Enhancing student support with a RAG powered chatbot-A novel approach to query resolution

Sindhu M[1], Shobha Chandra K[2], Druthi Mahesh[3], Shamitha L[4], Saniha M[5]

[1] *Department of CSE, Malnad College of Engineering, Hassan, Karnataka, India*

[2]*Assistant Professor, Department of CSE, Malnad College of Engineering, Hassan, Karnataka, India*

[3,4,5] *Department of CSE, Malnad College of Engineering, Hassan, Karnataka, India*

***Abstract*—This paper describes a RAG-based approach using AI to optimize the experience of students interacting with textbooks. Finding accurate answers to questions is one of the problems faced by students which is resolved by this project. Additionally, using pre-trained Large Language Models and Natural Language Processing allows the system to function as a great aid for self-education. The architecture of the system, its purposes, and the possibility of investigation into other areas of academia are being developed as a RAG-based system with particular importance to improve student's academic performance.**

***Index Terms*—Artificial Intelligence, Large Language Models, Natural Language Processing, Retrival-Augmented Generation.**

I. INTRODUCTION

Rapid developments in artificial intelligence are changing the way students learn and absorb information. Any textbook- based studying may suffer from problems such as inefficient and lengthy searching of content, difficulty comprehending core concepts, and poorly written summaries. This study proposes a retrieval-augmented generation on (RAG) based method that is adjoined to each student in order to rectify such shortcomings of manual reading.

The framework implements AI with retrieval techniques to pick out relevant information from textbooks, allowing users to get correct, relevant and easy-to-understand answers to their questions. Integration of retrieval and creativity assists in ensuring that the content given is adequate and accommodates the student. Additionally, the system utilizes advanced language models in a targeted manner with the help of academic materials to provide a more focused approach.

Constructive thoughts are complex but this method makes things easier and gives individual assistance which contributes to the betterment of the learning and problem solving. This approach allows for accurate retrival of academic information, which could reshape the manner in which one approaches self learning.

II. LITERATURE SURVEY

In this study, the methodology proposed makes use of a mix of different levels of retrieval, augmentation, and interface improvements from a variety of RAG frameworks. The proposed systems demonstrate promise for high accuracy but their high reliance on computation makes the systems difficult to scale up. Additionally, there is a significant dependency on the quality and coverage of the retrieval database. This limits deployment in dynamic environments unless amelioration of issues is carried out.

In order to evaluate the RAG, more than 100 retrieval tests were run. Each of the components such as retrieval, augmentation, and modular integration had more than 25 tests and was given an optimization strategy. The results were calculated and reviewed on many measures and boundaries establishing the application of RAG systems in the wider context of LLMs.

In order to put the theory into practice, GPT 3.5 and the OpenStax Prealgebra textbook were employed to demonstrate the construction of an RAG system. Questions were embedded into a given cosine similarity and compared to relevant sections of a textbook with various levels of prompt instruction to assess the quality of the answer provided. Human surveys were completed to estimate LLM's relevance and grounding in the generated responses.

The R^2-Former module is a low-rank adaptive transformer that serves to augment retrieval information without changing the retrieved texts. To deal with this information, it was retrieved-aware prompting that was applied to the LLM input embeddings, enhancing robustness and effectiveness, particularly, in scenarios where the resources available are few.

The system was also equipped with a lightweight retrieval evaluator that was able to estimate the confidence of the retrieved documents and help determine if to use, to throw away, or to even add useful content. Also, a decomposition-recomposition technique was implemented in order to remove unwanted data that positively affected the short-term memory and long-term memory content generation tasks. With these tasks, the RRAFT method utilized the chain-of-thought explanation reasoning band to improve retrieval efficiency by being able to differentiate between beneficial and irrelevant documents. This method was then tested against other techniques with datasets such as PubMed and HotpotQA and its consistent performance gain in document-based reasoning challenges proved its credibility.

