2025 Eighth International Women in Data Science Conference at Prince Sultan University (WiDS PSU) | 979-8-3315-2092-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/WIDS-PSU64963.2025.00041

Multi-Tiered RAG-Based Chatbot for Mental Health Support

Sadia Siddique
Arab Open University (KSA)
21462779KSA@aou.edu.sa

Fatimah Alsayoud
Arab Open University (KSA)
f.alsayoud@arabou.edu.sa

***Abstract*—Mental health challenges are increasing globally, and many are unable to access timely and affordable care due to financial, geographical, and societal barriers. Traditional mental health solutions often fall short in offering tailored support, relying on rigid, one-size-fits-all approaches that fail to address the unique needs of individuals, especially in underserved regions. To address these gaps, a multi-tiered chatbot leveraging Retrieval-Augmented Generation (RAG) provides dynamic, context-aware mental health assistance. Operating in four tiers, it offers general information, support for common issues, specialized guidance for conditions like anxiety and bipolar disorder, and personalized recommendations from user-uploaded documents. The RAG-based approach integrates the latest studies and techniques into its knowledge base, evolving in real-time without costly retraining of large language models (LLMs). This cost-effective system lowers financial barriers while improving access to quality mental health care in underserved regions. By enabling proactive and affordable mental health support, this model presents a sustainable solution to the global mental health crisis.**

***Index Terms*—Evidence-Based Therapy Chatbots, Cognitive Behavioral Therapy (CBT), Personalized Counseling, Mental Health Support, Conversational Agents, RAG, LLMs, Chatbot Counseling, Therapeutic Chatbots, AI for Mental Wellness, Generative AI, Multi-Query Retrieval, Multi-Query Graph RAG (MQG-RAG)**

I. Introduction

The global mental health crisis affects over 1 billion individuals, contributing to 14.3% of all deaths annually, with depression alone impacting over 300 million people [1]. Despite growing awareness, barriers such as stigma, high therapy costs, long waiting times, and a shortage of mental health professionals prevent timely access to care, particularly in underserved regions [2].

AI-driven chatbots provide a scalable and cost-effective alternative to traditional therapy. However, existing models rely on static, predefined responses, limiting their ability to provide personalized, up-to-date interventions. RAG addresses this limitation by combining real-time information retrieval with generative AI, ensuring responses are accurate, dynamic, and context-aware.

In addition to technological benefits, RAG-based chatbots align with cognitive and psychological theories, enhancing their effectiveness in mental health interventions. Cognitive Load Theory suggests that reducing mental effort improves engagement, making RAG's targeted retrieval crucial for users in distress [3]. Additionally, Dual-Process Theory highlights the need to shift users from emotion-driven reactions (System 1) to analytical reasoning (System 2)—a core principle in **Cognitive Behavioral Therapy (CBT)**. Furthermore, according to the Elaboration Likelihood Model (ELM), individuals are more likely to internalize and act upon mental health advice when presented with credible, personalized information [4], positioning **RAG-based retrieval as a powerful persuasive tool** in digital therapeutic interventions.

This research introduces Healio, an AI-driven mental health chatbot leveraging RAG architecture to provide scalable, personalized, and evidence-based support. Unlike traditional chatbots, Healio retrieves real-time information from trusted sources, ensuring responses remain clinically relevant and adaptive.

Healio operates across four tiers: Tier 1 provides general mental health guidance, Tier 2 retrieves evidence-based information, Tier 3 offers condition-specific support (e.g., anxiety, bipolar disorder), and Tier 4 analyzes user-uploaded documents for personalized interventions.

By integrating real-time retrieval with psychological frameworks, Healio offers a dynamic, accessible, and cost-effective mental health support system, addressing critical gaps in traditional care.

II. Related Works

Recent advancements in Retrieval-Augmented Generation (RAG) techniques have shown promise in enhancing AI-driven mental health chatbots. By retrieving external knowledge and integrating it with natural language generation, RAG can improve chatbot accuracy, responsiveness, and personalization. While existing mental health chatbots offer structured interventions, they are often limited by static frameworks. Integrating RAG can address these limitations by enabling real-time access to clinical guidelines, research, and tailored interventions.

