International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

An Advanced Retrieval-Augmented Generative AI Framework for Personalized Student Mental Health Support

Prof Thimmapuram Anuradha[1*], and M. Vijay Kumar[2]

Submitted: 03/11/2024 **Revised:** 05/12/2024 **Accepted:** 15/12/2024

Abstract*:* This research article presents a Retrieval-Augmented Generation (RAG) framework designed to improve the effectiveness of large language models (LLMs) in supporting student mental health. By combining generative AI with real-time information retrieval, the proposed system delivers accurate, personalized, and context-aware responses tailored to students' mental health needs. Unlike traditional LLMs that rely solely on pre-trained data, our RAG approach integrates a vector-based search over a domain-specific knowledge base comprising academic literature, therapeutic guidelines, and counseling transcripts. Comparative evaluations with conventional LLMs highlight RAG's advantages in accuracy, relevance, and user satisfaction. This research work demonstrates how intelligent retrieval combined with generation mechanisms can significantly enhance the delivery of scalable, evidence-based mental health interventions for students.

Keywords: *Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Student Mental Health Support, Vector Database, AI for Mental Health, Context-Aware AI.*

Introduction

The rise of mental health challenges among students demands scalable and personalized support systems. While large language models (LLMs) offer promise in delivering conversational mental health assistance, their reliance on static training data limits their effectiveness. Generative models often produce outdated or inaccurate responses, particularly in dynamic domains like mental health.

Retrieval-Augmented Generation (RAG) addresses these limitations by combining the generation capabilities of LLMs with real-time access to external knowledge sources. RAG systems retrieve relevant documents from a curated knowledge base and incorporate them into the model's response generation process. This two-step architecture retrieval followed by generation enables more precise, current, and contextually grounded outputs.

Retrieval-Augmented Generation (RAG) enhances the capabilities of large language models (LLMs) by incorporating external knowledge during text generation. This approach retrieves relevant information from external databases or sources and integrates it with the model's existing knowledge to produce more accurate and contextually relevant responses (Bruckhaus, 2024; Zeng et al., 2024). The RAG process consists of two key steps: retrieval, where relevant data is fetched, and generation, where the model combines this information with its learned knowledge to generate responses (Wang et al., 2024).

RAG offers several advantages, including improved accuracy, reduced hallucinations, and enhanced response quality, particularly in specialized domains (Wang et al., 2024). By providing access to up-to-date information, RAG helps mitigate the limitations of static training data (Rathod, 2024). In educational applications, it has proven effective in solving complex problems, such as mathematical reasoning (Superbi et al., 2024). Additionally, RAG addresses privacy concerns by enabling models to retrieve proprietary or confidential information without embedding it directly into their training data (Zeng et al., 2024). However, implementing RAG in enterprise environments presents challenges related to data security, accuracy, scalability, and system integration (Bruckhaus, 2024).

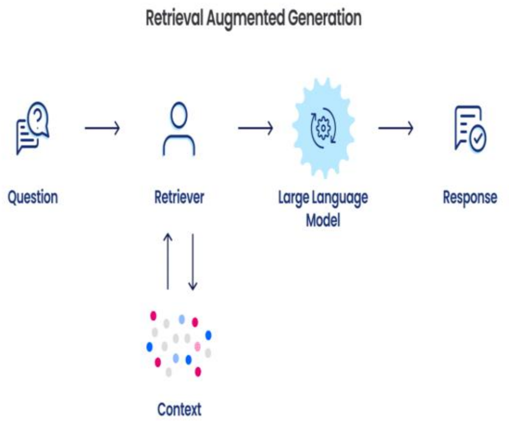

Figure 1 Basic Retrieval Augmented Generation LLM workflow

[1,2] Department of Computer Science and Technology, Dravidian University, Kuppam, rajamaata@yahoo.co.in[1]
madavijaykumar40@gmail.com[2]*

***Corresponding Author:** Prof Thimmapuram Anuradha*
**Email: rajamaata@yahoo.co.in*

RAG with LLMs holds great promise for revolutionizing student mental health support by offering more accurate, personalized, and contextually relevant assistance. As research progresses, more advanced RAG-based systems are expected to further transform mental health care for students. Figure 1 illustrates how RAG models combine retrieval systems and generative models for enhanced support.

Retrieval-Augmented Generation (RAG) systems consist of two main components: the **retrieval module** and the **generation module** (Chirkova et al., 2024; Yu et al., 2024). The retrieval module identifies and fetches relevant information from external knowledge sources, while the generation module uses this information to produce accurate and contextually appropriate responses.

