

ExamAce: Leveraging LLMs and RAG for Intelligent Academic Assistance

Shakthivel Arumugam
Artificial Intelligence
& Data Science

SIES Graduate School of Technology
Navi Mumbai, India
shaktivel888@gmail.com

Siddhesh More
Artificial Intelligence
& Data Science

SIES Graduate School of Technology
Navi Mumbai, India
mosiddhesh@gmail.com

Shivraj Murali
Artificial Intelligence
& Data Science

SIES Graduate School of
Technology
Navi Mumbai, India
shivrajmurali504@gmail.com

Arsh Pawar
Artificial Intelligence
& Data Science

SIES Graduate School of Technology
Navi Mumbai, India
pawar.arsh1333@gmail.com

Prof Rizwana Shaikh
Computer Engineering

SIES Graduate School of Technology
Navi Mumbai, India
rizwanas@sies.edu.in

Prof Deepali Jagtap
Artificial Intelligence
& Data Science,

SIES Graduate School of
Technology
Navi Mumbai, India
deepaligi@sies.edu.in

Abstract— Engineering students face challenges in accessing reliable information due to the proliferation of unverified online resources and the potential for inaccuracies in Large Language Model (LLM) outputs. This paper addresses this issue by presenting a novel question answering system that leverages a vector datastore of researched reference books and an LLM for generating responses. A key contribution of this work is the introduction of a confidence score metric designed to evaluate the relevance of the retrieved context to the user's query. This metric employs a combination of keyword matching algorithms, short-term filtering methods, and key term frequency analysis to ensure that the LLM is provided with pertinent information, thereby enhancing the accuracy and reliability of the generated answers. The paper analyses the limitations of existing answer evaluation methods in assessing context relevance and highlights the novelty of the proposed confidence score metric in its direct, interpretable, and keyword-focused approach. The potential impact of this research lies in improving the quality of information retrieval for engineering students, fostering deeper understanding, and mitigating the risk of misinformation in their learning process.

Keywords— Question Answering Systems, Large Language Models (LLMs), Vector Datastore, Context Relevance Assessment, Confidence Score Metric, Undergraduate Engineering Education, Reliability, Accuracy, Keyword Matching, Short-Term Filtering, Term Frequency Analysis, Retrieval Augmented Generation (RAG), Educational Technology, Information Retrieval, Answer Evaluation

I. INTRODUCTION

The academic journey of undergraduate engineering students increasingly involves navigating a vast landscape of online resources for learning fundamental concepts and tackling complex problems. While the internet offers unprecedented access to information, it also presents a significant challenge: discerning the credibility and reliability of sources. The ease with which inaccurate or unverified information can be found poses a considerable risk to students' understanding and the quality of their academic work. Furthermore, when students turn to advanced tools like Large Language Models (LLMs) for quick answers, they face the additional hurdle of potential inaccuracies

or fabrications, often referred to as hallucinations, generated by these models.

To address the critical need for reliable information in engineering education, this paper explores the development and evaluation of a question answering system grounded in a curated collection of researched reference books. This approach aims to mitigate the problems associated with unreliable online sources and the inherent limitations of LLMs when operating without a specific knowledge base. The system leverages a vector datastore to enable efficient retrieval of relevant information from these authoritative texts, and an LLM to synthesize coherent and contextually appropriate answers to students' queries. A core innovation of this project lies in the implementation of a novel confidence score metric. This metric is specifically designed to evaluate the relevance of the retrieved context from the reference books to the user's question, ensuring that the information provided to the LLM is pertinent and directly addresses the student's needs.

This research paper delves into the design and implementation of ExamAce, detailing its fine-tuning process, multi-agentic LLM framework, and evaluation methodology. The study also highlights the app's potential to revolutionize exam preparation, offering a structured, efficient, and reliable solution to meet the evolving needs of engineering students.

Motivation

1. The motivation behind this project is to bridge these gaps by creating an intelligent study assistant that leverages advanced AI techniques to offer personalized exam preparation support. By using fine-tuned Language Learning Models (LLMs) and a Retrieval-Augmented Generation (RAG) pipeline, this app aims to streamline the study process, providing students with concise and contextually relevant answers, lecture notes, topic summaries, and important questions.[4]
2. The goal is to not only enhance the quality of study materials but also save valuable time for students by delivering focused, exam-oriented content. This app

empowers students to prepare more effectively, ensuring they can grasp complex concepts, strengthen their understanding, and ultimately improve their academic performance. In a competitive academic environment, this tool can significantly reduce the stress of exam preparation and make quality education more accessible to all engineering students.[3]

II. LITERATURE SURVEY

A review of existing research on context-based relevance assessment in LLM-based question answering systems reveals a growing recognition of its importance. Several approaches have been explored to evaluate how well the retrieved context supports the generation of accurate and relevant answers. Some methods focus on using LLMs themselves to assess the relevance of the context given a question. For instance, the context-relevance assertion in promptfoo utilizes an LLM to break down the context into statements and evaluate the relevance of each statement to the query, ultimately calculating a relevance score based on the proportion of relevant statements. Similarly, Uptrain's Context Relevance evaluation determines if the extracted context can completely answer the query, give some relevant but incomplete answer, or contains no information to answer the query. These LLM-based evaluations often rely on prompt engineering to guide the model in making accurate relevance judgments.