Simpler RAG methods do not even need to go through the processes of learning. In these cases, the retrieved documents were placed on the model's input more or less as they were. Results from experiments carried out across various datasets have shown that this approach tends to be a positive one during language modeling, as the changes were deemed to be such that it is equivalent to increasing the model's size by 2-3 times.

The RAG Foundry platform platform embraces a range of differenti- ated workflows such as few-shot learning, template-based and fast manufacturing, data augmentation, and fine tuning, to cite but a few. This flexibility enables dataset creation, retrieval pipeline assembly, and model evaluation to be optimized for different RAG setups.Llama-3 and Phi-3 models have been stated to have fine-tuned settings for positive advancements both in terms of retrieval and generative capabilities as well. In their assessment of retrieval efficiency, the authors to this article used the Llama2-7b-chat model to perform the generating function and the MPNET when generating the embedding. The framework utilized 42 domain-specific queries sampled across 6 broad categories of keywords leveraging cosine similarity and hypothesis testing to determine keyword placement to explain retrieval performance appropriately.

In the FLARE framework, a confidence-based active retrieval stratergy is incorporated whereby preliminary phrases are created and their confidence evaluated after which additional retrieval is performed only if needed. This iterative technique incorporates explicit question creation and masked sentence questions on multi hop QA and summarizing tasks, especially on long form output data sets.

The Rewrite-Retrieve-Read (RRR) framework is based on three pillars, which include using a search engine, reading an LLM and using a trainable rewriter to rewrite the query. The performative feedback from the LLM helps the rewriter to shape its query in an improved manner with the help of reinforcement learning. This technique is evaluated across MCQ and open-domain QA tasks over a number of datasets, such as MMLU and HotPotQA, with evidence of improvements in task completion rates being found.

The Selfmem framework consists of a memory selector and a retrieval-augmented generator, with an aim to bolster iterative generation. The memory selector chooses the most appropriate output which will be used as a memory in the upcoming rounds, while the generator offers candidate to the outputs. The performance of this system was investigated on a number of datasets, including JRC-Acquis, XSum and BigPatent. ROUGE and BLEU were the success metrics as these confirm the effectiveness of the system.

In this study, LangChain alongside various enhanced embeddings including Word2Vec, GloVe and BERT are merged with databases containing vector embeddings for boost in contextual linkage and response accuracy. RAG and normal models are compared, which shows the significant growth in contextual linkage and response accuracy achieved with the implementation of advanced embedding in RAG models.

This study employs a BM25-based approach in conjunction with the more robust DPR model to retrieve the pertinent text, after which they integrate it into a Fusion-in-Decoder model. Collectively, they show how encoder and decoders can be employed in a model, where both passage embedding and language generation take place simultaneously. T5 and BART models were utilized, with performances analyzed through EM metrics.

UPRISE provides a retriever designed for smaller LLM models, such as the GPT-Neo-2.7B, complementary with larger models like GPT-3 and BLOOM. It employs task generalization with prompt scoring and retriever training via contrastive learning. The retriever aids in increasing the accuracy of the responses by comparing messages to prompts that have previously been set up, eliminating the need to retrain larger LLMs directly.

RAG pipelines rely on semantic chunking while also incorporating GROBID-based biographic data retrieval, along with specially tuned embedding algorithms and a searching system that takes abstracts as input integrated into one model. The framework for tests was RAGAS with novel prompting strategies for data-science-centric inquiries adhering to CRISP-DM categorization.

This method methodology incorporates techniques aimed at improving the system's RAG accuracy and efficiency tackling challenges in performance optimization for scaling and retrieval quality for domain-specific or general-purpose cases.

III. DESCRIPTION

The paper assesses the current trends surrounding Retrieval- Augmented Generation (RAG) architectures and how it inte- grated with Large language models for improved performance on different knowledge focused tasks. With RAG frameworks addressing the weaknesses associated with traditional question answering systems, they significantly improve the precision, resource utilization and flexibility.

