NEET-QA: Evaluating Retrieval Augmented Generation in Competitive Examinations

Dhiren Oswal[1, a)] Soham Tolwala,[1, b)] Pushpak Suryawanshi,[1, c)]
Dhairya Thakkar,[1, d)] and Swapnil Shinde[2, e)]

[1]*Student, Department of Artificial Intelligence and Data Science,*
Vishwakarma Institute of Information Technology, Kondhwa Bk, Pune, 4110037, Maharashtra, India

[2]*Assistant Professor, Department of Artificial Intelligence and Data Science,*
Vishwakarma Institute of Information Technology, Kondhwa Bk, Pune, 4110037, Maharashtra, India

[a)] dhiren.22011106@viit.ac.in
[b)] soham.22010224@viit.ac.in
[c)] pushpak.22010400@viit.ac.in
[d)] dhairya.22011097@viit.ac.in
[e)] swapnil.shinde@viit.ac.in

Abstract. Large Language Models (LLMs) have shown potential in various domains but can struggle with factual accuracy and knowledge gaps because of unseen or outdated data. Retrieval-Augmented-Generation (RAG) offers a solution by integrating information retrieval from reliable sources with LLM generation, leading to improved performance in knowledge-intensive domains. This paper explores the prowess of LLM+RAG for question-answering in specialized domains. We examine the state-of-the-art RAG framework, including retrieval, generation, and augmentation techniques, exemplified by its novel application in the National Eligibility-cum-Entrance Test (NEET), a standard Indian medical entrance exam for medical education, to assess the precision and contextual relevance of the answers generated by RAG workflow. We utilized Gemini-pro and gpt-3.5-turbo-0125 as the core LLMs for our RAG systems, in conjunction with GoogleGenerativeAIEmbeddings and Chroma Vector DB. The National Council of Educational Research and Training (NCERT) textbooks were fed as the primary data source to extend the knowledge base of LLM. The questions were taken from the 2022 NEET paper and the RAG-based systems were evaluated against the given ground truth answers. The GPT-incorporated model scored the highest with an overall 62.5% precision with the highest scores in Biology (71% precision). Through a comprehensive examination of the generated answers and state-of-the-art technologies and frameworks, we further elucidate the potentials and drawbacks of LLMs with RAG and jot down the necessary checkpoints in advancing question-answering systems across specialized domains.

Keywords: Question-answering systems, Large Language Models, Retrieval Augmented Generation Approach, Competitive Examinations, National Eligibility-cum-Entrance Test (NEET)

INTRODUCTION

In recent years, Large Language Models (LLMs) have garnered significant attention for their remarkable abilities in natural language processing tasks. Models such as Mistral, OpenAI's GPT series, Google's Gemini (previously known as Bert), and Meta's Llama have demonstrated proficiency in various domains, from general language understanding to specialized fields like medicine[1], finance[2], business[3], etc. For example, OpenAI's ChatGPT, supported by Microsoft, is famous for its flexibility. It can easily switch between tasks like creating complex codes and composing music. Its abilities aren't limited to just technical or literary tasks; it has also achieved significant success in academics. For instance, in specialized domains, it performed exceptionally well in the MBA program at the University of Pennsylvania and passed the law exam at Minnesota Law School[4]. Moreover, it successfully cleared the United States Medical Licensing Examination (USMLE)[5] in one attempt with 60% accuracy. Normally, aspiring doctors spend nearly four years in medical school and over two years in clinical rotations to pass this exam[6].

But, after a failed attempt at the Union Public Service Commission (UPSC) exam[7], AIM decided to check ChatGPT's medical prowess[8]. The NEET examination, a crucial milestone for aspiring medical professionals, presents a unique set of challenges. With questions spanning diverse topics in physics, chemistry, and majorly in

biology, NEET demands not only factual knowledge but also the ability to comprehend complex concepts and apply them in problem-solving scenarios. So, when ChatGPT took the NEET 2022 exam with 174 (originally 200) questions, it just passed, scoring 50.14%, while struggling particularly in biology[8]. It proves that traditional LLMs often struggle to meet these requirements, facing issues like hallucination, outdated information, and opaque reasoning processes, which can impede their accuracy and reliability [9], [10].

In response to these challenges, Retrieval-Augmented Generation (RAG) has emerged as a promising approach to

enhance the capabilities of LLMs in specialized domains [11], [12], [13], [14]. By integrating knowledge from external sources through retrieval mechanisms, RAG enables LLMs to generate contextually relevant responses grounded in credible information retrieved from an extended curated data source. As RAG has given promising results in complex and fact-dependent domains like biology, there is a lack of evaluation concerning how RAG systems handle questions that are highly complex and require nuanced understanding, synthesis of information from multiple sources, and critical reasoning.

