RAG Based Educational Chatbot GENIE

Riya Redkar
Department of AI&DS Engineering
Ajeenkya DY Patil School Of Engineering
Pune, India
riya.redkar@dypic.in

Lokesh Sonar
Department of AI&DS Engineering
Ajeenkya DY Patil School Of Engineering
Pune, India
lokesh.sonar@dpic.in

Diya Umale
Department of AI&DS Engineering
Ajeenkya DY Patil School Of Engineering
Pune, India
diya.umale@dypic.in

Shravani Kadu
Department of AI&DS Engineering
Ajeenkya DY Patil School Of Engineering
Pune, India
shravani.kadu@dypic.in

Prof. Varsha Babar
Department of AI&DS Engineering
Ajeenkya DY Patil School Of Engineering
Pune, India
varshababar@dypic.in

***Abstract*— This paper introduces GENIE (Guided Education with Neural Intelligence and Exclusivity), a novel system that revolutionizes document interaction by combining state-of-the-art Retrieval-Augmented Generation (RAG) architectures and large language models (LLMs). GENIE facilitates more efficient user interaction with scholarly PDFs using an easy-to-use question-answering interface, guaranteeing streamlined knowledge retrieval and effortless information access. GENIE utilizes a tiered PDF storage and retrieval system, allowing users to categorize documents effectively into general storage and a specialized "PDF Information Retrieval Box." This segregation allows for optimal retrieval performance and keeps the user interface simple to use. With LangChain for document-oriented querying, GENIE uses a RAG pipeline to process content in PDFs, extract and preprocess text, and translate it into semantic embeddings. These embeddings are cached in a vector database, supporting efficient semantic queries and accurate answer generation. For ensuring content credibility, GENIE does not make use of any external knowledge source but strictly uses the given documents for generating answers. In cases where document-specific content is limited, the system supports an optional facility for general-purpose querying driven by external LLMs. This facility equips users with the ability to opt between document-specific responses and general AI-generated responses. By tapping into state-of-the-art Natural Language Processing (NLP) technology, GENIE sets a new standard for smart and interactive knowledge retrieval systems. It shows how RAG models can generate accurate, varied, and factual responses, making GENIE a groundbreaking tool for the academic and research communities.**

Keywords*— *Retrieval Augmented Generation, Large Language Model, Chatbot, LangChain, PDF storage, Academic Question and Answering.

I. INTRODUCTION

The increasing number of digital files, especially PDFs, presents a considerable challenge the effective extraction of knowledge from their predominantly textual content. In recent decades, a variety of tools and methodologies have been created to tackle this problem, spanning from simple keyword search capabilities to more sophisticated text searching and natural language processing techniques. [1]. Most of the solutions to date, however, are unable to deliver contextually relevant data within time and in the right manner. The emergence of AIML, specifically with the creation of LLMs, has transformed this process, allowing for advanced and efficient retrieval of information from large amounts of digital files [2].

ChatGPT is commonly utilized as it is extremely popular, especially among learners who require aid with their assignments and have problems with plagiarism and academic dishonesty. It also assists instructors by allowing them to create course material and activities. Aside from academic uses, ChatGPT has numerous uses, including research, leisure, coding, explanation, analysis, and document creation. It can be utilized to assist in programming changes, integration tasks, data translation, and formatting. It also assists with virtual reality learning environments [3].

Despite the widespread use of these models, little study has been executed on how they can be utilized for particular interactions with PDFs. By joining LLMs with retrieval of documents in a communicative interface, this study aims to close this gap as well as provide a more personalized approach to document interactions. Recent growth in LLMs have significantly improved data retrieval and text extraction capabilities. Managing the increasing number of digital documents is made easier by this advancement, which makes it possible to effectively extract knowledge and generate insights. Recent growth of LLMs has made it possible to build RAG systems. The addition of a retrieval system to these improved

systems broadens the application of LLMs. RAG generates extremely accurate and context-specific replies by fusing the power of generating processes with the strength of information retrieval [4]. A RAG system begins by gathering relevant parts in response to the user query from a large document library. This feature enables the model to produce responses which can be infused with specific, appropriate data taken from the information obtained in addition to being based on a good grasp of the language. This technique greatly improves the model's capacity to provide in-depth answers to fact-based queries and navigate through enormous data repositories, making it an asset for precise and educational knowledge retrieval. In order to improve user interaction with documents through a conversational interface, this study presents a novel method for text extraction utilizing a LLM system.