RAG techniques have three beginning stages which are the Naive, Advanced, and Modular which highlight their use retrieval, generation, and augmentation respectively. In addition to that, retrieval-friendly modules like R2- Former and retrieval-prompts have been developed to help to develop belief concepts about the external acquisition of the LLMs. Moreover, domain based RAFT model can be utilized for fine-tuning to improve the access of niche content while controlling irrelevant information [6]. What's more , in-context RALM has the advantage of being more straightforward in its approach while making architectural changes quite necessary because of the rapid resources that are available for deployment.

The CRAG is yet another advance that improves on the earlier versions of this tool with the addition of an evaluat- ing module to assess the quality and trustworthiness of the information obtained using web searches and decomposition-recomposition protocols.

RAG education applications show the potential of students to appropriately engage in conversations while at the same time being able to rely on RAG models to recall information necessary in textbook accuracy.

Interrogating the Augmented Generation model of text retrievers' features to combine a variety of third parameters, and widgets as well as GAT on AI and machine learning apps, efforts are made on fundamental methodologies and frameworks. The RAG Foundry is a collaborative project intended to enable the construction and evaluation of RAG systems. It allows the combination of generative and retrieval paradigms in large language models, LLMs. It incorporates all required components for data generation, teaching, inferencing, and evaluation.

RAG systems applied to telecom or battery requirements emphasize on chunk size, embedding , and key word positioning. The research also suggests approaches for improving retrieval efficiency and using technical language and jargon in an efficient manner.

FLARE is shown to to enhance long form generation tasks by iteratively recalling information. It does so by limiting the extent for hallucinating and increasing the accuracy of their outputs.

To address the above challenges in retrieval-augmented LLM's RRR framework is introduced. RRR is distinctive from other approaches as it focuses on query modification and addresses the issue of the difference between the information that was expected to be retrieved and actually obtained which is not a strong retrieval strategy. To achieve this a retrainable language model is used to rewrite queries followed by a reinforcement learning modelling approach that enhances the training process as it builds on the LLM reader feedback.

State of the art outputs are produced by Selfmem framework from exemplifying across a range of benchmarks focused on generation tasks including abstract summarization, translation, and conversation modelling showcasing its applicability for optimizing retrieval based augmented models. The Selfmem framework stores the models generated outputs and uses them to produce successive generations.

The inclusion of AI in business and education for research purposes was limited self enhancing frameworks capabilities integration that are capable of adding value to the sector. More so highlighting how in both educational and business settings the effectiveness of LLMs can be greatly enhanced as placement of embeddings in business IT can drive development.

The authors explore the application of RAG for open-domain question answering and illustrate how the integration of text passage retrieval with sequence-to-sequence generative models results in top performance on Natural Questions and TriviaQA. This technique performs well with retrieved passages of multiple sizes hence enhancing the ability to retrieve relevant content to answer a multi-faceted query more efficiently.

The text describes how how UPRISE incorporates a simpler retriever that enhances the effectiveness of larger models such as GPT-3 and BLOOM. The retriever, trained under a smaller LLM (GPT-Neo-2.7B), utilizes contrastive learning for generalization testing and prompt scoring so as to increase response accuracy by providing the relevant prompts to the inputs without the requirement for tuning the larger models.

In conclusion, the work integrates these concepts to show- case the transformative potential that RAG frameworks have for QA systems to handle the challenges in the particular domain effectively. current designations.

IV. LIMITATIONS

While the advancements seem promising, the suggested RAG methods have certain limitations which pose scaling issues and affects real world applicability. One of the them is the requirement for extensive computational resources which pose difficulty in scaling these systems for greater usage. Moreover, the system’s performance is greatly dependent upon the breadth and richness of the retrieval database which highlights need for more work to solve dynamic business environment problems.

Even though the work offers a good synthesis of the theoretical and technology components of RAG, it does not extend into domain related cases. External database reliance and the granularity versus performance trade-off challenge realistic projects greatly, especially those in targeted areas.

