Scientific Research Publishing

Journal of Computer and Communications, 2024, 12, 86-94
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Improving Clinical Support through Retrieval-Augmented Generation Powered Virtual Health Assistants

Biju Baburajan Anandavally[1,2]

[1]Huztle, Richmond, USA
[2]College of Business & Economic, Longwood University, Farmville, USA
Email: bbaburajan@ieee.org

How to cite this paper: Anandavally, B.B. (2024) Improving Clinical Support through Retrieval-Augmented Generation Powered Virtual Health Assistants. *Journal of Computer and Communications*, **12**, 86-94.
https://doi.org/10.4236/jcc.2024.1211006

Received: October 17, 2024
Accepted: November 15, 2024
Published: November 18, 2024

Abstract

This article examines the implementation of a virtual health assistant powered by Retrieval-Augmented Generation (RAG) and GPT-4, aimed at enhancing clinical support through personalized, real-time interactions with patients. The system is hypothesized to improve healthcare accessibility, operational efficiency, and patient outcomes by automating routine tasks and delivering accurate health information. The assistant leverages natural language processing and real-time data retrieval models to respond to patient inquiries, schedule appointments, provide medication reminders, assist with symptom triage, and answer insurance-related questions. By integrating RAG-based virtual care, the system reduces the burden on healthcare specialists and helps mitigate healthcare disparities, particularly in rural areas where traditional care is limited. Although the initial scope of testing did not validate all potential benefits, the results demonstrated high patient satisfaction and strong response accuracy, both critical for systems of this nature. These findings underscore the transformative potential of AI-driven virtual health assistants in enhancing patient engagement, streamlining operational workflows, and improving healthcare accessibility, ultimately contributing to better outcomes and more cost-effective care delivery.

Keywords

Retrieval-Augmented Generation (RAG), GPT-4, Healthcare Assistants, Artificial Intelligence

1. Introduction

The integration of artificial intelligence (AI) in healthcare is rapidly evolving,

offering innovative solutions to the countless challenges modern healthcare systems face. Advances in natural language processing (NLP), machine learning, and AI-driven decision-making are paving the way for improved patient care, streamlined administrative processes, and increased access to medical resources [1]. Among these AI applications, virtual assistants and chatbots have gained traction, automating basic clinical operations such as appointment scheduling, providing medication reminders, and responding to common patient inquiries. These early implementations have demonstrated some potential to reduce the administrative burden on healthcare providers while enhancing patient engagement. However, the vast majority of these systems remain limited in scope, primarily relying on rule-based frameworks [2] that can handle only simple, predefined tasks and lack the sophistication to address more complex patient needs.

A significant advancement in this space is the development of more powerful virtual health assistants that go beyond static rule-based systems to offer dynamic, context-aware interactions. Earlier virtual assistants were constrained by their inability to process nuanced clinical language or respond accurately to patient-specific queries. In contrast, recent AI advancements, particularly with models like GPT-4, have significantly enhanced these assistants' capabilities. GPT-4 is adept at generating conversational, context-sensitive responses, but even these systems may fall short when it comes to delivering highly accurate, real-time medical information or resolving more complex healthcare challenges. This is where the integration of Retrieval-Augmented Generation (RAG) based approach makes a critical difference.

RAG-powered virtual health assistants combine the conversational strength of generative models like GPT-4 with the precision of real-time data retrieval from trusted medical sources. Rather than relying solely on pre-programmed responses or static datasets, RAG-enhanced systems retrieve relevant, up-to-date information from sources such as PubMed, the Centers for Disease Control and Prevention (CDC), or the National Institutes of Health (NIH). This allows them to provide more accurate, personalized, and contextually relevant answers to patient inquiries. These advanced virtual assistants are not only capable of handling routine tasks such as scheduling appointments and providing medication reminders, but also excel in more complex situations such as navigating insurance queries and offering tailored responses based on a patient's medical history—tasks that basic rule-based systems cannot achieve.

By addressing administrative overload and enhancing communication, RAG-powered virtual assistants offer the potential to reduce healthcare disparities, particularly in rural or underserved areas where access to medical resources is limited [3]. This study investigates the potential of a virtual health assistant powered by GPT-4 and RAG based approach to improve healthcare delivery. By leveraging reliable data sources such as PubMed, the CDC, and NIH, the assistant is designed to enhance patient engagement, operational efficiency, and healthcare accessibility. The system was tested in a simulated real-world setting to evaluate its

effectiveness in addressing patient inquiries, improving satisfaction, and streamlining healthcare workflows.