II. RELATED WORK

The use of deep learning methods has significantly advanced the fields of text understanding and information retrieval in recent years. In comparison to conventional keyword matching and rule-based approaches, which tend to fail to handle complicated documents, deep learning models provide much stronger solutions that can successfully deal with complicated structures. For example, M. Li and colleagues created the "BiomedRAG" model was designed to improve data extraction within biomedical contexts by leveraging diverse databases, leading to improved prediction accuracy [5]. Similarly, the "Almanac" framework, created by M. D. Cyril Zakka and colleagues, is designed to retrieve medical guidelines and treatment recommendations and outperforms traditional LLMs in terms of factual accuracy, completeness, user preference, and safety [6]. Additionally, "LeanDojo" a RAG-based LLM that uses a vast array of tools and data to speed up theory proving, was proposed by K. Yang and colleagues [7]. Lastly, in order to improve language development, P. Lewis and colleagues investigated a fine-tuning approach for RAG models that used both already trained parameterized and non-parameterized memory [8].

Major improvements have occurred in information retrieval, especially through the development of pre-trained neural language models closely related to our emphasis. There are works focused on improving the retrieval module to enable specific downstream tasks, such as question answering, through approaches such as search, reinforcement learning, or even a latent variable model, similar to our work. While such methods usually involve applying various retrieval architectures and optimization approaches to achieve competence in a single task, our work indicates that one retrieval-based architecture can be efficiently fine-tuned to be highly competent at many tasks.

A study by Zhang and colleagues presented a novel approach called the Multi-Modal Knowledge-aware Hierarchical Attention Network (MKHAN), aiming to enhance explainability in medical question answering through the integration of a knowledge graph [9]. However, many methods in this area are often customized to individual scenarios, leading to restrictions regarding the adaptability needed for more general document interaction applications.

This work enhances earlier studies with the description of RAG paradigm system which allows users to interactively navigate through their own interest PDF documents. In this study, advanced sentence embedding techniques are utilized to boost the effectiveness of the retrieval process. This context is fused into the outstanding feature of the LLM response generation. The user experience is highly improved. In this case, enhanced participatory engagement human-system interaction with the system is achieved which allows deep dialogues within the semantic context of the uploaded PDF documents and as such maximize information retrieval effectiveness and enhance user engagement.

A. Retrieval Augmented Generation (RAG)

RAG is one of the strongest methodologies in artificial intelligence, having the capability to provide updated and reliable external knowledge, thereby making it possible to reap significant advantages over tasks. Especially in the era of AIGC, RAG's high potential for retrieving supporting evidence makes it possible to improve existing generative AI systems, resulting in high-quality outputs. Recent research was performed in LLMs, which exhibited unmatched performance in language comprehension and generation; however, they even endure inherent issues, such as hallucinations and outdated internal knowledge. Since RAG's high potential for providing the latest and relevant supporting evidence, RA-LLMs have been proposed for leveraging external and reliable information foundations, thereby improving the generative quality of LLMs instead of relying on the internal knowledge of models [10].

B. Large Language Model (LLM)

LLMs are highly advanced AI systems capable of understanding, interpreting, and generating human-like text through training on vast datasets. These models are developed using extensive datasets, which allow them to understand context, semantics, and the subtleties of language. They are able to perform a variety of tasks, from responding to simple greetings to participating in complex conversations. LLMs learn from new data on a continuous basis, enhancing their capability to predict and generate coherent responses. Instead of being static programs, they are dynamically evolving digital cognitive systems that excel in natural language understanding, summarization, translation, and conversational AI applications across various domains.

C. LangChain Framework

LangChain is powerful tool that seeks to enhance LLMs through the ability to interact with external data sources, APIs, and information discovery methods seamlessly. LangChain enables developers to build applications that incorporate LLMs

to carry out tasks like viewing documents, answering questions, and task automation. LangChain enables the integration of models with databases, search engines, and utilities like RAG to enhance AI's capability to provide accurate and relevant responses. LangChain simplifies workflow automation, such as step-by-step reasoning and decision-making. LangChain, because of its extremely flexible architecture, allows developers to build intelligent chatbots, virtual assistants, and AI applications that interact with structured and unstructured data sources.