Further, the dependence on textbook data hinders the system’s ability to accommodate a variety of question types which is another constraint. Such dense responses would not be conversational enough for certain subjects, making them considerably less appealing in certain educational settings. Also, there exist the ethical dimensions relating to the risk of hallucinations affecting the ‘realism’ of LLM responses and consequently how students navigate such information.

The use of terminological neural networks, which is the technique used by the R2-Former module, has the potential of improving the semantic of translation although it tends to be architecture dependent which could potentially be a disadvantage at times. The additional computational burden may not be very great but could pose problems in applications that have restrained computational resources.

The use of web based information retrieval systems tend to delay the model and results do bring about issues with performance as a result of dependency on web based searches. The efficiency of the delivery evaluator and the compositionality method are additional factors that determine the effectiveness of the model.

In the training of the model they had a high- quality labelled dataset, that has gold and negative labelled amplifications too, which in every scenario is not possible or practical. Furthermore, the models which were trained has been limited to a specific domain, hence it would not have been generalizable to other applications of the trained models.

In the end, even though the performance is notable when previously retrieved documents are entered as inputs in the model, this approach overly burdens the model in the same way that it could be increased by 2-3 in size. Resources required to carry out certain language modeling tasks may be costly so this approach may not work for all activities.

The RAG Foundry framework is fully customizable for a specific domain but its RAG configuration and retrieval quality are crucial for its efficiency; furthermore, cloning the system may be hard due to hardware and data discrepancies.

To a large extent computational overheads as well as a system’s efficiency for big data were not studied, which leaves questions regarding the framework's efficiency on bigger data applications.

To a great extent, the study's technical focus and the small survey sample of queries limit the nature of the findings. They were also not followed by the broadening of the scope for bigger datasets and how to counter the degradation of retrieval quality with longer words and heavy use of acronyms. Such an approach also raises concerns regarding the negative impact of similarity score threshold on the accuracy of retrieval.

Due to the finalized finalized approach's iterative multiple retrieval operations, FLARE has been excluded since it raises operational costs and is not efficient compared to single pass strategies. The losses particularly affect the versatility of the approach because of the reliance on retrieval quality and low confidence on the recognition accuracies of the symbol which varies with the dataset or retrieval engine.

The less adaptability of the RRR framework also results from the conditions of having frozen retrievers and readers. The performance is determined by the quality of the pseudo data and the reinforcement learning set. Although the method moderates the problem of hallucinations, it does not solve it completely meaning that there is still room for improvement in the retrieval process.

In Selfmem, long streams of tasks have the potential to either carry over biases from previous generations or create a form of memory confusion which may have an effect on the generation task as well. The expectation that the memory produced on schedule x will always be better than the one produced on any previous schedule is indeed conceivably unrealistic while even a single generation cycle has an elevated cost using the latter shifted perspective.

The commercial and pedagogical nature of the study restricts the applicability of the methods to different spheres. To add to that, sophisticated embeddings and architectures such as LangChain are heavy in nature and for a low tech user or a small scale enterprise, the approach is hard to adopt.

The method covered in one of the paper is very computationally expensive because it leans towards the use of large pre-trained models. Retrieval is also a major factor along with how well the system can fetch relevant data in regards to performance. Fine tuning is also not a simple task as it requires a good dataset which makes the use of this approach somewhat specific domain restricted.

UPRISE is limited by its need for a set of prompt, which in some situations, might not be useful. Moreover, success in this framework is governed by the base LLM and the data used for training the retriever, resulting in the model being unable to generalize to other tasks or areas and also being subject to underlying dataset biases.

Ultimately, the semantic chunking based pipeline of Paper 16 will be limited by the availability of certain datasets and the large computational resources required for embedding and tuning. If the text segment modifier is not set correctly, then errors will occur, and because the system uses a first abstract retrieval architecture, it may not be able to locate important information that resides far in the depth of the articles being viewed.

These limitations describe the difficulty of scaling tasks or using RAG frameworks in the outside world, pointing out the necessity to further improve on these problems.

