PsyRA – A Retrieval-Augmented Dialogue System for Mental Health Support

Muhammad Hassaan[1], Muhammad Saad Ali Shah[1], Muhammad Uzaif[1], Yasir Saleem Afridi[1], Rehmat Ullah Khattak[1]

[1]Department of Computer Systems Engineering University of Engineering and Technology, Peshawar, Pakistan

***Correspondence:** 21pwcse1982@uetpeshawar.edu.pk, yasirsaleem@uetpeshawar.edu.pk, rehmatullah@uetpeshawar.edu.pk

Citation | Hassaan. M, Shah. M. S. A, Uzaif. M, Afridi. Y. S, Khattak. R. U, "PsyRA – A Retrieval-Augmented Dialogue System for Mental Health Support", IJIST, Vol. 07 Special Issue. pp 279-288, May 2025

Received | April 12, 2025 **Revised |** May 14, 2025 **Accepted |** May 16, 2025 **Published |** May 18, 2025.

Mental health support continues to face numerous challenges, including limited access to care, persistent social stigma, and a shortage of trained mental health professionals. In response to these issues, this paper introduces PsyRA, an innovative AI-powered system designed to enhance psychological assessments through a specialized retrieval-augmented generation (RAG) approach. Unlike conventional chatbots that often fail to capture the nuanced context of patient interactions, PsyRA leverages domain-specific psychological knowledge to deliver more accurate and in-depth assessments. It draws from a carefully curated knowledge base that includes psychological research, diagnostic guidelines, therapy exercises, and intervention strategies to inform its responses and suggestions. PsyRA is equipped to understand patient narratives more clearly, provide evidence-based assessments by retrieving relevant psychological information, and offer personalized intervention recommendations tailored to individual needs. Early evaluations indicate that PsyRA is capable of detecting subtle emotional cues within patient conversations and responding in alignment with established psychological practices. The system demonstrates promising potential to broaden access to mental health support, assist professionals in the assessment process, and reduce the barriers that often prevent individuals from seeking treatment. This work contributes to the expanding field of AI-assisted mental health care by illustrating how retrieval-based models can enhance both the depth and quality of psychological assessments, offering improved emotional sensitivity and reliable, evidence-driven guidance.

Keywords: Mental Health Technology, Retrieval-Augmented Generation, Psychological Assessment, Emotional Understanding, Artificial Intelligence, Conversational Agents.

Introduction:

Artificial Intelligence (AI) has shown great promise in transforming healthcare systems, especially in addressing the global mental health treatment gap caused by a shortage of qualified professionals and rising demand [1]. Around 970 million people worldwide suffer from mental health disorders, with a 13% increase in such conditions over the past decade. In response, AI-based psychological assessment tools have emerged as a way to make mental health services more accessible [2]. Modern AI applications in psychology range from basic chatbot assistants to advanced deep learning models that analyze emotions using text, speech, and physiological signals [3], [4]. However, despite these advances, many current systems struggle to understand patient concerns in context. As a result, they often provide generic responses that lack personalization and clinical accuracy [5], [6], [7].

Most traditional AI tools in mental health rely on fixed rule-based systems or pre-trained models, which limits their ability to adapt to individual mental health conditions. For example, commercial tools like Woebot and Wysa use cognitive-behavioral therapy (CBT) methods within set dialogue paths [8]. These tools can help people with mild symptoms but are less effective for complex or co-occurring conditions [9]. Sentiment analysis and natural language processing (NLP) have also been used to detect signs of emotional distress in language [10], achieving 70–85% accuracy in controlled settings [11]. However, these systems usually operate without real-time knowledge retrieval, reducing their ability to deliver in-depth, evidence-based assessments that meet current clinical standards [12], [13].

To overcome these challenges, our research introduces a next-generation AI-powered psychological assessment system using a Retrieval-Augmented Generation (RAG) pipeline. RAG is a major advancement in language model design. It combines transformer-based large language models with a retrieval system that accesses up-to-date psychological literature, diagnostic standards, and therapy guidelines [14].