Markey et al. (2024) [5] explored the impact of RAG on GPT-4's performance in generating clinical trial documents. Their study found that while standard GPT-4 produced medically relevant content, it lacked clinical reasoning and sometimes contradicted established guidelines. In contrast, the RAG-enhanced GPT-4 incorporated real-time external data, improving accuracy and alignment with medical standards. However, this research did not explore RAG's potential in mental health applications, presenting an opportunity to extend its use in AI-driven mental health support.

979-8-3315-2092-2/25/$31.00 ©2025 IEEE
DOI 10.1109/WiDS-PSU64963.2025.00041

Shah (2024) [6] introduced a user-centered mental health platform that personalizes support based on user profiles. The chatbot provides 24/7 accessibility, quick responses, and anonymity while analyzing user-provided data to identify mental health patterns. However, it does not utilize machine learning for real-time adaptation. Incorporating RAG could enhance the system by enabling dynamic retrieval of clinical guidelines and recent research, improving evidence-based interventions.

Kaif et al. (2024) introduced the Gemini MultiPDF Chatbot, which leverages RAG for document retrieval and question-answering. By embedding document content into a vector store using FAISS indexing and Langchain, the system efficiently retrieves relevant information. The research highlights potential applications in mental health, where chatbots could dynamically access clinical resources. However, this approach requires manual document uploads. Extending RAG to integrate external medical databases could eliminate this dependency, ensuring automated and up-to-date responses.

A. Existing Mental Health Chatbots: Woebot, Wysa, and Replika

The increasing demand for AI-driven mental health solutions has led to the development of chatbots such as **Woebot, Wysa, and Replika** [7]. While effective, these systems rely on predefined frameworks that limit adaptability to emerging research and personalized interventions.

Woebot applies CBT through structured dialogues, helping users reframe negative thought patterns. While clinical trials support its effectiveness in reducing depressive symptoms, its static nature restricts responsiveness to evolving mental health research.

Wysa offers a broader approach by integrating CBT, mindfulness, and motivational interviewing techniques. Research indicates that Wysa effectively aids stress and anxiety management . However, like Woebot, it depends on predefined conversational structures rather than real-time data retrieval.

Replika is designed for open-ended emotional companionship rather than structured mental health interventions. While users report benefits in emotional expression, studies highlight concerns about AI dependency and reinforcement of unhealthy attachment behaviors [8].

Unlike these existing models, **Healio** dynamically retrieves and integrates current clinical research, ensuring up-to-date and personalized recommendations. This adaptability allows Healio to address a wider range of mental health conditions beyond traditional chatbot architectures. Additionally, Healio prioritizes evidence-based mental health guidance over unstructured companionship, mitigating risks associated with AI dependency.

B. Comparison and Discussion

While Woebot and Wysa provide structured, evidence-based interventions, their static nature limits adaptability to new research. Replika, though engaging, lacks targeted therapeutic frameworks. By utilizing RAG-based retrieval, Healio delivers real-time, adaptive interventions, continuously incorporating the latest research.

From an ethical standpoint, concerns regarding data privacy, AI reliance, and regulatory compliance remain a challenge across all AI-driven chatbots. Healio mitigates hallucination risks by retrieving validated knowledge in real time; however, ensuring strong security measures and privacy safeguards will be essential for user trust and ethical AI deployment.

RAG enhances accuracy and trust in mental health chatbots by retrieving verified psychological resources, mitigating hallucinations common in fine-tuned models [9]. Studies on AI therapy chatbots like Woebot and Wysa show that structured, evidence-based exercises aid self-reflection and cognitive restructuring, improving mental health outcomes. By delivering relevant therapeutic content—such as CBT worksheets—precisely when needed, Healio follows just-in-time intervention principles. Additionally, RAG-based personalization fosters engagement by tailoring responses to individual user contexts, addressing the disengagement issues seen in static chatbot models.

C. Conclusion

RAG-based AI enhances mental health chatbots by enabling real-time, adaptive interventions. While Woebot and Wysa offer structured CBT-based support and Replika focuses on companionship, Healio dynamically retrieves credible clinical knowledge for personalized assistance. This research explores how RAG improves chatbot accuracy, adaptability, and personalization, with the next section detailing its implementation.

Building on these insights, this research will explore how RAG can optimize mental health chatbots by improving accuracy, real-time adaptability, and personalized interventions. The next section outlines the methodology for implementing these advancements.