III. PROPOSED METHODOLOGY

A. Storage for Academic PDFs

The first component of GENIE focuses on an efficient and structured storage system for academic PDFs. Users can upload and store their PDFs in a dedicated document storage system, similar to traditional PDF storage solutions. The system ensures proper organization of uploaded PDFs, allowing users to access and manage their documents seamlessly. When a document needs to be used for information retrieval, the user can manually select and upload the relevant PDFs into a dedicated "PDF Information Retrieval Box." This separation ensures that stored documents do not interfere with retrieval efficiency while maintaining an intuitive user experience.

1) Document Upload and Organization

The process begins when users upload their academic PDFs into the system. Upon upload, the system allows user to categorizes the documents. This categorization ensures that PDFs are organized in a manner that enhances later retrieval.

2) Separation of Storage and Retrieval

Once the documents are uploaded and stored, the system separates the documents into two categories: Storage and PDF Information Retrieval Box. The storage system holds all documents in an organized manner, while the retrieval system allows users to select and upload relevant PDFs for information extraction.

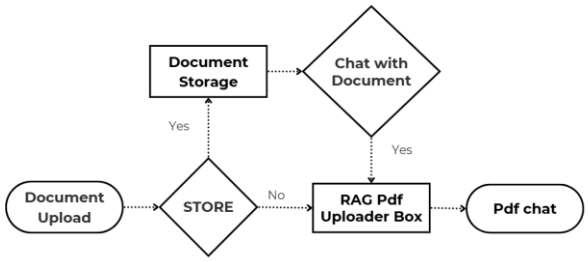

Fig .1. PDFs Storage and Q n A Flow Diagram.

B. PDF-Based Question Answering using LangChain

GENIE employs LangChain to facilitate document-specific querying, utilizing a Retrieval-Augmented Generation (RAG) pipeline.

1) Text Extraction and Preprocessing

The process begins with the upload of academic documents, such as PDFs. These documents are first processed through a step where text is retrieved from the PDFs. The content is parsed, and relevant text sections are extracted.

2) Embedding Generation and Storage

After extracting the content from the PDF, it is transformed into embeddings. Embeddings are "high-dimensional vector representations that encapsulate the semantic meaning of the content". The vectorized information is subsequently stored in a vector database, facilitating efficient and rapid retrieval through similarity searches.

3) Semantic Search and Answer Generation

The query submitted by the user is converted into embedding by the system. Then semantic search is performed in the vector database. The retrieved text chunks are passed to LangChain's LLM for response generation. The LLM processes the relevant text and generates a coherent answer based on the content of the uploaded PDFs.

4) Staying Strict to Document Content

An important requirement of the system was to ensure that the answers generated by the LLM were strictly based on the uploaded PDFs. This prevents the system from generating answers that are outside the scope of the user's provided materials.

C. General Question Answering with User Consent.

If the answer to user's query is not match with the PDFs content, GENIE includes a feature that allows users to opt for general question answering. This secondary feature enables the system to provide more broad, general-purpose responses when necessary.

1) User-Driven Choice for General Responses

If a query does not match any content within the uploaded PDFs, GENIE prompts the user with an option to opt-in for a generalized AI-generated response. This feature is designed to give users the flexibility to obtain answers even when document-specific information is not available.

2) Querying an External LLM

Once the user consents, the system queries an external LLM, which is trained on a wide range of general knowledge, to generate an answer. The model uses its broad training data to provide a response that, while not based on the uploaded documents, still tries to address the query in a coherent and informative way.

3) Flexibility for General and Document Based Answers

The system's design ensures that users can always choose between document-specific answers or more general, AI-generated insights. This flexibility was carefully integrated into the system's user interface, allowing users to make an informed decision about the type of response they wanted.