Developed by Facebook AI Research in 2020, RAG improves accuracy by dynamically referencing trusted sources outside the model's original training data [11], [14], [15]. Unlike conventional models with fixed knowledge, our system can fetch relevant psychological documents in real time, making its recommendations more adaptable and evidence-based [16]. This ensures that the system delivers context-aware assessments grounded in established psychological theories and clinical practices.

Our AI psychologist offers key innovations for digital mental health and computational psychiatry. First, by using the RAG architecture, it significantly improves assessment accuracy and the quality of therapy suggestions by referencing a verified knowledge base—reducing the risk of hallucinations and factual mistakes common in generative models [17]. Second, it enhances the therapeutic alliance by providing assessments that align with established clinical guidelines, addressing a major weakness of current mental health technologies [13], [18].

By combining RAG with a specialized psychological knowledge base, our system marks a major step forward in AI-based mental health support. Its ability to retrieve and apply domain-specific knowledge allows it to deliver accurate, personalized, and clinically sound assessments. This approach offers a scalable, technology-driven solution to help meet the global demand for mental health care.

Objectives and Novel Contributions:

The increasing demand for accessible mental health support has highlighted the limitations of existing AI-driven tools, particularly regarding contextual understanding and personalization. This study introduces PsyRA, an AI-powered psychological assessment assistant that leverages the Retrieval-Augmented Generation (RAG) to address these challenges.

The primary objectives of this study are as follows:

1. To develop a modular, context-aware conversational assistant for psychological assessment.
2. To enhance factual grounding and emotional alignment in AI-generated responses using evidence-based retrieval.
3. To address the limitations of static rule-based mental health chatbots through dynamic knowledge integration.

Novel Contributions:

This study offers several significant contributions to the existing literature on AI-assisted mental health tools. First, it introduces a domain-specific retrieval-augmented generation (RAG) architecture that is optimized specifically for mental health contexts. This architecture allows the seamless integration of curated psychological knowledge into AI-generated responses, enhancing both the relevance and accuracy of the system's output. Second, the study presents a novel method for reducing common issues associated with large language models (LLMs), such as hallucinations and repetitive outputs, by incorporating a retrieval layer that references authoritative and evidence-based psychological sources. This helps ensure that the system's responses are grounded in credible information. Third, the development of PsyRA as a lightweight psychological assistant demonstrates a balanced focus on both performance and explainability, making the system accessible and user-friendly while maintaining clinical integrity.

In contrast to earlier AI-based mental health tools like Woebot and Wysa, which predominantly rely on predefined cognitive-behavioral therapy (CBT) frameworks and rule-based interaction models, PsyRA stands out for its ability to dynamically adapt its responses. By referencing a curated knowledge corpus in real-time, PsyRA offers greater adaptability and clinical robustness. This dynamic approach allows for more personalized and contextually appropriate support, enhancing the user experience and increasing the system's potential utility in real-world mental health settings.

Clinical Use-Cases and Limitations:

PsyRA is designed as a psychological support tool to help users understand their emotions and access proven coping strategies. It is not meant to provide clinical diagnoses. Its main uses include:

1. Supporting individuals with mild to moderate mental health concerns (such as stress or anxiety) through personalized therapeutic exercises.
2. Helping mental health professionals by generating context-aware assessment summaries.
3. Educating users about mental health conditions using trustworthy, retrieved information.

PsyRA is not intended to replace professional therapy or diagnose complex psychiatric disorders, as it does not have the ability to perform deep clinical evaluations or manage serious crises (like suicidal thoughts). Its limitations include reliance on the quality of its knowledge base, possible misunderstanding of unclear user inputs, and lack of real-time human support during emergencies. These limitations emphasize that PsyRA should only be used as a supplementary tool under the guidance of a mental health professional.

Methodology:

System Overview:

PsyRA is a psychological assessment assistant built on a Retrieval-Augmented Generation (RAG) framework, designed to improve access to mental health support through structured yet flexible interactions. Unlike traditional diagnostic tools, PsyRA does not aim to diagnose; instead, it focuses on assessing user conditions and offering personalized support using retrieved, evidence-based mental health knowledge (as detailed in Section I-A).

