva and Dina Demner-Fushman. 2022. Explainable AI for Medical Text: Challenges and Opportunities. *Frontiers in Artificial Intelligence* 5 (2022), 891779. https://doi.org/10.3389/frai.2022.891779

[23] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288* (2023). https://arxiv.org/abs/2307.09288

[24] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous. 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry* 64, 7 (2019), 456–464. https://doi.org/10.1177/0706743719828977

[25] Giulia Vilone and Luca Longo. 2021. Explainable Artificial Intelligence: a Systematic Review. *arXiv preprint arXiv:2006.00093* (2021). https://arxiv.org/abs/2006.00093

[26] L. Yang, Y. Zhao, T. Sun, X. Li, and H. Lin. 2022. A survey on deep learning for textual crisis management. *Expert Systems with Applications* 190 (2022), 116162. https://doi.org/10.1016/j.eswa.2021.116162

[27] Jin Zhan, Suchin Srinivas, and Laks V. S. Lakshmanan. 2023. Keyword Search on Graphs: A Survey. *Proceedings of the VLDB Endowment* 16, 12 (2023), 4031–4034. https://doi.org/10.14778/3611540.3611598

[28] Y. Zhang, K. Chen, J. Poon, and S. K. Poon. 2020. Detecting emergency situations from social media messages using deep learning models. *IEEE Access* 8 (2020), 198689–198700. https://doi.org/10.1109/ACCESS.2020.3034899