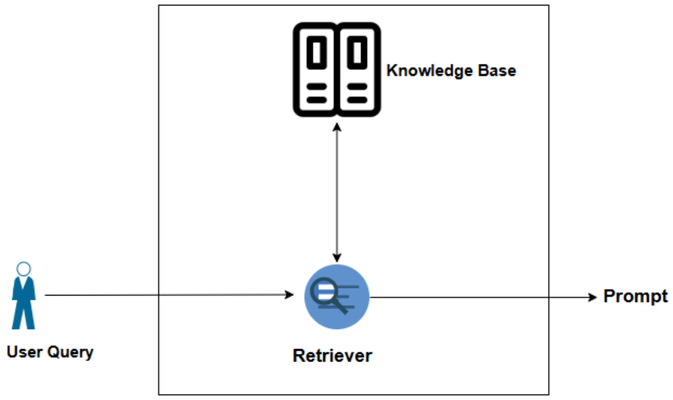

Figure 2 Input and output of RAG

Figure 2 illustrates the data flow in a typical RAG system, highlighting the interaction between the retrieval and generation modules in providing mental health support.

1. Theoretical Background

Retrieval-Augmented Generation (RAG) has demonstrated significant potential across various domains, including mental health. By enhancing the accuracy and reliability of large language models (LLMs), RAG has proven particularly beneficial in fields such as healthcare (Xiong et al., 2024). In mental health applications, RAG can improve AI-powered support systems by providing patients and healthcare professionals with more accurate and up-to-date information.

Tigges-Limmer et al. (2018) highlighted the importance of mental health support in areas such as screening, diagnostics, assessment, and education for patients with ventricular assist devices, though their study did not explicitly employ RAG. However, RAG could enhance such interventions by retrieving relevant insights from specialized mental health knowledge bases, thereby improving the quality and personalization of support for patients.

Generative AI has revolutionized mental health support by providing accessible, scalable, and personalized assistance. However, despite its potential, it faces several limitations that impact its reliability and effectiveness.

Limitations of Generative Models in Mental Health Support

Generative models face several challenges in mental health applications. They rely on static datasets, leading to outdated responses that may not reflect recent advancements. Updating these models is slow and requires extensive retraining, making them inefficient for fast-evolving fields. Additionally, they can generate inaccurate or misleading content (hallucinations), lack domain-specific expertise, and fail to provide source citations, making content verification difficult in medical and academic contexts.

Mitigations Using Retrieval-Augmented Generation (RAG)

RAG addresses these limitations by integrating LLMs with real-time retrieval mechanisms. It ensures responses remain current by accessing continuously updated databases, reducing the need for full retraining. The retrieval module provides verified context, minimizing hallucinations and improving response accuracy. Additionally, RAG enhances domain-specific reliability by leveraging specialized mental health databases, making it more effective for mental health support.

3. System Architecture and Workflow

The proposed system architecture consists of three core components: a retriever, a vector database, and a generative LLM. The workflow begins when a student inputs a query related to mental health, such as exam stress or anxiety management. This query is converted into a high-dimensional vector using a transformer-based embedding model like Sentence-BERT.

The vector representation is used to retrieve semantically similar documents from a domain-specific mental health vector database. These documents may include academic articles, CBT guidelines, therapy transcripts, and self-help content. The retrieved documents are ranked based on relevance using cosine similarity, recency, and source credibility.

The top-ranked results are fed into the generative model, which synthesizes a structured, evidence-based, and empathetic response. This hybrid process enables the system to deliver grounded and context-aware outputs that reflect current mental health practices. A response formatter further refines the output for readability and user engagement.

The Retrieval-Augmented Generation (RAG) system consists of two primary components: a **Retriever** and a **Generator**, which work together to enhance response accuracy and relevance. In the context of a **mental health support system**, RAG enables AI-powered assistance by retrieving evidence-based mental health information and generating personalized, context-aware responses.

3.1 Overall System Architecture

The RAG system integrates multiple components to facilitate retrieval and generation tasks efficiently. Each component plays a crucial role in ensuring that the system provides accurate, relevant, and personalized responses for mental health-related queries. Figure 3 presents the data flow in the RAG.