At its core, PsyRA combines a vector-based knowledge retrieval system with a large language model (LLM) to increase the accuracy and relevance of its responses. The system actively retrieves relevant mental health literature, diagnostic criteria, and therapeutic strategies from a structured database, integrating this information into its conversations. This method ensures that PsyRA's guidance is both contextually meaningful and grounded in well-established psychological research [4], [14].

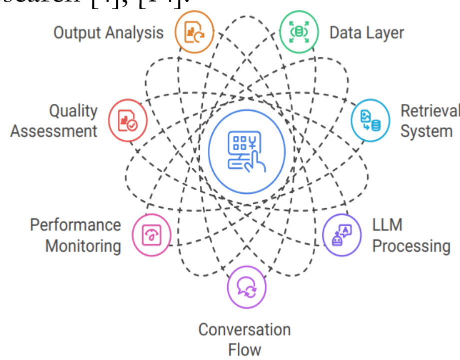

Figure 1. System Architecture Overview

Data Layer & Knowledge Base:

Knowledge Acquisition and Storage:

The effectiveness of PsyRA's psychological assessments relies heavily on a well-structured and dynamically retrievable knowledge base. PsyRA uses ChromaDB, a vector database optimized for high-dimensional similarity searches [19], to store and retrieve embedded psychological literature. This ensures that its assessments are based on clinically relevant information rather than just generative model outputs. To further strengthen the system, future versions will expand the knowledge base with a wider and more diverse range of sources, including peer-reviewed psychological studies and expert-annotated clinical guidelines. Collaboration with mental health professionals will be key in validating the accuracy and relevance of this content, ensuring that PsyRA's assessments remain aligned with real-world clinical scenarios [13].

Vector-Based Retrieval Mechanism:

To optimize document retrieval, PsyRA uses a multi-step embedding and retrieval process. Raw psychological texts are first preprocessed using the Recursive Character Text Splitter, which breaks documents into 1,000-character chunks with a 200-character overlap [20]. This overlap helps maintain context across segments. Each chunk is then converted into vector embeddings using the all-MiniLM-L6-v2 model, which is well-suited for dense passage retrieval [21]. These vector representations are stored in ChromaDB, where they are indexed for semantic similarity searches. When a user submits a query, PsyRA performs a k-nearest neighbor (k-NN) search to retrieve the most semantically relevant text chunks. The retrieved segments are ranked using cosine similarity, ensuring responses are contextually accurate and relevant to the user's psychological concerns. This chunk-based retrieval approach supports efficient and focused knowledge extraction [15]. However, the system faces challenges with vague or unclear queries. In such cases, the similarity search may return less relevant content, forcing PsyRA to rely more heavily on the generative capabilities of its language model. This can introduce bias or reduce precision if sufficient context is lacking. To improve reliability, future updates will include confidence-based filtering and the use of expert-annotated datasets [4].

Retrieval and Response Generation:

Knowledge Retrieval and Contextual Relevance:

The system retrieves relevant information from ChromaDB using all-MiniLM-L6-v2 embeddings, which help align responses semantically with user queries. A k-nearest neighbor (k-NN) search ranks the retrieved text chunks using cosine similarity, selecting the most contextually appropriate content [4], [22]. To maintain factual accuracy, an Answer Relevancy Score is used to decide whether a retrieved chunk should be included, giving priority to high-confidence content. Additionally, faithfulness assessments are applied to ensure that the generated responses remain consistent with the retrieved knowledge [22].

LLM-Based Response Generation and Summarization:

The retrieved content is processed using Llama-3.1-8b-instant, which generates coherent, well-structured, and adaptive responses. To preserve dialogue continuity, the system uses summarization techniques that condense prior interactions, allowing PsyRA to retain and build on user context [23]. Through structured prompt engineering, the system maintains professional communication standards—ensuring responses are neutral, evidence-based, and emotionally supportive. PsyRA also aligns with established ethical guidelines for AI-driven mental health interventions [24], [25].