This paper explores the maiden utilization of RAG in the question-answering system for the National Eligibility-cum-Entrance Test (NEET) —a highly competitive undergraduate-level Indian Medical Entrance Exam, known for its rigorous standards and comprehensive coverage of subjects, particularly in medical and engineering fields— with a focus on assessing the accuracy and potential to tailor responses to complex inquiries. When a Large Language Model i.e. ChatGPT wrote the NEET exam individually, it achieved nearly a 50% accuracy rate with low scores, especially in Biology[8]. However, by leveraging RAG techniques, our QA system scored an overall accuracy rate of 61.1%, and a precision rate of 62.5% on the same exam, with the highest scores in the Biology domain (70% accuracy, 71% precision). In addition to highlighting the efficacy of RAG in enhancing LLMs' performance in specialized domains like NEET, this paper explores the findings from the score obtained from GPT and Gemini to help navigate future developments for specialized QA domains. By leveraging the RAG-enabled responses, we seek to provide methodologies in addressing the unique challenges posed by involuted domains like NEET, ultimately contributing to the development of more reliable and robust question-answering systems. Out assessment uses metrics like precision and accuracy, as the NEET answers are objective i.e. have a ground truth. However, an overview of state-of-the-art evaluation methodologies/frameworks for a more subjective comprehensive understanding is provided below.

LITERATURE REVIEW

Large Language Models (LLMs) have emerged as powerful tools for various applications, including Question-Answering (QA) systems. However, specialized domain QA systems with LLMs face drawbacks like hallucinations (generating irrelevant or incorrect information) and limited interpretability of their answers. These limitations can be particularly detrimental in critical domains like education and healthcare [10], [15].

However, several studies advocate for RAG techniques to improve the accuracy and reliability of LLM-based QA systems. To gain a deeper understanding of RAG functionalities, [11] provides a comprehensive survey of Retrieval-Augmented Generation paradigms. They delve into various RAG frameworks like Naive RAG, Advanced RAG, and Modular RAG, outlining their core components (retrieval, generation, augmentation) and evaluation methods. It particularly focuses on discussing various evaluation methodologies, as due to the complexity of LLMs, evaluating their performance poses challenges. Therefore, evaluation methodologies are necessary to provide insights into their effectiveness and limitations. The survey emphasizes three quality scores: context relevance (ensures retrieved information is highly relevant), answer faithfulness (ensures generated answers adhere to retrieved-context), and answer relevance (ensures generated answers directly address the question). These metrics provide a comprehensive framework for evaluating both retrieval and generation components of RAG systems[11].

Applying LLMs in educational contexts, particularly Math Q&A, presents unique challenges like the need for clear explanations and step-by-step solutions, limited ability to address open-ended or novel problems, data bias, and fairness concerns [14]. Research by [14] demonstrates the effectiveness of RAG in mitigating LLM shortcomings such as hallucinations and enhancing answer interpretability. [14] implements a RAG system for a Math chatbot, achieving improved answer accuracy by retrieving relevant passages from a pre-defined corpus and feeding them into an LLM for generation. This highlights the potential of RAG for educational Q&A systems.

While RAG offers significant benefits, there's room for further optimization. [16] explores using Reinforcement Learning (RL) to optimize RAG-based chatbots. Their study achieves a reduction in resource consumption by reducing the number of LLM token calls compared to traditional RAG approaches. This suggests that exploring diverse optimization strategies beyond RAG itself can yield further improvements in Q&A system efficiency [15]. Active learning is another promising optimization strategy for RAG systems. This approach allows the system to identify the most informative context to retrieve for each specific question, potentially leading to improved accuracy and efficiency [17]. Results demonstrate that ACTIVERAG surpasses previous RAG models, achieving a 5% improvement on question-answering datasets[17].

RAG can also be applied to specific QA domains, such as the MIRAGE benchmark and MedRAG toolkit provide valuable resources for researchers and developers working with RAG systems in the medical domain [18]. MIRAGE focuses on real-world medical question-answering scenarios and evaluates RAG systems' ability to answer these questions using zero-shot learning and question-only retrieval for multi-choice tasks. MedRAG, on the other hand, is an easy-to-use toolkit that allows researchers to explore various RAG configurations with different corpora, retrievers, and LLMs [18]. This combination of resources can facilitate the development and evaluation of effective RAG systems for medical QA.