III. Framework

Healio is developed using the RAG framework, which offers several advantages over traditional approaches like fine-tuning. The chatbot is designed across four tiers to provide progressively more personalized and specific mental health support.

Tier 1: Healio functions as a general chatbot, similar to a standard LLM, providing responses without specialized mental health knowledge. At this level, GPT-3.5-Turbo generates clear and coherent answers across a broad range of topics.

Tier 2: The second level incorporates documents from credible and reliable sources related to mental health issues, coping strategies, treatments covering topics like schizophrenia, stress, PTSD, suicidal tendencies etc, offering up-to-date and evidence-based information. By leveraging RAG, Healio can retrieve information from trusted mental health sources, ensuring the advice is both relevant and novel, especially when addressing newer techniques or interventions in the field.

Tier 3: At this level, the Healio chatbot is specifically tailored to address mental health disorders like anxiety and

bipolar disorder. By leveraging RAG, Healio can access a specialized knowledge base, which includes documents stored in the Chroma vector store. These documents cover crucial topics such as rapid cycling, mania, hypomania, mood stabilizers, scopophobia, neurosis, and various types of anxiety disorders. Healio provides targeted advice and interventions based on the latest research and clinical guidelines, offering personalized treatment suggestions, coping techniques, and information on effective therapies.

Tier 4: The fourth level enables users to upload their own documents, such as medical reports or therapy transcripts. Healio can then fetch relevant information, summarize the contents, and answer user queries based on this personalized data. This allows the chatbot to provide highly tailored responses, offering insights from medical reports and therapy sessions, ensuring that the advice and support are specifically aligned with the user's unique mental health journey.

A. *Why RAG Over Fine-Tuning*

Traditional fine-tuning is computationally expensive, time-consuming, and prone to obsolescence as new research emerges [10]. In contrast, Retrieval-Augmented Generation (RAG) updates its external knowledge base and generates new embeddings without retraining the entire model. This approach not only ensures responses remain accurate and up-to-date but also minimizes hallucinations by transparently indicating when sufficient information is lacking. As a result, RAG offers a more resource-efficient and reliable method for integrating emerging research compared to conventional fine-tuning.

B. *Comparison with Knowledge-Augmented Generation (KAG) and Reinforcement Learning*

RAG was chosen as the foundation for Healio due to its flexibility, adaptability, and ability to access real-time information, making it particularly well-suited for addressing open-ended and evolving mental health discussions. Unlike Knowledge-Augmented Generation (KAG), which relies on predefined knowledge graphs, RAG allows the system to retrieve and integrate the latest psychological research, therapeutic methodologies, and coping strategies without requiring frequent manual updates. This is particularly important in mental health applications, where treatment approaches are continuously evolving and user queries tend to be highly individualized and unstructured.

Although KAG offers advantages in precision and structured reasoning, making it beneficial for scenarios that demand clinical accuracy and multi-step decision-making—RAG is better suited for delivering dynamic, context-aware, and scalable mental health support. Studies suggest that RAG-based models improve diagnostic and therapeutic capabilities in AI-driven healthcare systems [11].

Moreover, while Reinforcement Learning (RL) has been widely applied in AI, it presents challenges such as high computational demands, slow learning rates, and difficulty in dynamically incorporating new knowledge. To address these limitations, Goyal et al. [12] propose Retrieval-Augmented Reinforcement Learning (R2A), which enables RL agents to retrieve and leverage prior experiences from external datasets. This approach enhances decision-making efficiency and overall performance. Empirical evaluations across multiple environments, including gaming tasks and AI-driven simulations, demonstrate that R2A significantly outperforms conventional RL techniques, reinforcing the benefits of retrieval-augmented methodologies.