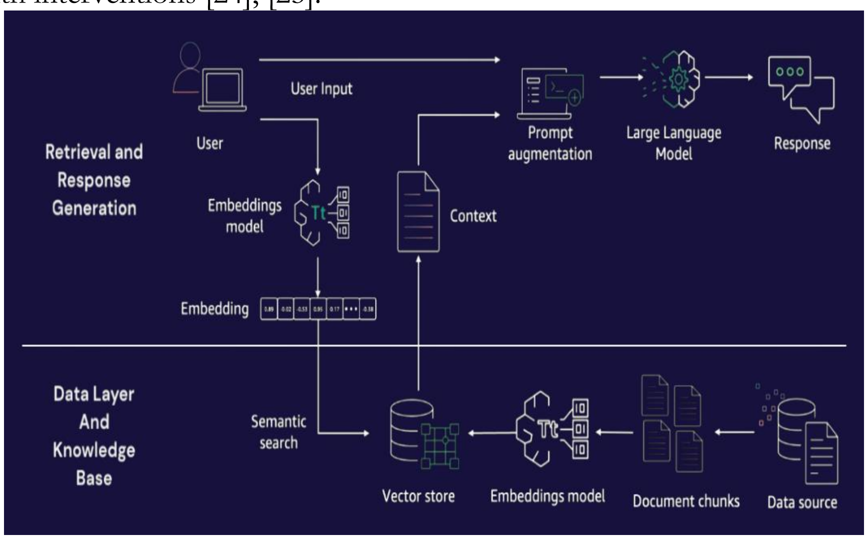

Figure 2. Flow diagram of the proposed RAG-based psychological assessment system.

The architectural pipeline of the proposed system is illustrated in Figure 2, showcasing the interactions between the user interface, retrieval mechanisms, knowledge base, and language model.

Ethical Considerations and Risks:

PsyRA's deployment in mental health support raises several ethical concerns that must be addressed to ensure user safety and system integrity. Key risks include:

1. Data privacy and confidentiality, as sensitive user information must be securely handled and protected from unauthorized access.
2. Misuse or over-reliance, where users might treat PsyRA as a substitute for professional care, especially in severe or crisis situations.
3. Bias and fairness, since training data or retrieval content could unintentionally reflect cultural, gender, or socioeconomic biases.
4. Transparency and accountability, requiring that users are clearly informed of PsyRA's capabilities, limitations, and non-clinical role.

Addressing these issues involves implementing robust data governance, user consent protocols, bias mitigation strategies, and ongoing collaboration with mental health professionals to ensure ethical AI deployment in sensitive domains like mental health.

Emotional Harm:

Inaccurate or insensitive responses can worsen user distress, especially for individuals experiencing severe mental health issues. While PsyRA addresses this risk by using evidence-based content retrieval and structured prompt engineering, limitations remain in fully interpreting subtle emotional cues and complex psychological states. System Misuse Users might rely on PsyRA for important mental health decisions, potentially bypassing professional care. To prevent this, PsyRA clearly communicates its role as a supportive tool and advises users to seek professional consultation for more serious concerns. Privacy and Data Security User inputs containing sensitive personal information must be safeguarded. While PsyRA uses secure data handling practices, the risk of data breaches remains a concern in real-world deployment.

Over-Reliance on AI:

Prolonged use of PsyRA may lead to dependency, potentially reducing users' engagement with human therapists. To mitigate this risk, PsyRA is designed to emphasize its role as a supplementary tool.

These risks highlight the importance of continuous monitoring, user education, and ongoing collaboration with mental health professionals to ensure the ethical deployment of the system [25].

Results and Discussion:

Empirical Validation Using User Queries:

A pilot study tested PsyRA with synthetic user queries simulating mental health concerns like emotional exhaustion and stress. These queries assessed PsyRA's ability to detect emotional cues and provide evidence-based responses. Compared to a baseline (Llama-3.1-8b-instant), PsyRA showed a 16.9% higher TextBlob Sentiment score (0.152 vs. 0.130) and an 18.2% lower Empath Fear score (9.0 vs. 11.0), indicating more supportive responses. Average response time was 1.14 seconds, slightly faster than the baseline's 1.18 seconds. Conversation data are available at
https://github.com/saadshah8/PsyRA/blob/main/README.md.