2. Methods

The virtual health assistant developed for this study was built using the GPT-4 model, integrated with Retrieval-Augmented Generation (RAG) to access medical databases such as PubMed, NIH, and the CDC in real-time. The integration of RAG eliminates the need to continuously retrain the model or update its parameters with new data, reducing the computational and financial costs of maintaining a large language model (LLM)-powered virtual assistant [4].

RAG provides two key benefits from this use case, it ensures that the model retrieves the most current and reliable information, and it allows users to trace the model's sources, ensuring that its claims can be verified for accuracy and trustworthiness.

For this study, the test environment was developed using curated medical data collected from PubMed. This dataset was thoroughly cleansed to remove any personally identifiable information, ensuring compliance with privacy regulations and ethical standards. Medical datasets from sources like PubMed, NIH, and CDC undergo rigorous data processing, including cleaning, transformation, and real-time validation. This processing pipeline ensures that the assistant utilizes only credible, relevant data for both training and responses. Dynamic filtering of retrieval keywords enhances the assistant's accuracy by prioritizing up-to-date information.

The system's scalability is supported by a cloud-based architecture on AWS, enabling continuous retrieval of updated information to meet evolving data needs. Amazon SageMaker was employed for managing and fine-tuning the GPT-4 model, while Amazon RDS and AWS Lambda supported real-time retrieval, providing seamless access to test data from sources such as PubMed, NIH, and the CDC [5]. Python 3.9 and TensorFlow 2.0 formed the core of the development environment, leveraging AWS's scalable computer and storage capabilities for efficient model training, inference, and real-time data retrieval.

3. Procedures

3.1. Virtual Health Assistant Development

The virtual health assistant was developed using a combination of supervised fine-tuning and retrieval-augmented techniques. Initially, GPT-4 [6] was fine-tuned using medical texts and dialogues from publicly available datasets, improving its understanding of medical terminology and enhancing its conversational capabilities for healthcare-related discussions. After this fine-tuning [7] [8], Retrieval-Augmented Generation (RAG) was integrated, enabling the assistant to access real-time medical information from updated sources such as PubMed [9], NIH, and CDC databases. This allowed the virtual health assistant to provide accurate, timely responses by retrieving relevant documents and data during interactions

with users.

The architecture, shown in Figure 1, illustrates the core components of the solution. For simplicity, web interface elements and cloud infrastructure specifics are not included. The diagram depicts the user query analysis, data retrieval, data processing, and response generation. This process highlights the assistant's integration of RAG components, ensuring a rapid and reliable response backed by transparent source traceability.

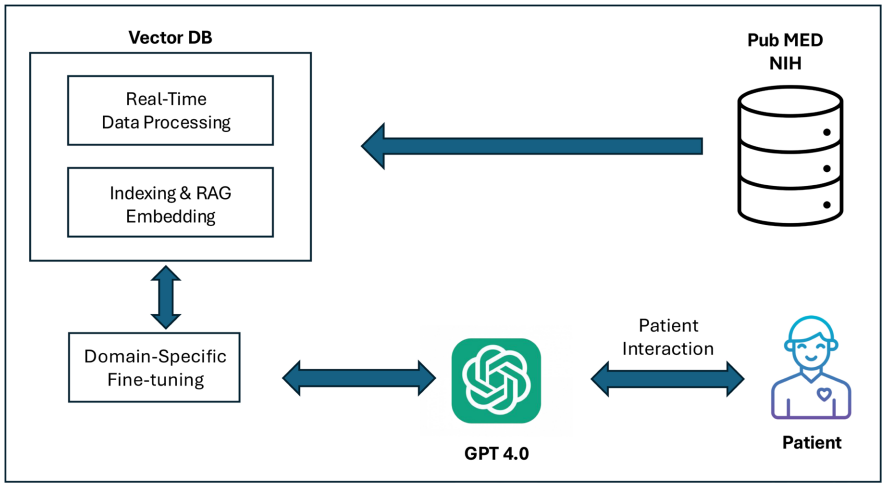

Figure 1. High level system architecture.