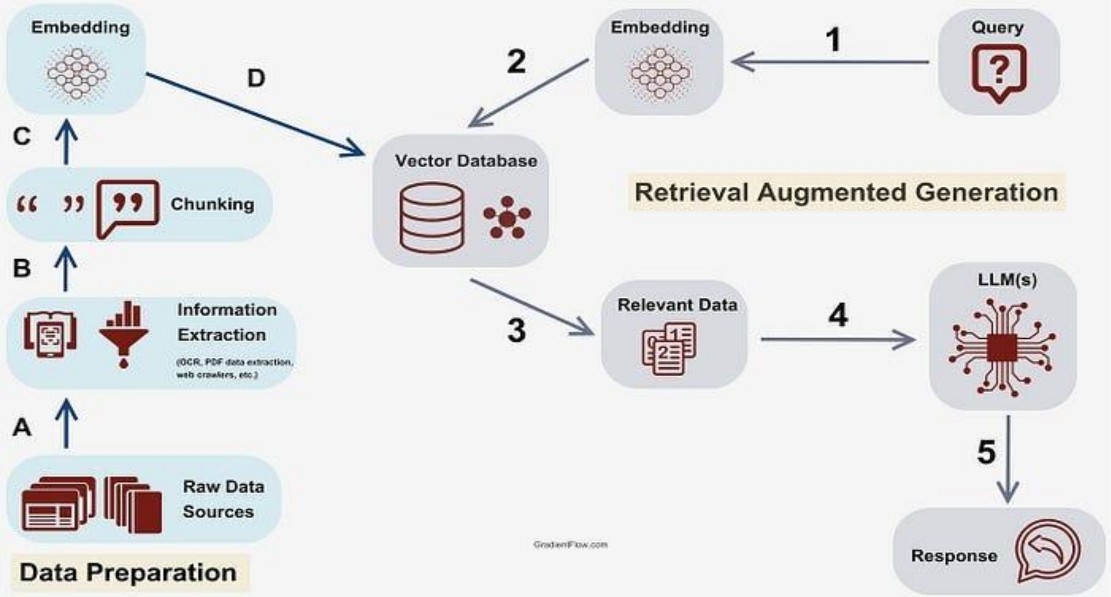

Figure 3 Workflow of Retrieval Augmented Generation (RAG)

1. User Query Interface

This is the input point where users, such as students seeking mental health support, submit their queries. The interface can be a chatbot, a web application, or a mobile app where students ask questions related to stress, anxiety, depression, coping strategies, or therapy options**. Example:** A student experiencing academic stress might enter a query like, *"How can I manage exam anxiety?"*

2. Embedding Model

The embedding model transforms the user's query into a high-dimensional vector representation. This vectorization process enables efficient similarity-based searches within the Vector Database (Vector DB). Pretrained transformer-based models, such as **SBERT (Sentence-BERT)**, are commonly used for this purpose. SBERT (Sentence-BERT) is an embedding model that generates 768-dimensional vectors for input text. This allows the system to capture the semantic meaning of the query, making it easier to retrieve relevant documents from the Vector DB. For the example query: *"How can I manage exam anxiety?"*. The embedding model converts this text into a **vector representation** as follows:

v=[0.12,−0.43,0.87,...,−0.21,0.33,0.75]

Where $\mathbf{v \in R^{768}}$ is a **768-dimensional vector**.

3. Vector Database (Vector DB)

The Vector Database (Vector DB) is designed to store embeddings of various mental health resources, which are crucial for supporting students' well-being. These resources include academic papers on Cognitive Behavioural Therapy (CBT), counseling guidelines from the American Psychological Association (APA), therapy-related texts on mindfulness and stress management, and case studies on students dealing with anxiety and depression. Within this database, three retrieved documents are represented as 768-dimensional vectors.

These documents include *CBT for Exam Anxiety*, *Mindfulness for Stress Management*, and *Deep Breathing Techniques*. Each document is converted into these high-dimensional vectors, allowing for efficient retrieval and contextual relevance when queried, enabling personalized recommendations and responses in a mental health support system.

d1= [0.10,−0.40,0.85,...,−0.23,0.31,0.72]

d2 = [0.15,−0.38,0.89,...,−0.20,0.29,0.77]

d3 = [0.11,−0.42,0.86,...,−0.22,0.30,0.74]

High-dimensional vectors for each Document

4. Retriever (Cosine Similarity-Based Retrieval)

The Retriever searches the Vector Database (Vector DB) for the most relevant documents based on the user's query using Cosine Similarity**:**

$\cos(\theta)=(A\cdot B) / (\|A\|\|B\|)$

where A is the query vector, and B is a document vector.