IV. Experiment

A. *Naïve Rag*

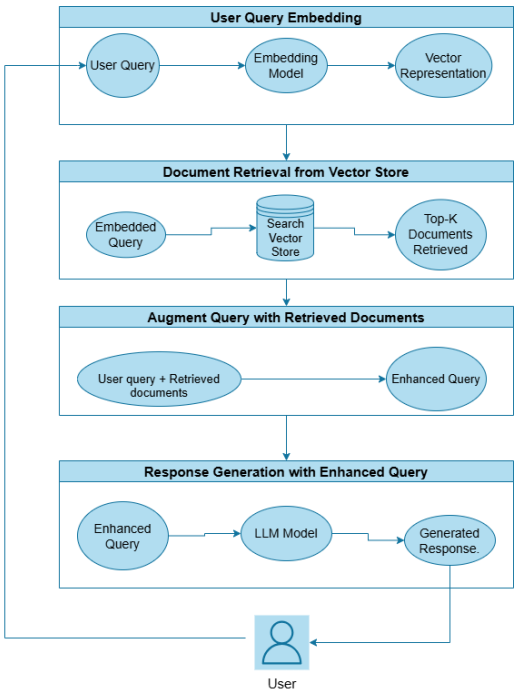

Fig. 1: RAG Pipeline for Healio: The diagram shows the process of converting a user query into a response, including query embedding, document retrieval, query augmentation, and response generation.

To evaluate the effectiveness of a RAG-based chatbot in delivering accurate, context-aware, and empathetic mental health support, a series of experiments were conducted focusing on document retrieval, embedding techniques, and response generation to ensure alignment with evidence-based mental health guidance.

*1) **Data Preparation and Document Embedding**:* The RAG pipeline begins by processing mental health documents as a knowledge base. PDFs are loaded using `PyPDFLoader` from Langchain and split into semantically meaningful chunks with the `RecursiveCharacterTextSplitter` to facilitate efficient retrieval. Each chunk is embedded using `OpenAIEmbeddings` (model: `text-embedding-ada-002`) and stored in a persistent Chroma vector database.

*2) **User Query Embedding and Retrieval**:* When a user submits a query, it is embedded using the same model, ensuring that both query and document embeddings share the same vector space. The Chroma vector store then performs a cosine similarity search, retrieving the top k most relevant document chunks. This ensures that responses are based on contextually accurate and pertinent information.

*3) **Context Augmentation and Conversation History**:* The retrieved documents are combined with the user query and conversation history. This augmentation guarantees that the chatbot provides contextually relevant, coherent, and empathetic responses while maintaining continuity across conversation turns.

*4) **Response Generation**:* The final response is generated using the pre-trained `ChatOpenAI` model (GPT-3.5-Turbo) via a `ChatPromptTemplate`. The system prompt guides the model to deliver empathetic, informative responses based on CBT principles, ensuring cultural sensitivity and respect. The chatbot's answer is then displayed to the user, with the option to view the ranked list of relevant documents for transparency.

B. Multi-Query Graph-RAG (MQG-RAG)

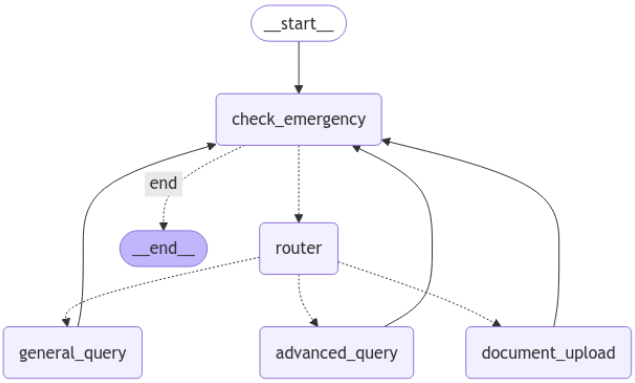

Fig. 2: Architecture of Multi-Query Graph-RAG (MQG-RAG).

This enhanced RAG pipeline integrates **LangGraph**, a flow-based decision-making framework, to improve retrieval accuracy and user safety. Key features include:

- **Structured Query Routing:** Directs user queries to one of three retrieval paths (general, advanced, or document-based) for improved precision. *Emergency checks are performed prior to routing to any level.*
- **Multi-Level Vector Store:** Organizes information hierarchically, ensuring queries are processed at the most relevant level.
- **Crisis Detection:** The system prioritizes user safety by detecting emergencies, such as suicide risk, and providing immediate contact information. Before processing each query, it assesses for crisis indicators and continues engaging the user with supportive prompts to explore their feelings and concerns. This process continues until the user confirms they are safe or expresses willingness to seek professional help.
- **Memory Management:** Utilizes *ExtendedState* to maintain conversation continuity and preserve multi-turn context.
- **MultiQueryRetriever:** Generates alternative query variations to enhance semantic matching and reduce retrieval bias.

Overall, this approach streamlines the retrieval process, enhances response accuracy, and ensures that urgent situations are appropriately handled, contributing to a more effective mental health support chatbot.