Another paper [19] introduces PaperQA, a novel agent-based architecture for scientific question answering. It utilizes three separate agents: 1) Search Tool: retrieves relevant research papers using keyword search and vector embedding techniques. 2) Gather Evidence Tool: extracts key passages from identified papers, building a context library for answer generation. 3) Answer Question Tool: analyzes the evidence and formulates a final answer with proper citations. This design allows for dynamic adjustments throughout the retrieval and information-gathering process, leading to improved accuracy in QA tasks compared to standard LLMs, enhanced answer grounding through

retrieved evidence and citations, and more accurate and contextually grounded answers compared to standard LLMs. It also introduces the LitQA dataset for PaperQA evaluation. LitQA evaluates the ability both to retrieve necessary information and to present an accurate answer based on that information[19]. While the focus here has been on RAG for QA systems, the potential applications extend to other areas. For instance, RAG can be used for tasks like claim verification, where retrieving relevant evidence is crucial for determining the veracity of a claim [20].

In conclusion, the reviewed literature overwhelmingly supports the potential of Retrieval-Augmented Generation (RAG) for enhancing the capabilities of LLM-based question-answering (Q&A) systems. This is particularly true in educational contexts, specialized QA domains, and broader domains like claim verification where the interpretability of relevant information provided by RAG is crucial. While traditional RAG approaches offer significant benefits, exploring optimization strategies like Reinforcement Learning (RL) and Active Learning holds promise for further improvements in efficiency and accuracy. Furthermore, a comprehensive understanding of RAG frameworks, as outlined in [11], is essential for successful integration within our project.

METHODOLOGY

The Retrieval-Augmented Generation (RAG) workflow initially involves breaking down large amounts of relevant text data (corpus) into smaller pieces and encoding them into a format that allows for efficient searching. When a query is received, the retrieval agent of RAG searches the encoded data for similar pieces based on their content and appends this information to the Large Language Model's (LLM's) information base. Using this extended database, the LLM crafts a response. Ultimately, this process allows RAG to both find relevant information and use it to generate well-informed responses. We will now explore our application of the RAG framework for the NEET QA system in more detail. FIGURE 1. gives the architecture used

User Input

The RAG workflow, illustrated in Figure I, commences with a user query derived from the NEET question paper, posed to an LLM. Here, we have employed the NEET 2022 Question set as the primary source for extracting questions. These questions were carefully selected to ensure relevance to the NEET exam. However, it's important to note that questions containing complex formatted LaTeX text, diagrams, or images were excluded from our dataset. The reason to skip the questions was due to the usage of freely available rudimentary models that only worked on text, and not complex equations or visual content. Once the relevant questions were identified, they were combined with their corresponding ground truth answers in a JSON format. This format allowed us to organize the data efficiently, with each question-answer pair representing a single entry in the dataset. For example: "Which of the following is a characteristic feature of the phylum Porifera? Options: A) Radial symmetry B) Presence of a true coelom C) Cellular level of organization D) Segmented body plan". While ChatGPT serves as a widely used LLM, its knowledge is confined to pretraining data. RAG mitigates this limitation by making use of up-to-date document excerpts from external knowledge repositories, such as NCERT[21] textbooks for physics, chemistry, and biology, Oswal Revision notes[22] for biology, and Concepts of Physics[23] by Prof. H.C. Verma for physics in making the vector store, which encompass all study materials relevant to the NEET exam. These books are most preferred by students for the NEET exam preparation. Hence, by combining questions from the NEET 2022 Question Set with relevant external resources, we created a comprehensive dataset that encompassed all study materials relevant to the NEET exam. This dataset served as the foundation for evaluating our Retrieval-Augmented-Generation (RAG) model, enabling it to generate well-informed responses to user queries derived from the NEET question paper. These retrieved articles, coupled with the initial query, compose an enriched prompt that enables the LLM to generate a well-informed response. This illustration highlights the effectiveness of RAG in enhancing the capabilities of LLMs by integrating external knowledge sources.