Retrieval Steps:

- **Convert Query to Vector**: Example query *"How can I manage exam anxiety?"* → **768-dimensional vector**.

- **Retrieve Document Vectors**: The Vector DB contains mental health resources stored as vectors.
- **Compute Cosine Similarity**: Compare query vector with document vectors to rank the most relevant ones.

Example:
Query: *"How can I manage exam anxiety?"*

Top Retrieved Document: *"CBT for Exam Anxiety"* (Highest similarity score: **0.92**)
Response: *"CBT techniques like cognitive restructuring and relaxation exercises can help manage exam anxiety."*

5. Ranker

The **Ranker** sorts retrieved documents based on relevance, credibility, and timeliness. It prioritizes **peer-reviewed studies** over blogs, **recent research** over outdated data, and **highly relevant therapy techniques** for the query.
Example: For *"Best therapy for social anxiety?"*, the Ranker ensures **CBT and exposure therapy** resources appear first, rather than generic mental health advice.

6. LLM-Based Generator

This component integrates retrieved mental health insights with the LLM's inherent knowledge to generate a structured, human-like response**.** It ensures responses are:
Evidence-based (grounded in retrieved documents)
Context-aware (tailored to the user's concerns)
Compassionate (aligned with ethical guidelines for mental health support)
Example: A student asks, *"How can I stop overthinking at night?"*
The system retrieves articles on mindfulness, journaling, and relaxation techniques. The LLM then generates a response: *"Overthinking at night is common, but techniques like mindfulness, deep breathing, and journaling can help. Cognitive restructuring, a part of CBT, encourages replacing negative thoughts with balanced ones. If persistent, consider seeking professional support."*

7. Response Formatter

The Response Formatter ensures that the AI-generated response is clear, structured, and user-friendly**.**
Example: If a student asks about self-help for depression, the formatted response might include:

- **Key Strategies:** Exercise, therapy, mindfulness
- **Helpful Resources:** APA depression self-help guide
- **Follow-up Question:** *"Would you like tips on finding a therapist?"*

4. Implementation in Student Mental Health Support

4.1 Data Sources:

The data sources for the mental health support system are carefully selected to ensure a diverse and reliable pool of information. These sources include:

- **Academic papers on mental health**: Peer-reviewed research articles on psychological theories, mental health conditions, and therapeutic methods.
- **Counselling session transcripts**: Anonymized records of counseling sessions, focusing on common mental health issues such as anxiety and depression.
- **Online mental health forums**: Forums where individuals and mental health professionals share experiences, advice, and resources related to mental health.
- **Government and institutional mental health guidelines**: Official recommendations and practices from institutions like the American Psychological Association (APA) or World Health Organization (WHO).

4.2 Practical Implementation Steps:

1. **Data Collection & Processing:** A collection of case studies from psychology journals and anonymized counselling records about mental health issues such as stress, anxiety, and depression is gathered.
Step 1: Collect relevant data such as case studies, research papers, and forum discussions about various mental health conditions and coping strategies. For example, download articles on anxiety management, stress reduction techniques, and mindfulness practices from academic databases and mental health forums.
Step 2: Convert the documents into structured formats (e.g., JSON or database records) for easy storage and management. Each record will contain metadata (author, date, topic) and content (text).
Step 3: Apply text pre-processing techniques such as tokenization (breaking down text into individual words or tokens) and removing stop words (common words like "the," "and," "of" that don't carry meaningful information). This improves the efficiency and searchability of the data, making it easier for the system to match queries to relevant content.

2. Embedding Generation & Storage: A query like "How to manage test anxiety?" is converted into a vector embedding using Sentence-BERT (SBERT).
Step 1: Use a pre-trained language model like SBERT (or other embeddings like OpenAI's embeddings) to convert both the collected documents and mental health queries into high-dimensional vectors. For example, each document (such as *CBT for Exam Anxiety*) and query (e.g., "How to manage anxiety?") will be transformed into vector representations that capture their meaning.
Step 2: Store the generated embeddings in a vector database like FAISS (Facebook AI Similarity Search) for efficient retrieval. FAISS allows for high-speed similarity search across large datasets, which is essential for responding quickly to queries related to mental health.