V. Model Performance and Evaluation

A. Naïve RAG

1) Evaluation Framework: To assess the performance of the Healio chatbot, the following three key metrics were used: **Role Adherence, Answer Relevancy, and Faithfulness**. These metrics evaluate the chatbot's ability to function as a mental health assistant by ensuring responses are contextually relevant, truthful, and empathetic.

a) Role Adherence Metric: Role Adherence measures whether the chatbot remains within its designated role as a mental health assistant, providing only therapeutic support without deviating into unrelated topics. It is calculated as:

$$\text{Role Adherence} = \frac{\text{Exchanges in Assigned Role}}{\text{Total Exchanges}}$$

b) Answer Relevancy Metric: Answer Relevancy evaluates how well the chatbot's responses align with user queries. This ensures the chatbot remains contextually appropriate and useful. It is computed as:

$$\text{Answer Relevancy} = \frac{\text{Number of Relevant Statements}}{\text{Total Number of Statements}}$$

c) Faithfulness Metric: Faithfulness assesses the chatbot's ability to generate responses that are consistent with retrieved knowledge, preventing hallucinations. It is defined as:

$$\text{Faithfulness} = \frac{\text{Number of Truthful Claims}}{\text{Total Number of Claims Made}}$$

A higher score in all three metrics indicates stronger alignment with intended chatbot behavior. The **DeepEval framework** was used to measure these metrics across multiple test cases, ensuring a comprehensive evaluation.

2) Testing Methodology: Healio was tested using **seven predefined test cases**, each representing a real-world mental health query. User inputs were generated using GPT-3.5-Turbo and processed through the RAG pipeline, retrieving the top three most relevant contexts. The chatbot's responses were evaluated using **Role Adherence, Answer Relevancy, and Faithfulness**.

For each test case: - The chatbot was queried **10 times**, and the average score was recorded. - These averages were aggregated across all test cases to determine the overall metric score. - The process was repeated across **four tiers**, ensuring robust evaluation.

TABLE I: Baseline Questions for Chatbot Evaluation

Question ID	User Query
Q1	I've been having thoughts of not wanting to live anymore. What should I do?
Q2	I've been having flashbacks and nightmares about a traumatic event. How can I cope?
Q3	I feel anxious in social situations and avoid them. Any advice?
Q4	I just had a baby and feel overwhelmed. Is this normal?
Q5	I've been feeling down and unmotivated. How can I shift my mindset?
Q6	I feel isolated and disconnected from friends and family. How can I reconnect?
Q7	My job is highly demanding, and I feel completely burned out. What can I do?

TABLE II: Evaluation Results Across Four Tiers

Metric	Tier 1	Tier 2	Tier 3	Tier 4
Role Adherence	1.0000	1.0000	1.0000	1.0000
Answer Relevancy	0.9691	0.9912	0.9954	0.8451
Faithfulness	0.9976	0.9158	0.9201	0.9031

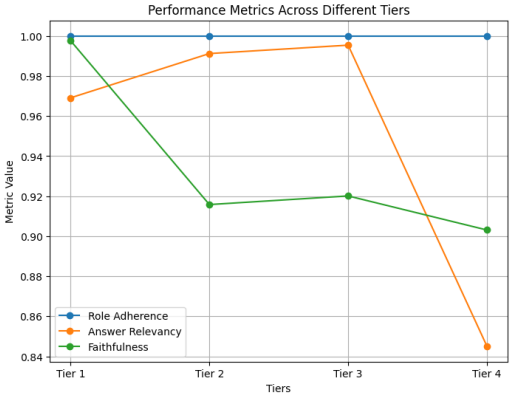

Fig. 3: Evaluation results across four tiers, illustrating chatbot performance.

3) Evaluation Results:

a) Role Adherence: The chatbot maintained a perfect **Role Adherence** score of **1.0** across all tiers, demonstrating that it consistently aligned with its intended role due to well-structured system prompts.

b) Answer Relevancy: The chatbot performed exceptionally well in **Tiers 1–3** with relevancy scores exceeding **0.95**. However, Tier 4 saw a dip to 0.8451, likely due to its reliance on user-uploaded documents, which may not always contain relevant context.

c) Faithfulness: Faithfulness was highest in **Tier 1** due to reliance on a fixed system prompt, with no risk of inconsistencies. **Tiers 2–4**, which depended on retrieved knowledge, showed slight drops in faithfulness, emphasizing the need for **better alignment between retrieval and response generation**.