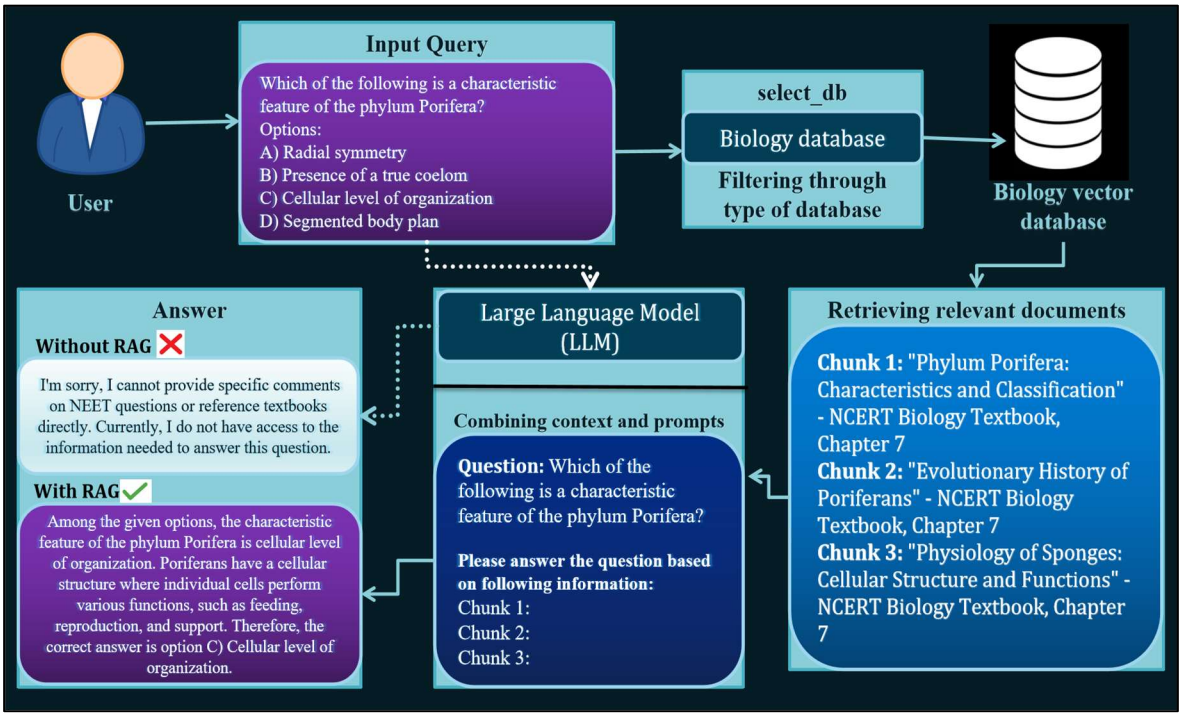

FIGURE 1. Architecture: The RAG process applied to NEET question-answering

Creating Vector Database

This stage is fundamental in preparing the data, serving as the initial step in offline operations. It involves several phases aimed at transforming raw data into a format suitable for analysis. Initially, the data undergoes cleansing and extraction processes. Various file formats such as PDF, HTML, Word, and Markdown which are used for storing or presenting content related to the required domain specialization are converted into standardized plain text to ensure consistency across documents. We have obtained the NCERT e-books from [21] and parsed them to obtain the texts from the textbooks. To address the context-window limitations of language models, this text is then segmented into smaller, manageable chunks through a process known as chunking, which can be done on various criteria like character or limits, specific titles, etc. These chunks are further transformed into vector representations using an embedding model selected for its balance between inference efficiency and model size. Here we use GoogleGenerativeAIEmbeddings. The embeddings are then stored in a database, also known in this case as a vector store as it stores the vector embeddings. When storing text in a vector database like ChromaDB, the text is encoded into numerical vector formats within a high-dimensional space. This transformation enables similarity comparisons during the retrieval phase, facilitated by tools like Chroma vector store's similarity search. We are storing the source of the document as metadata in the vector database i.e. the default schema for storing documents in the vector database. Finally, we create a vector database to store these retrieved document chunks in a vectorized format, enabling efficient and scalable search capabilities. In the process of constructing the vector database, we strategically incorporate three distinct vector databases, each dedicated to a specific subject as shown in Figure II.

This approach is designed to mitigate the overloading of computational resources, thereby enhancing the efficiency and performance of the system. On average, each subject encompasses approximately twenty-five PDF documents, each consisting of 20 pages. This substantial volume of data underscores the complexity and richness of the information associated with each subject. The subsequent sections will delve into the methodologies employed to efficiently process and manage this data. The entire vector database creation process, illustrated in Figure II, lays the groundwork for efficient retrieval and processing of information necessary for tasks like Q&A queries in large documents or document libraries, particularly in the context of Retrieval-Augmented Generation (RAG). Through innovative approaches such as context-aware chunking and tools like DocumentLoader and TextSplitter, meaningful document segmentation is achieved, leading to improved outcomes in RAG applications.