Evaluation of Text Chunk Quality:

To evaluate the quality of text chunks generated from PDF documents, three metrics were used: average chunk length, semantic coherence, and entity continuity. The average chunk length was 2716.93 characters, providing a balance between readability and computational efficiency. Semantic coherence, assessed using a large language model (LLM), scored between 0.7 and 0.9, reflecting logical consistency. Entity continuity scores showed smooth transitions, with one instance of discontinuity (score of 0). These results validate the effectiveness of the text splitting process for subsequent NLP tasks (see Table 1)

Table 1. Text Chunk Quality Metrics:

Metric	Value
Average Chunk Length	2716.93 characters
Coherence Scores	0.90
	0.80
	0.80
	0.85
	0.85
	0.70
	0.80
	0.80
	0.80
	0.80
Entity Continuity	1
	1
	1
	1

Metric	Value
	1
	0
	1
	1
	1

Retrieval Consistency Evaluation:

Retrieval consistency was assessed using cosine similarity across multiple trials for test questions. The results showed an average similarity of 0.80 for questions such as "What are the symptoms of depression?" and "Best therapies for PTSD?", and 0.75 for "How to manage anxiety attacks?". The standard deviation was 0.216, indicating stable and reliable retrieval performance (see Table 2).

Table 2. Retrieval Consistency Metrics

Test Question	Avg. Similarity	Std. Dev.
What are the symptoms of depression?	0.800	0.200
How to manage anxiety attacks?	0.750	0.250
Best therapies for PTSD?	0.800	0.200

Comparison of Baseline Model and PsyRA:

A comparative evaluation against Llama-3.1-8b-instant assessed PsyRA's effectiveness using both linguistic and emotional metrics. PsyRA showed an 11.7% reduction in the Self-BLEU score (from 0.801 to 0.707), indicating more diverse responses. TextBlob Sentiment improved by 16.9% (from 0.130 to 0.152), and Subjectivity increased by 4.5% (from 0.441 to 0.461), suggesting more encouraging and human-like responses. The Empath Fear score decreased by 18.2% (from 11.0 to 9.0), though negative emotion and anger remained unchanged (see Table 3). To further validate response quality, future evaluations will incorporate a broader dataset, including expert annotations from mental health professionals, ensuring alignment with clinical standards.

Table 3. Evaluation Metrics Comparison of Baseline Model vs. PsyRA

Metric	Baseline Model	PsyRA	Improvement
Self-BLEU Score ↓	0.801	0.707	11.7% Reduction
TextBlob Sentiment ↑	0.130	0.152	+16.9%
TextBlob Subjectivity ↑	0.441	0.461	+4.5%
Empath Negative Emotion ↓	4.0	4.0	0%
Empath Anger ↓	2.0	2.0	0%
Empath Fear ↓	11.0	9.0	18.2% Reduction