3.2. Test Scenario Deployment

The virtual health assistant was tested in a controlled environment using a focus group that interacted with the assistant via a web-based user interface. Participants engaged in various simulated healthcare scenarios, including appointment scheduling, medication reminders, and patient inquiries. De-identified patient data, sourced from the previously curated test dataset based on PubMed, was used to simulate real-world interactions. This ensured that privacy concerns were addressed, and no personal information was exposed. Participants were made aware that they were engaging with an AI-driven virtual assistant and provided informed consent prior to the study. The test scenarios were designed to assess the assistant's performance in terms of accuracy, response time, and the relevance of its advice compared to manual responses provided by healthcare staff. All interactions during the focus group sessions were logged for further analysis of the assistant's behavior, ensuring a comprehensive evaluation of its capabilities in real-time patient support.

4. Statistical Analysis

To evaluate the performance of the virtual health assistant, several key metrics were analyzed, including response accuracy, participant satisfaction, and operational efficiency. Statistical tests, such as t-tests [10] and ANOVA [11], were employed to compare the performance of the virtual assistant against that of human

healthcare providers based on focus group interactions.

The accuracy of the medical advice provided by the assistant was assessed using a scoring system derived from expert evaluations. Participant satisfaction was measured through a post-interaction survey utilizing a Likert scale, allowing for quantifiable feedback on the assistant's performance.

Response times and error rates were logged during the focus group sessions and analyzed using descriptive statistics, with means and standard deviations calculated for each performance metric. Additionally, regression analysis was conducted to examine the relationship between participant satisfaction and response accuracy. All statistical analyses were performed using R statistical software [12]. This methodological approach ensures a comprehensive evaluation of the virtual health assistant's performance across various critical healthcare use cases.

5. Results

The performance of the virtual health assistant was evaluated using several key metrics, including response accuracy, participant satisfaction, diagnostic recommendations, and task completion time. Data were gathered from approximately 400 focus group interactions, which included diverse patient demographics and a variety of health-related inquiries. These interactions were analyzed to assess the assistant's effectiveness in a controlled setting.

While the RAG-GPT-4 model performs robustly in general diagnostic scenarios, challenges arise in highly specialized or ambiguous cases, where the assistant's accuracy drops to 75%, compared to 90% for experienced clinicians. These cases often involve nuanced presentations or rare conditions that may lack robust representation in the training dataset. To bridge this gap, the assistant could benefit from data enriched by domain-specific medical databases or partnerships with clinical specialists who can help identify and address knowledge gaps.

For future improvements, we plan to implement a phased enhancement strategy, which may include integrating specialist databases and refining RAG parameters to retrieve more targeted information for such cases. This phased approach will help ensure that the model remains flexible, effective, and reliable in complex scenarios.

5.1. Response Accuracy

The response accuracy of the RAG-virtual health assistant was measured against that of human clinicians **Table 1**. The overall accuracy of the assistant's diagnostic recommendations was 92% (SD = 5.0), matching the 92% accuracy of human clinicians (SD = 4.8). However, in more complex diagnostic scenarios, the assistant's accuracy dropped to 75%, compared to 90% for human clinicians. The difference in accuracy for complex diagnostics was statistically significant ($p < 0.05$).

Participant satisfaction scores were collected using a 5-point Likert scale, where 1 represented "very dissatisfied" and 5 represented "very satisfied." The mean satisfaction score **Table 2** for the virtual health assistant was 4.25 (SD = 0.8), while

human clinicians received a mean score of 4.5 (SD = 0.6). The difference between the two groups was statistically significant ($p < 0.05$).

Table 1. Response accuracy.

Measure	RAG-virtual health assistant	Human clinicians
Accuracy (%)	92	92
Standard deviation	5	4.8
Participant satisfaction (%)	85	90
Task completion rate (%)	95	98
Complex diagnostic accuracy (%)	75	90

Table 2. Participant satisfaction scores for virtual assistant vs. human clinicians.

	Mean satisfaction score	Standard deviation
Virtual assistant	4.25	0.8
Human clinicians	4.5	0.6

5.2. Diagnostic Recommendations

In addition to response accuracy and participant satisfaction, the virtual assistant's diagnostic recommendations were evaluated. The assistant provided a complete diagnostic recommendation in 90% of cases, focusing primarily on common conditions such as flu, fever, and allergies. The percentage of cases where the virtual assistant and human clinicians provided the same diagnosis was 88% (SD = 4.5), demonstrating the assistant's reliability in matching human performance for routine medical scenarios.