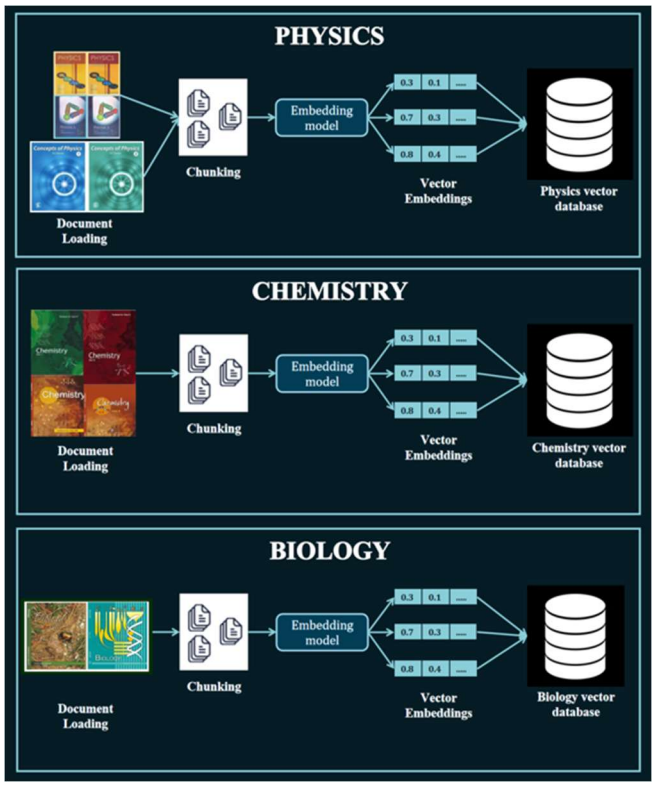

FIGURE 2. The vector database creation process

Retrieving Relevant Documents

Upon receiving a user query, the system leverages the encoding model previously utilized during the indexing phase to encode the query into a vector representation. Subsequently, the system finds relevant chunks by calculating similarity scores between the query vector and the vectorized chunks stored within the vectorized corpus using the Chroma vector store's similarity search. Our system retrieves the top K chunks demonstrating the highest similarity to the query. These selected chunks serve as the expanded contextual basis for addressing the user's request, ensuring that the synthesized response is grounded in relevant information retrieved from the indexed corpus. This process enables the system to provide accurate and informed responses tailored to the user query, enhancing the overall effectiveness of the Retrieval Augmented Generation (RAG) framework.

Answer Generation

In our methodology, the process of generating answers involved several steps to ensure accuracy and relevance. Initially, questions from the NEET 2022 Q1 question set were iteratively processed through a prompt instance. Each question served as a query, and this query was passed to the model, including both the Google Gemini-pro and GPT-3.5-turbo-0125 large language models (LLMs), sequentially. Upon receiving a query, the model underwent a series of operations. The query was embedded and then subjected to vector search to retrieve the top K documents from the corpus. In our case, K was set to 1, ensuring that only the most relevant document was retrieved. Once the context from the retrieved document was obtained, it was combined with the original prompt template corresponding to the subject area (e.g., biology template).

The combined prompt, enriched with contextual information, served as the basis for generating the response. The model then synthesized a well-informed answer using the integrated knowledge from both the prompt and the retrieved document.

In instances where the model failed to find an answer or responded with "Sorry, I don't have much information about it," we implemented a fallback strategy. Specifically, we switched to the RAG-free mode, allowing the model to generate a response based solely on its internal knowledge base. This iterative process ensured that each question was thoroughly processed and that the generated responses were contextually relevant and accurate to the best extent possible, leveraging both the internal knowledge of the LLMs and the additional information

retrieved through the RAG framework.

Moreover, the challenges encountered during the initial testing phase with GPT-3.5-turbo-0125 highlighted the necessity for the adoption of RAG. These challenges included restricted access to specialized domain knowledge and the inability to interpret complex queries, which underscored the importance of leveraging external knowledge sources through RAG to enhance the model's performance.

```
prompt_template_bio = """
You are a question answering bot which is specialized
in BIOLOGY. Answer the following Multiple Choice based
question which has one single correct answer from the
four options given in the question itself. you have
to accurately choose the correct option and return it.
Choose the answer according to the context provided
below.
GOLDEN RULE: You will be evaluated by your predicted
option and hence focus on returning only the option.
Context: {context}
Question: {question}
Don't answer if the context provided to you is
irrelevant, if you are not sure and confident and
decline to answer and say "Sorry, I don't have much
information about it."
Include a brief reason for why you chose a particular
option.
Answer:
"""
```

EXPERIMENTS

Experimental Details

Our experiments were carried out on two LLMs, one on Gemini-pro and another on gpt-3.5-turbo-0125. The model was tested on a selection of questions from the NEET 2022 code (Q1) paper across physics, chemistry, and biology. It's important to note that we exclusively included questions composed in standard text ASCII characters, omitting those containing equations, charts, or graphs, and complex latex texts due to the inherent complexities they introduce. The experiment comprised 42 physics questions out of 50, 39 chemistry questions out of 50, and 99 biology questions out of 100 given questions.