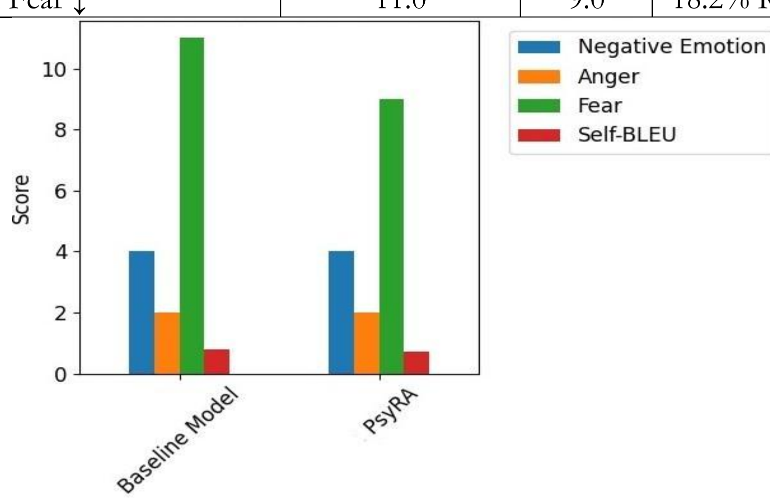

Figure 3. Comparison of Empath Categories and Self-BLEU Between the Systems

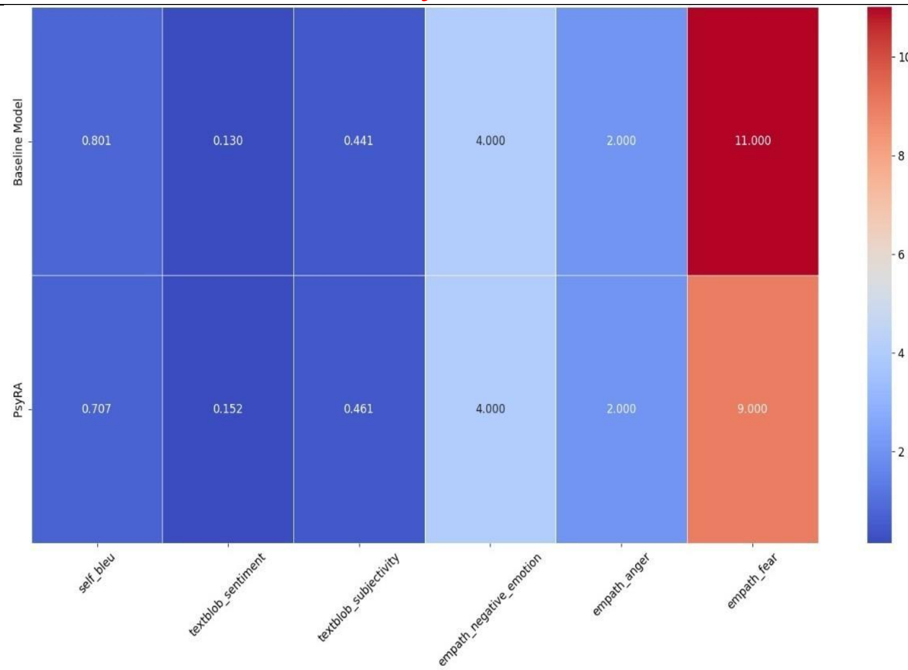

Figure 4. Heatmap of All Metrics

Discussion and Comparison with Existing Approaches:

Strengths of PsyRA Over Prior Systems:

Existing AI mental health tools, including Woebot and Wysa, primarily utilize rule-based frameworks grounded in CBT principles [26], [27]. While effective for basic mood tracking and stress management, these systems often lack the adaptability to handle complex or comorbid conditions. PsyRA addresses this limitation by integrating a RAG architecture that dynamically retrieves and incorporates domain-specific psychological knowledge into its responses.

Compared to baseline model, PsyRA demonstrated

1. Empath reported a 16.9% improvement in emotional expressiveness.
2. 11.7% reduction in repetitive responses, as indicated by the self-BLEU scores.
3. Enhanced coherence in topic transitions and factual consistency.

These improvements suggest that PsyRA offers more diverse, emotionally attuned, and contextually appropriate interactions that align more closely with established psychological practices.

Limitations in Prior Retrieval-Based Systems:

Previous implementations of RAG in healthcare have predominantly focused on factual question-answering tasks, often neglecting the emotional nuances essential for psychological assessments [28], [29]. PsyRA distinguishes itself by tailoring retrieval prompts to achieve empathetic alignment, ensuring that responses are not only factually accurate, but also emotionally supportive. Furthermore, unlike systems that utilize open-source retrieval from platforms such as Wikipedia and PubMed, PsyRA employs a curated knowledge base. This strategy enhances contextual precision and ensures that the information provided aligns with the current clinical standards.