4) Specialized Evaluation for Tiers 3 and 4: Tiers 3 and 4 underwent additional evaluation using more targeted queries on mental health conditions. - **Tier 3** was tested with six questions focusing on **anxiety and bipolar disorder**. - **Tier 4** was tested with **hypothetical medical documents** and responses were evaluated for faithfulness.

TABLE III: Performance Metrics for Tier 3 and Tier 4

Tier	Answer Relevancy	Faithfulness	Role Adherence
Tier 3	0.9832	0.9388	1.0000
Tier 4	0.9446	0.9624	0.9778

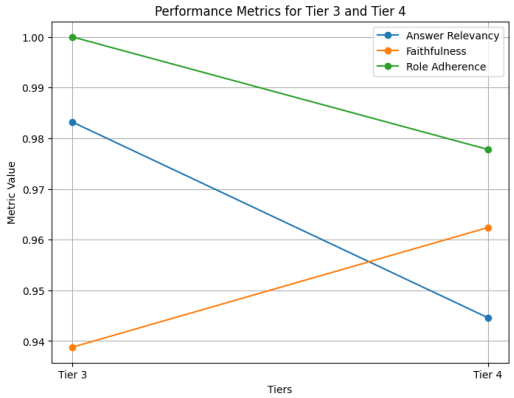

Fig. 4: Comparison of performance metrics for Tier 3 and Tier 4.

B. MQG-RAG Evaluation

MQG-RAG was evaluated using the **RAGAS framework**, focusing on **Contextual Precision, Contextual Recall, Faithfulness, and Answer Relevancy**.

TABLE IV: Performance Metrics for MQG-RAG

Metric	Score
Contextual Precision	0.9929
Contextual Recall	0.9057
Faithfulness	0.88
Answer Relevancy	0.86

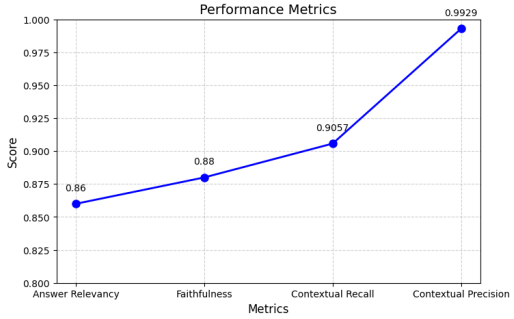

Fig. 5: Architecture of MQG-RAG

C. Conclusion: Comparing Naïve RAG and MQG-RAG

The evaluation of Healio's chatbot across different frameworks reveals the advantages and limitations of both Naïve RAG and MQG-RAG in mental health applications.

a) Performance Comparison: Naïve RAG demonstrated strong **Role Adherence**, maintaining a perfect score across all tiers. However, **Answer Relevancy** and **Faithfulness** showed variability, particularly in Tier 4, where responses relied on user-uploaded documents. This highlights Naïve RAG's sensitivity to context availability, making it less adaptable to complex, real-world scenarios.

In contrast, MQG-RAG, which integrates **LangGraph and Multi-Query Retrieval**, significantly improved **Contextual Precision (0.9929)** and **Recall (0.9057)**. The ability to generate and retrieve multiple queries allowed MQG-RAG to produce more contextually accurate and semantically rich responses. While **Faithfulness (0.88)** and **Answer Relevancy (0.86)** were slightly lower than the highest-scoring Naïve RAG tiers, MQG-RAG's consistency across queries and its ability to leverage structured retrieval mechanisms made it more reliable for dynamic interactions.

b) Key Takeaways:

- **Naïve RAG excels in structured settings** where clear system prompts and well-defined contexts are available, but its performance degrades when retrieving from diverse document sources.
- **MQG-RAG enhances retrieval robustness** through multi-query mechanisms, improving contextual accuracy and reducing dependence on single-query retrieval failures.
- **Naïve RAG is more role-consistent**, whereas **MQG-RAG adapts better to complex, multi-turn interactions**, making it a stronger choice for real-world applications that require dynamic context updates.
- **Both models require further optimization** to improve faithfulness in open-domain retrieval settings.