Experimental Details

In the context of our research study, the results were calculated by checking our predicted answer against the ground truth available in the NEET answer booklet. For each multiple-choice question with 4 options, a single correct choice was available. To evaluate the results, we have taken inspiration from the standard ML metric - accuracy and tweaked it from the conventional machine learning formula typically used in evaluating machine learning models, because we have a deterministic (i.e. single-class) prediction. While the standard accuracy formula for binary class prediction incorporates true positives, true negatives, false positives, and false negatives, our approach focuses on a simpler metric: the ratio of true positives i.e. total correct answers to total questions, expressed as a percentage, as we do not possess the false negative values. As NEET penalizes incorrect answers, we have included precision, where incorrect answers are considered false positives.

Our equations for accuracy and precision are as follows:

$$Accuracy = \frac{True\ Positives}{Total\ questions} \quad (1)$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (2)$$

By defining accuracy as the ratio of correct answers to total questions attempted, we aimed to provide a straightforward measure of the model's performance relative to the questions presented. This metric allows for a

direct assessment of the model's effectiveness in answering NEET exam-style questions without the need for complex evaluation formulas that may not align with the nature of LLM-based systems.

Moreover, it's important to note that this research marks the first attempt at evaluating the performance of LLM+RAG models in the context of NEET question-answering systems, and as such, there is no official comparison available for reference. However, insights from previous experiments based solely on LLMs, such as those reported by Analytics India Magazine[8], indicate that models like ChatGPT achieved a performance level of approximately 50% in similar assessments, with poor scores in botany because of the unavailability of precise knowledge.

Table I displays the outcomes, and Table II displays the metrics evaluation of our answer-generation phase. The results from both models show a close comparison, with GPT slightly outperforming. Notably, the GPT-based QA system managed to achieve precision rates of 50% and above for all three sections and a 62.5% overall precision. Particularly in Physics, GPT not only provided superior answers but also abstained from hunching to prevent incorrect responses. Remarkably, both models excelled in Biology questions, achieving an accuracy and precision rate of nearly 70%, a significant improvement compared to GPT's previous performance in this subject [8].On the other hand, the overall performance QA system, especially for Gemini was below average in physics, which had a higher number of numerical and application-based question.

The LLM models operate majorly on probabilities, generating responses based on their pre-trained knowledge. Evaluating LLM performance using traditional accuracy metrics may not accurately reflect its capabilities or limitations. Hence, with our results, we do not assess the LLMs but the question-answering system developed over the LLMs. LLMs, particularly when augmented with RAG, exhibit nuanced behaviour influenced by factors such as the quality and relevance of external knowledge sources, the complexity of the queries, and the model's inherent biases and limitations. In this particular use case, LLM's decisive power can be tested by understanding the patterns of confidence level of predicted answers for a particular question.

Table I:Results for NEET 2022 Q1 Paper

		CORRECT	WRONG	UNANSWERED
GEMINI	PHY	12	30	0
	CHEM	19	20	0
	BIO	69	21	9
	OVERALL	100	71	9
GPT	PHY	20	20	2
	CHEM	20	18	1
	BIO	70	28	1
	OVERALL	110	66	4

Table II: Metric scores for NEET 2022 Q1 Paper

		GEMINI	GPT
PRECISON	PHY	0.28571429	0.5
	CHEM	0.48717949	0.52631579
	BIO	0.76666667	0.71428571
	OVERALL	0.59064327	0.625
ACCURACY	PHY	0.28571429	0.47619048
	CHEM	0.48717949	0.51282051
	BIO	0.6969697	0.70707071
	OVERALL	0.55555556	0.61111111

Discussion

When RAG is employed alongside LLMs, the effectiveness of the model hinges significantly on the quality of the vector store, which directly influences the efficiency of the retrieval process. For instance, our observations indicate that the accuracy of responses varied across subjects, with biology exhibiting higher accuracy compared to physics. This discrepancy can be attributed to differences in the quality of the data present in the vector store. In the case of biology, where the vectors were built from comprehensive and highly dense information, the RAG framework could retrieve relevant context more effectively, resulting in more accurate responses.