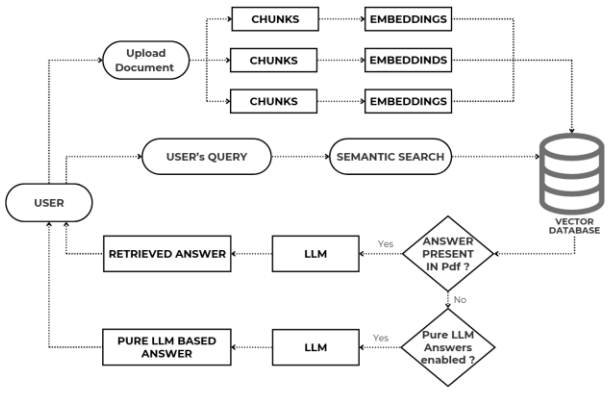

Fig. 2. GENIE Chatbot Operation Diagram

IV. RESULT AND ANALYSIS

The GENIE prototype provided valuable insights into its performance in processing uploaded PDFs and generating responses based on user queries. The evaluation focused on three main aspects: the accuracy of retrieved information, the system's ability to filter answers based solely on document content, and the impact of toggling the Pure LLM-based response mode. For testing phase, we provided a pdf on "OOP with C++.

For the first test, a query about "Object-Oriented Programming (OOP) in C++" was submitted. The system retrieved a detailed explanation, outlining fundamental concepts like classes, encapsulation, inheritance, and polymorphism. The structured and informative response indicated that the document-processing feature efficiently extracts relevant knowledge when the necessary information is available in the uploaded file.

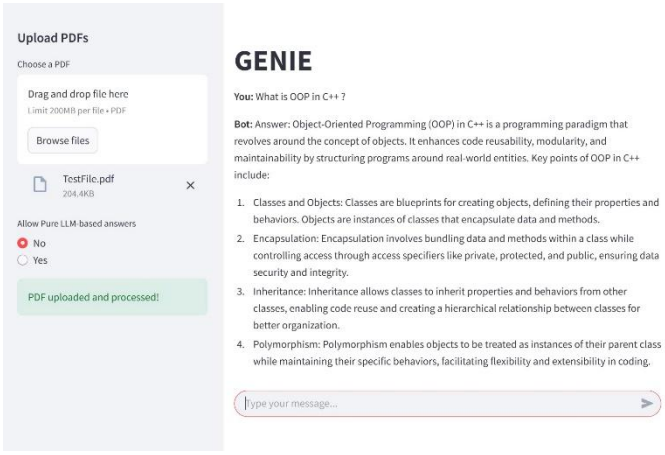

Fig. 3. Answer Generated Strictly from PDF (Direct LLM based Answer Disabled)

The second test, a query about "Agile Methodology" was made, but with the Pure LLM-based mode disabled. This time, the system was unable to generate a response, displaying a message that no relevant content was found. This confirmed that when restricted to document-based responses, the AI strictly adheres to the information available in the uploaded file and does not rely on any external knowledge.

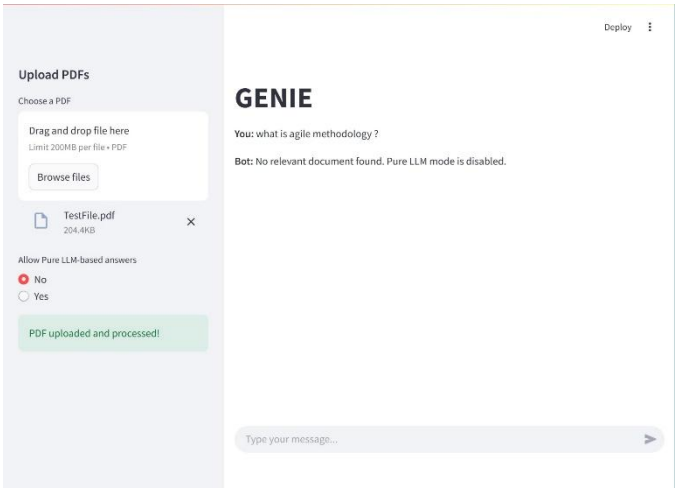

Fig. 4. Query Outside the PDF is not answered when LLM Based Answering is Disabled.