VI. Real-World Deployment and Integration

For effective deployment, the chatbot can collaborate with *mental health professionals, psychologists, and government agencies* to ensure clinically validated responses and compliance with mental health regulations. Partnering with *hospitals, counseling centers, and mental health organizations* enables real-world testing, allowing professionals to monitor performance and refine interventions. Security and privacy measures, including *HIPAA and GDPR compliance*, data encryption, and crisis escalation protocols, must be implemented to protect user confidentiality. Continuous monitoring, ethical AI development, and regular updates based on clinical research will ensure the chatbot remains reliable, effective, and aligned with best practices in digital mental health care.

VII. Conclusion

This research introduces a Multi-Tiered RAG approach for mental health support, enabling the Healio chatbot to provide scalable, context-aware, and personalized assistance. By leveraging retrieval-based methods, Healio ensures accuracy, real-time information access, and individualized interventions. The tiered framework enhances flexibility, allowing for general advice, evidence-based insights, specialized mental health guidance, and user-specific document processing.

Evaluation results demonstrate that Healio maintains strong role adherence, delivers relevant and accurate responses, and upholds faithfulness to retrieved information. The findings underscore RAG's superiority over traditional fine-tuning by enabling continuous integration of new research, ensuring dynamic and reliable support.

This work establishes a foundation for more advanced, empathetic, and accessible mental health chatbots. Future research can refine this framework by incorporating enhanced personalization and adaptive features to further improve mental health support systems.

References

[1] P. Chodavadia, I. Teo, D. Poremski, D. S. S. Fung, and E. A. Finkelstein, "Prevalence and economic burden of depression and anxiety symptoms among singaporean adults: results from a 2022 web panel," *BMC Psychiatry*, vol. 23, no. 1, p. 104, 2023, published February 14, 2023.

[2] A. A. Ahad, M. Sanchez-Gonzalez, and P. Junquera, "Understanding and addressing mental health stigma across cultures for improving psychiatric care: A narrative review," *Cureus*, vol. 15, no. 5, p. e39549, 2023.

[3] J. Schmidhuber, S. Schlogl, and C. Ploder, "Cognitive load and productivity implications in human-chatbot interaction," in *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*. IEEE, Sep. 2021, pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/ICHMS53169.2021.9582445

[4] Y. K. Dwivedi, J. Balakrishnan, A. M. Baabdullah, and R. Das, "Do chatbots establish "humanness" in the customer purchase journey? an investigation through explanatory sequential design," *Psychology & Marketing*, Aug. 2023. [Online]. Available: https://doi.org/10.1002/mar.21888

[5] N. Markey, I. El-Mansouri, G. Rensonnet, C. van Langen, and C. Meier, "From rags to riches: Using large language models to write documents for clinical trials," 2024. [Online]. Available: https://arxiv.org/abs/2402.16406

[6] S. Shah, "Navigating wellness: Chatbot-powered solutions for mental health," *International Journal for Research in Applied Science and Engineering Technology*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271140903

[7] M. D. R. Haque and S. Rubya, "An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews," *JMIR Mhealth Uhealth*, vol. 11, p. e44838, May 2023.

[8] I. A. of Computer Psychology, "Emotional dependence on ai: A psychological perspective on replika and ai companions," 2025.

[9] R. Shusterman, A. C. Waters, S. O'Neill, M. Bangs, P. Luu, and D. M. Tucker, "An active inference strategy for prompting reliable responses from large language models in medical practice," *NPJ Digital Medicine*, vol. 8, no. 1, p. 119, Feb. 2025.

[10] O. Ovadia, M. Brief, M. Mishaeli, and O. Elisha, "Fine-tuning or retrieval? comparing knowledge injection in llms," in *Conference on Empirical Methods in Natural Language Processing*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266162497

[11] IEEE, "Advancements in ai for healthcare diagnostics and therapeutics," *IEEE Transactions on Artificial Intelligence*, 2023.

[12] A. Goyal, A. L. Friesen, A. Banino, T. Weber, N. R. Ke, A. P. Badia, A. Guez, M. Mirza, P. C. Humphreys, K. Konyushkova, L. Sifre, M. Valko, S. Osindero, T. Lillicrap, N. Heess, and C. Blundell, "Retrieval-augmented reinforcement learning (r2a)," in *Conference Proceedings*, 2022.