Conversely, in subjects like physics, where textbooks possessed overlong passages for topics, the retrieval process encountered challenges, leading to decreased accuracy in responses. This highlights the critical role of robust and comprehensive knowledge databases accessible to RAG in ensuring the accuracy and reliability of LLM-based question-answering systems. Our system utilizes Chroma vector store's similarity search mechanism to compute similarity scores between the query vector and the vectorized chunks stored within the corpus. This process aims to retrieve relevant text from the corpus based on the query and provide pertinent information to the model. However, in domains like physics and chemistry, where the information is predominantly presented in the form of indirect contexts like lengthy derivations in physics, the effectiveness of similarity search may be compromised. The challenge arises from the fact that formulas and key concepts often get intertwined within these derivations, making it difficult to pinpoint specific sections for retrieval. Moreover, the complexity of physics questions further exacerbates this issue, as the relevance of retrieved chunks may not always align with the intended context of the query. Furthermore, when utilizing RAG with LLM, the autonomy of the LLM's intelligent agent may be compromised. Consequently, there may be instances where LLMs without RAG outperform RAG-equipped LLMs, as the responses generated by the latter are influenced primarily by the retrieved content of the retrieval agent rather than the inherent intelligence of the LLM itself. This underscores the need for careful consideration of the balance between RAG augmentation and the LLM's native capabilities to optimize performance.

Pre-processing queries: In specialized standardized exams, not all questions are straightforward. Such twisted questions might require understanding over multiple layers to reach the crux of the question’s meaning. For such questions, extensive analysis by breaking down the query appropriately to extract the questions is necessary.

Optimizing retrievers: In the discussion section, we highlighted a limitation related to the effectiveness of similarity search in retrieving relevant information in domains like physics. To address this limitation, alternative search methods, such as keyword search, can be explored. Unlike similarity search, which relies on content similarity, keyword search can effectively retrieve relevant formulas and concepts based on specific keywords or terms present in the query while considering academic question-answering systems. Moreover, a combination of different searches in retrieval methods can assist in providing accurate information to the LLM chain.

Post-processing relevant documents: Summarization or contextual refinement, can be applied to the retrieved chunks to distill the information and extract the required data from the most relevant retrieved content. This post-processing step can help streamline the information retrieved from the corpus, making it more concise and aligned with the user query. By integrating these alternative search methods and post-processing techniques, the system can enhance its ability to retrieve and present relevant information, particularly in domains characterized by complex, indirect, and redundant contexts.

Integration of Vision Language Models (VLMs): Incorporating VLMs into the RAG framework will enable the system to handle questions containing images and equations. By leveraging VLMs, the model can interpret visual information, such as diagrams, charts, and graphs, commonly found in educational materials like NCERT textbooks. This enhancement will broaden the scope of questions that the system can effectively answer, providing more comprehensive assistance to users preparing for exams like NEET.

Enhanced Multimodal Capabilities: Beyond images and equations, future iterations of RAG can explore incorporating various methods to extract tables, complex equations, diagrams, and other visual representations, along with methods to map them into vector databases. By seamlessly integrating these diverse forms of information, the system can offer a more immersive and interactive educational environment, catering to diverse learning preferences and enhancing comprehension.

Domain Expansion and Specialization: Expanding the application of RAG by using vivid study resources for NEET and combining multiple chunks for cross topic knowledge bank can assist in answering questions having multiple layers of information.

CONCLUSION

Our research analyses the question-answering system with LLMs in integration with Retrieval-Augmented Generation (RAG) techniques with a maiden attempt at Indian Competitive examinations like NEET (National Eligibility-cum-Entrance Test). By leveraging external knowledge sources such as NCERT textbooks and adopting innovative methodologies like context-aware chunking and similarity-based retrieval, our proposed RAG framework demonstrates significant scores —though not human-like— in accuracy and relevance compared to traditional LLM approaches. Our system achieved more than 60% in terms of accuracy and precision rates, with commendable performance in Biology for which GPT individually couldn't answer well, showcasing its efficacy in providing contextually relevant and accurate responses to user queries and underscoring the importance of augmenting LLMs with external knowledge sources to overcome inherent limitations such as outdated information and lack of domain expertise. It also revealed challenges, such as limitations in the quality of available knowledge and retrieving methods, which impacted the model's performance, particularly in subjects like physics.

Furthermore, our study highlights the limitations of RAG systems and the necessary techniques to bridge them such as simplifying queries through pre-processing, formulating advanced retrieval methods, and post-processing retrieved documents which can also be considered as key points to design new RAG-based chatbots.

In essence, our study contributes to the growing body of literature on the integration of external knowledge sources with LLMs, providing valuable insights into the potential applications and benefits of RAG techniques in specialized domain contexts.