In the third test, the same query was made while the Pure LLM-based mode was enabled. The system successfully retrieved a structured response that not only defined Agile but also highlighted its key principles, such as customer collaboration, iterative development, and flexibility. This result showed that when the AI is allowed to incorporate its own knowledge along with the document, it generates comprehensive answers that align well with industry standards.

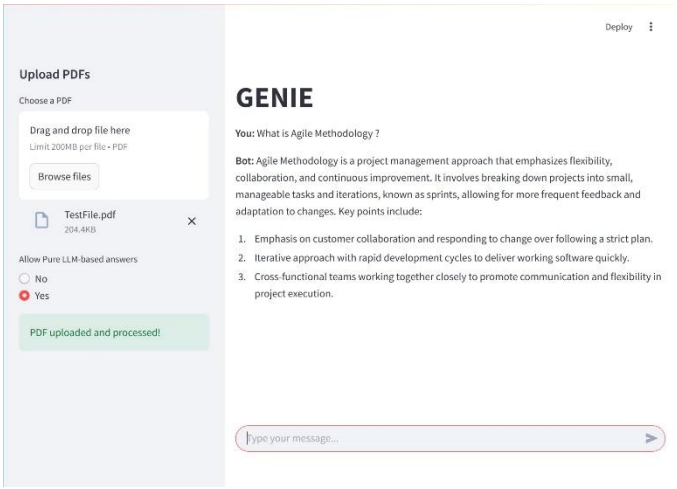

Fig. 5. Answers Generated from outside the PDF (Direct LLM based Answer Enabled)

To compare GENIE's accuracy with a general AI model, we asked the same question, "Classification of requirements?" to both GENIE and ChatGPT.

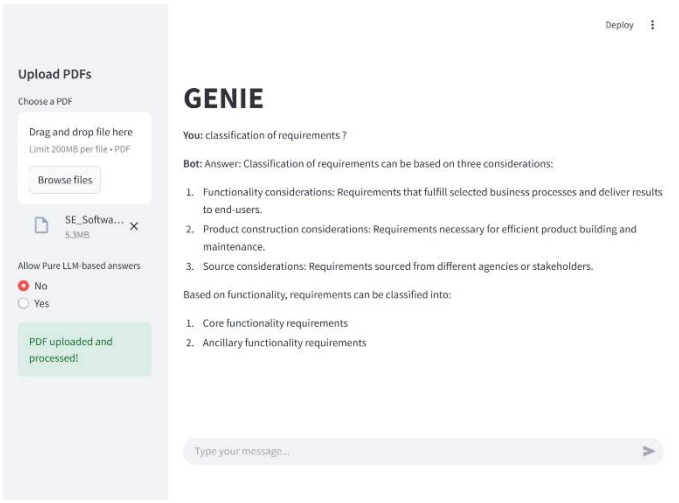

Fig. 6. Answer Generated by GENIE Strictly from PDF (Pure LLM-based Answer Disabled)

In GENIE, a Software Engineering PDF was uploaded, and LLM-based responses were turned off, forcing it to answer only from the document. As shown in Fig. 6, the reply was directly pulled from the PDF, listing three types of considerations and their breakdown exactly as written in the material.

On the other hand, ChatGPT's response (Fig. 7) came from its internal knowledge. While the answer made sense, it used a different classification that didn't match the content of the uploaded document.

This test showed that GENIE, when restricted to document-only mode, gives answers that are source-accurate, while ChatGPT, though informative, may not always align with study material. This makes GENIE ideal for academic tasks were sticking to the document matters most.