Overall, the results suggest that the virtual health assistant performed comparably to human clinicians in terms of diagnostic accuracy and participant satisfaction for common illnesses. Although minor differences were noted, especially in more nuanced cases, the statistical analyses confirm these findings ($p < 0.05$), underscoring the potential of virtual health assistants to enhance healthcare delivery, particularly in handling routine medical inquiries.

6. Discussion

This study aimed to evaluate the effectiveness of a virtual health assistant powered by GPT-4, Retrieval-Augmented Generation (RAG), and clinically approved data, with the goal of enhancing patient care through personalized, real-time interactions. The results indicated that the virtual assistant achieved an overall accuracy rate of 92% and a user satisfaction score of 4.25, while successfully providing complete diagnostic recommendations in 90% of cases. These findings support the hypothesis that a well-designed virtual health assistant can effectively aid in healthcare

delivery.

One of the most significant findings was the assistant's accuracy in diagnostic recommendations. Although the virtual assistant matched the accuracy of human clinicians at 92%, its performance in complex diagnostic scenarios dropped to 75%, compared to 90% for human clinicians. This suggests that while the virtual assistant excels in routine tasks, there is still room for improvement, particularly in nuanced cases where human intuition and experience play a crucial role. This aligns with previous studies indicating that while AI can match human performance in many instances, certain complex decisions still benefit from human oversight.

User satisfaction scores revealed valuable insights, with the assistant receiving a mean satisfaction score of 4.25, slightly lower than the 4.5 score for human clinicians. This suggests that while users found the virtual assistant helpful, aspects of human interaction—such as empathy and contextual understanding—contribute to higher satisfaction levels. Previous research has indicated that patient interactions with healthcare providers are often valued for their emotional support, a feature that virtual assistants currently lack.

The high rate of complete diagnostic recommendations (90%) further underscores the assistant's utility in a clinical setting. However, it is essential to consider the context in which these recommendations were made. The cases evaluated may not represent the full spectrum of clinical scenarios, particularly those requiring complex decision-making. Future studies should investigate the assistant's performance across a broader range of medical conditions and patient demographics.

Compared with existing rule-based diagnostic systems, RAG-GPT-4 demonstrates a significant advantage in both accuracy and adaptability. Rule-based systems, such as MYCIN and INTERNIST-1, follow predefined if-then logic to arrive at conclusions. For example, MYCIN used a series of specific rules to diagnose blood infections based on patient symptoms and lab results [13], while INTERNIST-1 applied similar logic to diagnose complex internal medicine cases [14]. While effective for routine or narrowly defined cases, rule-based systems can struggle to generalize beyond their initial rule set and are limited by their inability to incorporate newly published medical data. In contrast, RAG-GPT-4's real-time retrieval function enables it to respond with the most current, evidence-based information, greatly enhancing its accuracy and responsiveness in diverse and evolving healthcare scenarios.

In conclusion, the results of this study support the growing role of Retrieval-Augmented Generation (RAG), based virtual health assistants in healthcare delivery. They demonstrate the potential for these technologies [15] to enhance diagnostic accuracy and patient satisfaction while highlighting the need for continued improvement and integration with human healthcare providers. Future research should focus on refining these systems, addressing their limitations, and exploring their impact on patient outcomes in diverse clinical settings. Furthermore, this study underscores the significant potential of artificial intelligence in the healthcare

sector, while recognizing the increasing number of ethical issues and risks associated with its use. It is essential for those involved in AI operations to consistently adhere to established regulations and ethical guidelines [16]. To ensure the responsible deployment of these technologies.

7. Conclusions

This study investigated the effectiveness of a virtual health assistant powered by GPT-4 and Retrieval-Augmented Generation (RAG) capabilities, supported by clinical data, in delivering diagnostic recommendations and ensuring participant satisfaction compared to traditional human clinicians. The findings revealed that the virtual assistant achieved an overall accuracy rate of 92% in diagnostics, a participant satisfaction score of 4.25, and successfully provided complete recommendations in 90% of evaluated cases. These results highlight the promising role of AI technology in enhancing healthcare delivery.