ACKNOWLEDGMENTS

We take this opportunity to thank the Head of the Department **Dr. P.N.Mahalle** and our project guide **Prof. Swapnil Shinde** for their valuable guidance and for providing all the necessary facilities, which were indispensable in the completion of this project report. We are also thankful to all the staff members of the Department of Artificial Intelligence and Data Science, VIIT, Pune for their valuable time, support, comments, suggestions and persuasion.

REFERENCES

1. H. Jung et al., "Enhancing Clinical Efficiency through LLM: Discharge Note Generation for Cardiac Patients," pp. 1–10, 2024, [Online]. Available: http://arxiv.org/abs/2404.05144
2. H. Zhao et al., "Revolutionizing Finance with LLMs: An Overview of Applications and Insights," pp. 1–37, 2024, [Online]. Available: http://arxiv.org/abs/2401.11641
3. E. Haaralahti, "Utilization of local large language models for business applications," 2024.
4. S. M. Kelly, "ChatGPT passes exams from law and business schools," CNN Business, 2023. [Online]. Available: https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html
5. J. Lubell, "ChatGPT passed the USMLE. What does it mean for med ed?," American Medical Association(AMA), 2023. [Online]. Available: https://www.ama-assn.org/practice-management/digital/chatgpt-passed-usmle-what-does-it-mean-med-ed
6. L. Varanasi, "ChatGPT is on its way to becoming a virtual doctor, lawyer, and business analyst. Here's a list of advanced exams the AI bot has passed so far.," Business Insider. [Online]. Available: https://www.businessinsider.in/tech/news/chatgpt-is-on-its-way-to-becoming-a-virtual-doctor-lawyer-and-business-analyst-hereaposs-a-list-of-advanced-exams-the-ai-bot-has-passed-so-far-/slidelist/97388435.cms
7. D. Sharma, "ChatGPT passed Wharton's MBA but failed UPSC prelims: here is what you need you know," India Today, New Delhi, Mar. 06, 2023. [Online]. Available: https://www.indiatoday.in/technology/news/story/chatgpt-passed-whartons-mba-but-failed-upsc-prelims-here-is-what-you-need-you-know-2343171-2023-03-06
8. S. Saha, "ChatGPT Takes NEET; Will it Pass with Flying Colors or Flunk?," AI Trends & Future. [Online]. Available: https://analyticsindiamag.com/chatgpt-takes-neet-will-it-pass-with-flying-colors-or-flunk/
9. Dr. Sajjad Mahmood, "Understanding the Limitations of Language Models." [Online]. Available: https://www.linkedin.com/pulse/understanding-limitations-language-models-dr-sajjad-mahmood-zslsf/
10. E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," Learn. Individ. Differ., vol. 103, p. 102274, Apr. 2023, doi: 10.1016/J.LINDIF.2023.102274.
11. Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.10997
12. X. Du and H. Ji, "Retrieval-Augmented Generative Question Answering for Event Argument Extraction," Proc. 2022 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2022, no. 1, pp. 4649–4666, 2022, doi: 10.18653/v1/2022.emnlp-main.307.
13. P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," Adv. Neural Inf. Process. Syst., vol. 2020-Decem, no. NeurIPS, 2020.
14. Z. Levonian et al., "Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference," Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.03184
15. C. Wu, W. Lin, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "PMC-LLaMA: Towards Building Open-source Language Models for Medicine," 2023, [Online]. Available: http://arxiv.org/abs/2304.14454
16. M. Kulkarni, P. Tangarajan, K. Kim, and A. Trivedi, "Reinforcement Learning for Optimizing RAG for Domain Chatbots," Jan. 2024, [Online].

Available: http://arxiv.org/abs/2401.06800
17. Z. Xu et al., "ActiveRAG: Revealing the Treasures of Knowledge via Active Learning," 2024, [Online]. Available: http://arxiv.org/abs/2402.13547
18. G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking Retrieval-Augmented Generation for Medicine," 2024, [Online]. Available: http://arxiv.org/abs/2402.13178
19. J. Lála, O. O'Donoghue, A. Shtedritski, S. Cox, S. G. Rodriques, and A. D. White, "PaperQA: Retrieval-Augmented Generative Agent for Scientific Research," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.07559
20. H. Liu et al., "Retrieval augmented scientific claim verification," JAMIA Open, vol. 7, no. 1, 2024, doi: 10.1093/jamiaopen/ooae021.
21. Textbooks PDF. [Online]. Available: https://ncert.nic.in/textbook.php
22. NEET (UG) Notes Biology. Oswal Publications. [Online]. Available: https://oswaalbooks.com/pages/neet-ug-notes-biology
23. H.C Verma, Concepts of Physics. [Online]. Available: https://hcverma.in/books