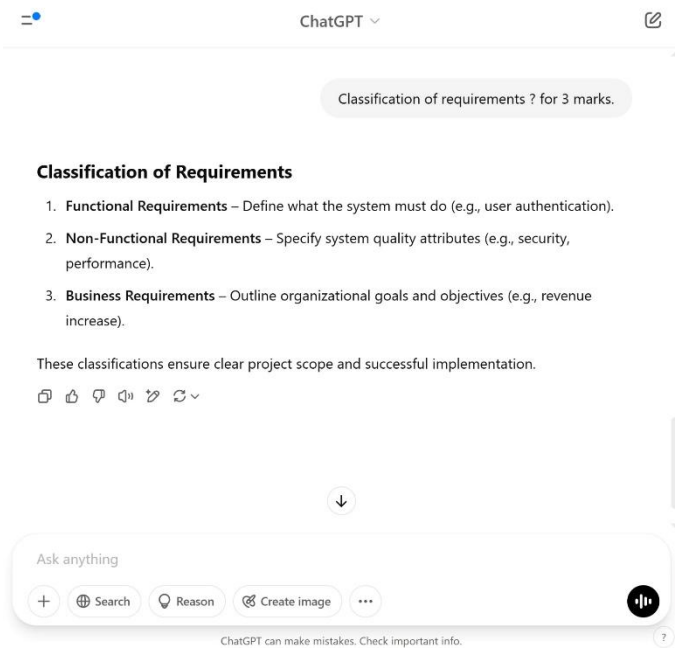

Fig. 7. Answer Generated by ChatGPT

V. CONCLUSION AND FUTURE SCOPE

GENIE is a powerful AI driven academic assistant that allow the user or students interact with study material or any documents they wised. Unlike normal chatbot, it gives responses based on resources or documents provided, ensuring accuracy and relevant solution. With features like personalized library, interactive assistant and containerized data organization, GENIE build up academic productivity. By keeping datasets separate, it reduces confusion and helps users manage different subjects effectively. This model is built to optimize learning, save time, and provide a structured approach to academic research. GENIE make the accessing and organizing academic materials more efficient.

The Future Scope of GENIE (Guided Education with Neural Intelligence and Exclusivity) hold significant potential for expanding its capabilities, particularly in domain of research and academic study. There can be some filed where GENIE can evolve to enhance its performance. GENIE collect all the document and find interconnected concepts, and establish semantic net between different resources. Which allows to integrate multi document querying & Cross Referencing. With advance powered in LLM we can allow GENIE as virtual academic assistant that permits interactor to use voice command interact for hand off research. Upcoming iteration can also assist in providing statistical analysis explanation and can suggest possible methodologies on priority trends by understanding complex datasets. As model can continue to evolve, the integration of container - based dataset differentiation creates an opportunity to enhance personalization, security and accuracy. It makes sures that interactor on multiple resources can organize, manage and maintain information without crossing the content. Future scope of GENIE can allow support multi language. It implements language translation for any languages. It also allows multi-language querying and short summary.

ACKNOWLEDGMENT

The authors of this paper would like to express their sincere gratitude to Ajeenkya DY Patil School of Engineering for providing the necessary resources and facilities to carry out this research. Special thanks are extended to Dr. Bhagyashree Dhakulkar (Head of Department), prof. Varsha Babar (Mentor) for their invaluable guidance and support throughout the study. We also acknowledge the encouragement and constructive feedback provided by our peers and colleagues during the development of this research.

REFERENCES

[1] Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. Multimedia tools and applications, 82(3), 3713-3744.

[2] Bahl, L. R., Brown, P. F., De Souza, P. V., & Mercer, R. L. (1989). A tree-based statistical language model for natural language speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 37(7), 1001-1008.I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. Education sciences, 13(4), 410.

[4] Xu, L., Lu, L., Liu, M., Song, C., & Wu, L. (2024). Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology. Heritage Science, 12(1), 118.

[5] Li, M., Kilicoglu, H., Xu, H., & Zhang, R. (2025). Biomedrag: A retrieval augmented large language model for biomedicine. Journal of Biomedical Informatics, 162, 104769.

[6] Hiesinger, W., Zakka, C., Chaurasia, A., Shad, R., Dalal, A., Kim, J., ... & Nelson, J. (2023). Almanac: Retrieval-augmented language models for clinical medicine.

[7] Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., ... & Anandkumar, A. (2023). Leandojo: Theorem proving with retrieval-augmented language models. Advances in Neural Information Processing Systems, 36, 21573-21612.

[8] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.

[9] Zhang, Y., Qian, S., Fang, Q., & Xu, C. (2019, October). Multi-modal knowledge-aware hierarchical attention network for explainable medical question answering. In Proceedings of the 27th ACM international conference on multimedia (pp. 1089-1097).

[10] Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., ... & Li, Q. (2024, August). A survey on rag meeting llms: Towards retrieval-augmented large language models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 6491-6501).