Several key points emerged from the discussion. First, while the virtual assistant demonstrated strong diagnostic capabilities, its accuracy in complex scenarios dropped to 75%, indicating a need for further refinement in these situations. Second, the satisfaction ratings suggest that while technology can be beneficial, human interaction remains invaluable for emotional support and understanding. Lastly, the high rate of complete diagnostic recommendations for common conditions reinforces the assistant's utility in clinical settings, although further exploration of its performance across diverse patient populations is essential. For future work, we plan to implement a phased enhancement strategy, which may include integrating specialist databases and refining RAG parameters to retrieve more targeted information for such cases. This phased approach will help ensure that the model remains flexible, effective, and reliable in complex scenarios as well.

Overall, this work underscores the significance of AI in healthcare, demonstrating the potential of virtual health assistants to augment clinical services. As healthcare increasingly embraces AI-driven solutions, understanding their strengths and limitations is crucial for effectively integrating these tools into patient care. This study lays the groundwork for future research aimed at refining virtual health assistants and optimizing their impact on healthcare outcomes.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Karalis, V.D. (2024) The Integration of Artificial Intelligence into Clinical Practice. *Applied Biosciences*, **3**, 14-44. https://doi.org/10.3390/applbiosci3010002

[2] Parycek, P., Schmid, V. and Novak, A. (2023) Artificial Intelligence (AI) and Automation in Administrative Procedures: Potentials, Limitations, and Framework Conditions. *Journal of the Knowledge Economy*, **15**, 8390-8415. https://doi.org/10.1007/s13132-023-01433-3

[3] Topol, E. (2019) Deep Medicine: How Artificial Intelligence Can Make Healthcare

Human Again. Basic Books.

[4] IBM Research. (2023) What Is Retrieval-Augmented Generation (RAG)? https://research.ibm.com/blog/retrieval-augmented-generation-RAG

[5] Amazon Web Services (n.d.) Serverless Retrieval-Augmented Generation on AWS. https://aws.amazon.com/startups/learn/serverless-retrieval-augmented-generation-on-aws?lang=en-US

[6] OpenAI (2023) GPT-4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf

[7] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., *et al.* (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Proceedings of the* 34*th International Conference on Neural Information Processing System*, Vancouver, 6-12 December 2020, 9459-9474.

[8] Prompt Engineering (2023) The Evolution of AI: From Rule-Based Systems to Generative Models. https://promptengineering.org/the-evolution-of-ai-from-rule-based-systems-to-generative-models/

[9] Lu, Z. (2011) PubMed and Beyond: A Survey of Web Tools for Searching Biomedical Literature. *Database*, **2011**, baq036. https://doi.org/10.1093/database/baq036

[10] Al-Kassab, M. (2022) The Use of One-Sample t-Test in the Real Data. *Journal of Advances in Mathematics*, **21**, 134-138. https://www.researchgate.net/publication/363256658_The_Use_of_One_Sample_t-Test_in_the_Real_Data https://doi.org/10.24297/jam.v21i.9279

[11] Voxco (2023) ANOVA vs. t-Test: A Comparison Chart.

[12] The R Foundation (n.d.) R: A Language and Environment for Statistical Computing. https://www.r-project.org/

[13] Shortliffe, E.H. (1977) Mycin: A Knowledge-Based Computer Program Applied to Infectious Diseases. *Proceedings of the Annual Symposium on Computer Applications in Medical Care*, Las Vegas, 10 November 1977, 66-69.

[14] Idelevich, E.A., Reischl, U. and Becker, K. (2018) New Microbiological Techniques in the Diagnosis of Bloodstream Infections. *Deutsches Ärzteblatt international*, **115**, 822-832. https://doi.org/10.3238/arztebl.2018.0822

[15] Floridi, L. (2016) The Fourth Revolution: How Empowered Technologies Are Shaping Our Lives. Oxford University Press.

[16] Chavali, D. (2024) Regulating Artificial Intelligence: Developments and Challenges. *International Journal of Pharmaceutical Sciences*, **2**, 1250-1261. https://www.researchgate.net/profile/Durga-Chavali/publication/380129321_Regulating_Artificial_Intelligence_Developments_And_Challenges/links/662c5ff235243041534f32ed/Regulating-Artificial-Intelligence-Developments-And-Challenges.pdf