Model and Dataset Biases:

PsyRA's performance may be affected by biases in its knowledge base. For example, the current knowledge base may underrepresent certain demographics (e.g., non-Western populations) or overemphasize specific therapeutic approaches (e.g., Cognitive Behavioral Therapy (CBT)). These biases could result in less relevant recommendations for a diverse user base. Additionally, the LLM's generative capabilities may introduce subtle biases in response phrasing, especially when retrieval yields low-confidence results. In real-world deployment, challenges include ensuring scalability across various clinical settings, handling multilingual inputs, and maintaining system reliability during high user loads. Future work will focus on diversifying the knowledge base and implementing bias detection mechanisms to improve fairness and applicability [13].

Conclusion:

This study introduces PsyRA, an AI-powered system designed to enhance psychological assessment through retrieval-augmented generation (RAG). By utilizing a domain-specific knowledge base, PsyRA delivers context-aware responses that align with psychological best practices. Initial evaluations show that PsyRA effectively recognizes emotional patterns and generates diverse, supportive responses, demonstrating improvements in coherence and emotional intelligence compared to baseline models.

Future developments include transitioning to a structured data model, expanding the knowledge base with expert-annotated resources, and addressing ethical risks and biases. Large-scale, real-world evaluations will assess PsyRA's effectiveness across a range of clinical scenarios. This research contributes to the field of AI-assisted mental health by highlighting how retrieval-augmented models can enhance psychological support and increase accessibility.

References:

[1] E. G. Lattie, E. C. Adkins, N. Winquist, C. Stiles-Shields, Q. E. Wafford, and A. K. Graham, "Digital Mental Health Interventions for Depression, Anxiety, and Enhancement of Psychological Well-Being Among College Students: Systematic Review," *J Med Internet Res 2019;21(7)e12869 https//www.jmir.org/2019/7/e12869*, vol. 21, no. 7, p. e12869, Jul. 2019, doi: 10.2196/12869.

[2] W. H. Organization, "World mental health report: Transforming mental health for all," *Geneva WHO Publ.*, 2022.

[3] A. A. Abd-alrazaq, M. Alajlani, A. A. Alalwan, B. M. Bewick, P. Gardner, and M. Househ, "An overview of the features of chatbots in mental health: A scoping review," *Int. J. Med. Inform.*, vol. 132, Dec. 2019, doi: 10.1016/J.IJMEDINF.2019.103978,.

[4] and H. W. Gao, Yunfan Xiong, Yun Gao, Xinyu Jia, Kangxiang Pan, Jinliu Bi, Yuxi Dai, Yi Sun, Jiawei Meng Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv:2312.10997*, 2023, [Online]. Available: https://arxiv.org/pdf/2312.10997

[5] W. Zhang and J. Zhang, "Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review," *Math. 2025, Vol. 13, Page 856*, vol. 13, no. 5, p. 856, Mar. 2025, doi: 10.3390/MATH13050856.

[6] L. Laranjo *et al.*, "Conversational agents in healthcare: a systematic review," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 9, pp. 1248–1258, Sep. 2018, doi: 10.1093/JAMIA/OCY072.

[7] M. W. Liu, Q. Zhu, and Y. Yuan, "The role of the face itself in the face effect: Sensitivity, expressiveness, and anticipated feedback in individual compliance," *Front. Psychol.*, vol. 9, no. JAN, Jan. 2019, doi: 10.3389/FPSYG.2018.02499.

[8] K. Daley, I. Hungerbuehler, K. Cavanagh, H. G. Claro, P. A. Swinton, and M. Kapps, "Preliminary Evaluation of the Engagement and Effectiveness of a Mental Health Chatbot," *Front. Digit. Heal.*, vol. 2, p. 576361, Nov. 2020, doi: 10.3389/FDGTH.2020.576361/BIBTEX.

[9] M. V. Kathleen Kara Fitzpatrick, Alison Darcy, "Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial," *JMIR Publ.*, vol. 4, no. 2, 2017, [Online]. Available: https://mental.jmir.org/2017/2/e19/

[10] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," *Psychol. Med.*, vol. 49, no. 9, pp. 1426–1448, Jul. 2019, doi: 10.1017/S0033291719000151,.