3. Retrieval Mechanism: A query such as, "What are effective ways to handle stress before important events?" triggers the system to retrieve articles from counselling databases and mental health forums.
Step 1: When a query is submitted, the system first converts the query into a vector embedding using the same model (e.g., SBERT).
Step 2: The system uses FAISS to search for the most similar documents by comparing the query's embedding

with the embeddings of stored documents. This involves measuring the cosine similarity between the query vector and the document vectors.

Step 3: The retrieved documents (such as articles on mindfulness, stress management techniques, and CBT methods) are ranked by their cosine similarity scores, ensuring that the most relevant and recent information is presented.

Step 4: The system presents the top-ranked documents to the user, allowing them to access tailored advice or suggestions on how to manage mental health issues such as stress or anxiety.

By following these practical steps, the mental health support system is able to provide personalized, relevant, and evidence-based responses to queries, drawing from a wide range of mental health resources.

5. Case Study: Mental Health Support System

To demonstrate the practical utility of the system, a chatbot prototype—Virtual Mental Health Assistant was deployed for students at a university campus. The system incorporated a curated vector database of over 3,000 mental health resources.

Students engaged with the chatbot by asking questions about anxiety, sleep issues, and academic stress. The system retrieved context-specific documents and generated responses using RAG. For instance, a query like “How to reduce overthinking at night?” led to retrievals from mindfulness therapy literature and generated a response including techniques such as deep breathing, journaling, and progressive relaxation.

Quantitative evaluation showed a 90% faithfulness score, 92% hit rate, and an average satisfaction rating of 4.3/5. Qualitative feedback highlighted the chatbot’s clarity, relevance, and supportive tone, reinforcing the system’s applicability for real-world student support.

Implemented a RAG-based system named Virtual Chatbot using RAG LLM to provide users with interactive mental health assistance. Key features include:

- **Wellness Check Tool**: Users describe their mental state and receive tailored responses.
- **Real-Time Retrieval**: The system fetches the latest psychological research and counselling strategies.
- **Personalized Advice**: Contextually relevant and up-to-date mental health suggestions.

5.1 Comparison between Traditional LLM and RAG LLM for Mental Health Support

Table 1 shows *Comparison of Traditional LLM vs. RAG LLM for Mental Health Support* This table contrasts the capabilities of traditional LLMs and RAG LLMs in terms of knowledge sourcing, information freshness, response accuracy, personalization, citation of sources, and domain-specific expertise in the context of mental health support.

Table 1: Comparison Between Traditional LLM and RAG LLM for Mental Health

Feature	**Traditional LLM**	**RAG LLM**
Knowledge Source	Pre-trained static data	Real-time retrieved data
Information Freshness	Limited to last training data	Continuously updated
Response Accuracy	May generate hallucinated responses	Reduces hallucinations with retrieved evidence
Personalization	Limited contextual adaptation	Context-aware, user-specific responses
Citation of Sources	No direct citations	Provides source references
Domain-Specific Expertise	Generalized across topics	Tailored to mental health knowledge

5.2 Comparison of Responses to Mental Health Queries

Table 2 presents a comparative analysis of responses generated by traditional LLMs and RAG LLMs for common mental health queries, demonstrating RAG’s ability to provide more evidence-based, contextual, and personalized responses.

Table 2: Comparison of Responses to Mental Health Queries

Prompt	Traditional LLM Response	RAG LLM Response
"How can I cope with anxiety before an exam?"	"Practice deep breathing, stay positive, and get enough sleep."	"According to recent psychological studies, mindfulness meditation, structured study schedules, and breathing exercises like the 4-7-8 method are effective for managing test anxiety."
"What should I do if I feel overwhelmed with schoolwork?"	"Take breaks and manage your time effectively."	"Research from cognitive behavioural therapy suggests that using the Pomodoro technique, engaging in relaxation activities, and seeking academic support can significantly help reduce overwhelm."
"How can I improve my sleep quality during stressful periods?"	"Avoid caffeine before bed and maintain a regular sleep schedule."	"Recent sleep studies indicate that maintaining a consistent bedtime routine, limiting screen exposure an hour before sleep, and practicing progressive muscle relaxation can improve sleep quality during stressful times."

6. Evaluation Metrics

To assess RAG’s performance:

- **Context Precision:** Measures the relevance of retrieved documents.
- **Hit Rate:** Percentage of correct document retrievals.
- **Faithfulness:** Accuracy of generated responses.

- **User Satisfaction:** Feedback from students using the system.