V. PRELIMINARY DESIGN

A. Libraries and Tools Used:

Gradio: This is used for building the user interface.
OpenAI Models: Embedding models and GPT-4 for natural language understanding and generation.
LangChain: For Architecting the RAG-based Structure.
Elasticsearch: For indexing and searching the content.

Functionality Overview:
The chatbot connects to a database that is loaded with subject-specific PDFs. Users have the ability to upload further more documents while the chat continues. Retrieves and summarizes for suitable content from the PDFs. Provides the content retrieved back to the user for verification.

B. Working of RAG:

Data Ingestion: Subject specific PDFs are uploaded into the system. Perform chunking of the documents. Embeddings for each chunk are generated and then later indexed for efficient retrieval.
Content Retrieval: Process the users query. Index embeddings match the query. The most relevant k chunks are then retrieved.

Synthesis: LLM instructions, the user's question, content obtained, and the previous conversations are all added together. Make use of the LLM to provide a reasonable and orderly answer.

VII. CONCLUSION

This project clearly illustrates the role of a RAG enabled chatbot for catering to the informational queries of students. This Chatbot integrates retrieval and generative AI capabilities with the assist of which accurate answers related to context are obtainable from a broad pool of subject-relevant PDF files. Future work will expand the knowledge base, enhance the processing of complex queries and enable support of many languages for a wider range of students.

VIII. REFERENCES

[1] ”Retrieval-Augmented Generation Approach: Document Question An- swering using Large Language Model” by Kurnia Muludi, Kaira Milani Fitria, Joko Triloka, Sutedi.

[2] ”Retrieval-Augmented Generation for Large Language Models: A Sur- vey” by Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, Haofen Wang.

[3] ”Retrieval-augmented Generation to Improve Math Question- Answering: Trade-offs Between Groundedness and Human Preference” by Zachary Levonian, Chenglu Li, Wangda Zhu, Anoushka Gade, Owen Henkel, Millie-Ellen Postle, Wanli Xing.

[4] ”R2AG: Incorporating Retrieval Information into Retrieval Augmented Generation” by Fuda Ye, Shuangyin Li, Yongqi Zhang, Lei Chen.

[5] ”Corrective Retrieval Augmented Generation (CRAG)” by Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, Zhen-Hua Ling.

[6] ”RAFT: Adapting Language Model to Domain Specific RAG” by Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, Joseph E. Gonzalez.

[7] ”In-Context Retrieval-Augmented Language Models” by Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton- Brown, Yoav Shoham.

[8] ”RAG Foundry: A Framework for Enhancing LLMs for Retrieval- Augmented Generation” by Daniel Fleischer, Moshe Berchansky, Moshe Wasserblat, Peter Izsak.

[9] ”Observations on Building RAG Systems for Technical Documents” Sumit Soman and Sujoy Roychowdhury.

[10] ”Active Retrieval-Augmented Generation (FLARE)” by Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, Graham Neubig.

[11] ”Query Rewriting for Retrieval-Augmented Large Language Models” by Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, Nan Duan.

[12] ”Lift Yourself Up: Retrieval-Augmented Text Generation with Self- Memory ” by Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, Rui Yan.

[13] ”Artificial Intelligence Text Processing Using Retrieval-Augmented Generation: Applications in Business and Education Fields” by Bogdan-Stefan Posedaru, Florin-Valeriu Pantelimon, Mihai-Nicolae Dulgheru, Tiberiu- Marian Georgescu.

[14] ”Leveraging Passage Retrieval with Generative Models for Open Do- main Question Answering” by Gautier Izacard and Edouard Grave.

[15] ”UPRISE: Universal Prompt Retrieval for Improving Zero-Shot Eval- uation” by Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Denvy Deng, Qi Zhang.

[16] ”A Retrieval-Augmented Generation Framework for Academic Litera- ture Navigation in Data Science” by Ahmet Yasin Aytar, Kamer Kaya, Kemal Kilic.