[11] A. Seyeditabari, N. Tabari, S. Gholizadeh, and W. Zadrozny, "Emotion Detection in Text: Focusing on Latent Representation," Jul. 2019, Accessed: May 21, 2025. [Online]. Available: https://arxiv.org/pdf/1907.09369

[12] D. D. Luxton, "Ethical implications of conversational agents in global public health," *Bull. World Health Organ.*, vol. 98, no. 4, p. 285, Apr. 2020, doi: 10.2471/BLT.19.237636.

[13] E. E. Lee *et al.*, "Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom," *Biol. Psychiatry Cogn. Neurosci. Neuroimaging*, vol. 6, no. 9, pp. 856–864, Sep. 2021, doi: 10.1016/J.BPSC.2021.02.001.

[14] A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, and T. Althoff, "Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach," *Web Conf. 2021 - Proc. World Wide Web Conf. WWW 2021*, pp. 194–205, Apr. 2021, doi: 10.1145/3442381.3450097;PAGE:STRING:ARTICLE/CHAPTER.

[15] P. Lewis et al, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Adv. neural Inf. Process. Syst.*, pp. 9459–9474, 2020.

[16] C. J. Cai, S. Winter, D. Steiner, L. Wilcox, and M. Terry, "'Hello Ai': Uncovering the onboarding needs of medical practitioners for human–AI collaborative decision-making," *Proc. ACM Human-Computer Interact.*, vol. 3, no. CSCW, Nov. 2019, doi: 10.1145/3359206.

[17] and S. F. G. Ramesh, T. Liu, "Retrieval-augmented generation for AI-generated content: A survey," *arXiv Prepr. arXiv2401.10733*, 2024.

[18] V. Ta *et al.*, "User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis," *J. Med. Internet Res.*, vol. 22, no. 3, p. e16235, Mar. 2020, doi: 10.2196/16235.

[19] I. G. Nils Reimers, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Assoc. Comput. Linguist.*, pp. 3982–3992, 2019, doi: 10.18653/v1/D19-1410.

[20] "RecursiveCharacterTextSplitter — 🦜🔗 LangChain documentation." Accessed: May 21, 2025. [Online]. Available: https://python.langchain.com/api_reference/text_splitters/character/langchain_text_splitters.character.RecursiveCharacterTextSplitter.html

[21] W. Y. Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, "Dense Passage Retrieval for Open-Domain Question Answering," *Assoc. Comput. Linguist.*, pp. 6769–6781, 2020, doi: 10.18653/v1/2020.emnlp-main.550.

[22] S. S. Shahul Es, Jithin James, Luis Espinosa-Anke, "Ragas: Automated Evaluation of Retrieval Augmented Generation," *arXiv:2309.15217*, 2023, doi: https://doi.org/10.48550/arXiv.2309.15217.

[23] D. K. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *arXiv:2005.11401*, 2021, doi: https://doi.org/10.48550/arXiv.2005.11401.

[24] Y. He *et al.*, "Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials," *J. Med. Internet Res.*, vol. 25, no. 1, p. e43862, Apr. 2023, doi: 10.2196/43862.

[25] A. S. Miner, A. Milstein, and J. T. Hancock, "Talking to Machines About Personal Mental Health Problems," *JAMA*, vol. 318, no. 13, pp. 1217–1218, Oct. 2017, doi: 10.1001/JAMA.2017.14151.

[26] "Woebot: The Mental Health Ally - Apps on Google Play." Accessed: May 21, 2025. [Online]. Available: https://play.google.com/store/apps/details?id=com.woebot&hl=en&pli=1

[27] "Wysa - Everyday Mental Health." Accessed: May 21, 2025. [Online]. Available: https://www.wysa.com/

[28] "How Retrieval-Augmented Generation (RAG) Supports Healthcare AI Initiatives." Accessed: May 21, 2025. [Online]. Available: https://www.makebot.ai/blog-en/how-retrieval-augmented-generation-rag-supports-healthcare-ai-initiatives

[29] "Retrieval-augmented generation - Wikipedia." Accessed: May 21, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Retrieval-augmented_generation