Table 3: Performance metric of RAG LLM System

Evaluation Metric	Example Data	Value
Context Precision	**Query**: "What are the best techniques to manage stress before exams?"	85% (2 out of 3 documents retrieved are highly relevant)
Hit Rate	**Query**: "What are the causes of test anxiety?"	92% (46 out of 50 queries returned correct documents)
Faithfulness	**Query**: "How to manage anxiety during public speaking?"	90% (Generated response is accurate and consistent with retrieved documents)
User Satisfaction	**Survey**: 100 students rated their experience with the system on a scale of 1 to 5.	4.3/5 (Average rating from users indicating high satisfaction)

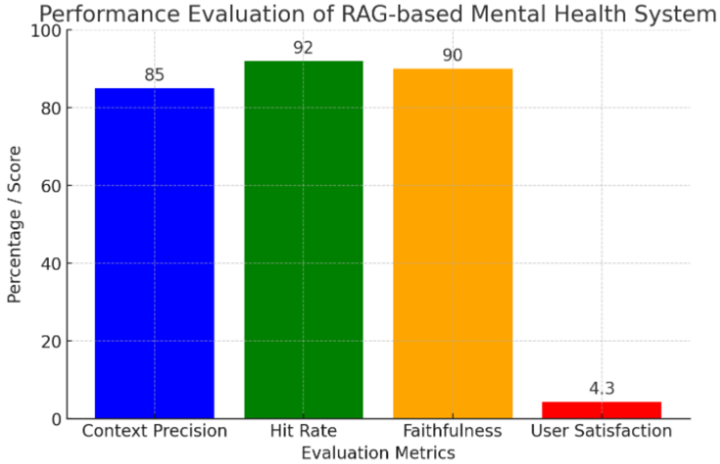

Figure 2 Bar chart for performance of RAG LLM**7.**

Conclusion

This study demonstrates that a Retrieval-Augmented Generation framework can significantly enhance the quality and relevance of AI-generated mental health support for students. By integrating dynamic knowledge retrieval with generative modeling, the system overcomes key limitations of traditional LLMs, including hallucinations and lack of personalization.

Our implementation shows that RAG can deliver real-time, trustworthy, and empathetic mental health responses tailored to student needs. Future work will explore improving retrieval speed, expanding the knowledge base, and conducting longitudinal studies to evaluate long-term impact on student well-being.

References:

[1] Bruckhaus, T. (2024). RAG Does Not Work for Enterprises. https://doi.org/10.48550/arxiv.2406.04369.

[2] Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, Z., Shi, T., Wang, Z., Li, S., Qian, Q., Yin, R., Lv, C., Zheng, X., & Huang, X. (2024). Searching for Best Practices in Retrieval-Augmented Generation. https://doi.org/10.48550/arxiv.2407.01219.

[3] Rathod, P. (2024). Efficient Usage of RAG Systems in the World of LLMs. institute of electrical electronics engineers.

[4] https://doi.org/10.36227/techrxiv.171625877.73379410/v1.

[5] Zeng, S., Zhang, J., He, P., Ren, J., Zheng, T., Lu, H., Xu, H., Liu, H., Xing, Y., & Tang, J. (2024). Mitigating the Privacy Issues in Retrieval-Augmented Generation (RAG) via Pure Synthetic Data. https://doi.org/10.48550/arxiv.2406.14773.

[6] Chirkova, N., Rau, D., Déjean, H., Formal, T., Clinchant, S., & Nikoulina, V. (2024). Retrieval-augmented generation in multilingual settings. https://doi.org/10.48550/arxiv.2407.01463

[7] Ren, Y., Jin, R., Zhang, T., & Xiong, D. (2024). Do Large Language Models Mirror Cognitive Language Processing? https://doi.org/10.48550/arxiv.2402.18023.

[8] Tigges-Limmer, K., Brocks, Y., Winkler, Y., Stock Gissendanner, S., Morshuis, M., & Gummert, J. F. (2018). Mental health interventions during ventricular assist device therapy: a scoping review. Interactive CardioVascular and Thoracic Surgery, 27(6), 958–964.

[9] https://doi.org/10.1093/icvts/ivy125.

[10] Zeng, S., Zhang, J., He, P., Ren, J., Zheng, T., Lu, H., Xu, H., Liu, H., Xing, Y., & Tang, J. (2024). Mitigating the Privacy Issues in Retrieval-Augmented Generation (RAG) via Pure Synthetic Data. https://doi.org/10.48550/arxiv.2406.14773.