Other approaches involve using metrics that compare the question and the context. For example, RAGAS (Retrieval Augmented Generation Assessment) employs a Response Relevancy metric that generates artificial questions based on the response and then computes the cosine similarity between the embeddings of these generated questions and the embedding of the original user input to determine relevance. This method operates on the principle that if the answer is relevant to the question, it should be possible to reverse-engineer the question from the answer.

While these existing methods provide valuable ways to assess context relevance, they often differ from our proposed approach in their reliance on semantic embeddings or the use of another LLM for evaluation. Our confidence score metric, in contrast, emphasizes a more direct, keyword-based analysis combined with short-term filtering and term frequency. While some related work might incorporate keyword analysis as part of a broader evaluation strategy, the specific combination of these three techniques, tailored for the undergraduate engineering education domain, appears to be a novel aspect of our work. For example, some systems might use BM25 scores, which consider term frequency and inverse document frequency, as part of their relevance ranking. However, the explicit inclusion of short-term filtering to refine the keyword analysis and the specific combination of these elements for the purpose of generating a confidence score for context relevance evaluation distinguishes our approach. By focusing on the explicit lexical overlap and the frequency of key terms, our metric offers a potentially more transparent and interpretable way to assess the immediate relevance of the retrieved context, which can be particularly beneficial in an educational setting where understanding the connection between the source material and the answer is important for learning.

Existing Systems

1. Traditional Learning Management Systems (LMS): Blackboard, Moodle, Canvas.

Limitations: Often lack personalized learning paths and intelligent answer generation capabilities. Primarily focus on content delivery and assignment submission.

2. Online Tutoring Platforms: Chegg, Khan Academy.

Limitations: Can be expensive, may not always provide contextually relevant answers specific to a student's curriculum. Answers might lack depth or come from unverified sources.

3. AI-powered Chatbots: Solutions integrated into some LMS or standalone apps.

Limitations: As noted by Smutny & Schreiberova (2020), many educational chatbots are in early stages and lack the sophistication to serve as true AI teaching assistants. Often rely on pattern matching rather than deep understanding.

4. General-Purpose LLMs: ChatGPT, Bard, etc.

Limitations: While powerful, they can hallucinate answers (provide incorrect information), lack specific domain knowledge, and don't always provide answers tailored to a particular course or institution.

Limitations of Current Systems:

- Lack of Personalization:** Most systems offer generic content, failing to adapt to individual learning styles, paces, or specific curriculum requirements.
- Inaccurate or Irrelevant Information:** General-purpose LLMs can provide incorrect or misleading answers, especially in specialized domains. Content from online tutoring platforms may be unverified.
- Limited Contextual Understanding:** Many systems struggle to understand the specific context of a student's question, leading to generic or irrelevant responses. They may not consider the student's course, professor's teaching style, or specific lecture notes.

Proposed System

1. Fine-Tuned LLMs for Accuracy and Relevance:

ExamAce utilizes state-of-the-art Large Language Models (LLMs) such as Llama 3.1, Llama 2, and Mistral, which are fine-tuned on carefully curated datasets. These datasets include lecture notes, past exams, and presentations from professors, ensuring that the responses generated are not only accurate but also contextually relevant to the specific subjects being studied. This targeted fine-tuning process enhances the model's ability to provide subject-specific answers, addressing a common limitation of generic LLMs that may not understand nuanced academic content.

- Retrieval-Augmented Generation (RAG) for Knowledge Integration:** The implementation of a Retrieval-Augmented Generation (RAG) pipeline is a cornerstone of ExamAce's approach to knowledge integration. This system allows ExamAce to retrieve pertinent information from a vector database, specifically FAISS (Facebook AI Similarity Search), which stores embeddings created using Nomic Embedding. Research by Lewis et al. (2020) highlights

the benefits of RAG in enhancing the relevance and accuracy of generated content by combining retrieval with generative capabilities.

3. **Multi-Agentive Framework for Enhanced Response Quality:** ExamAce employs a multi-agentive framework, where multiple LLMs collaboratively generate responses, significantly enhancing accuracy and completeness for engineering education. Unlike traditional single-model systems, which are simpler but suffer from limited perspective, higher hallucination risk, and poor performance on unfamiliar topics, ExamAce leverages the collective strengths of multiple models. This multi-agentive design offers key advantages: improved accuracy through cross-validation, broader domain coverage, and greater response completeness. While it introduces increased system complexity, higher resource consumption, and slightly higher latency compared to single-model systems, these trade-offs are justified by the critical need for precision in engineering education.
4. **Focus on Engineering Education:** Designed specifically for engineering students, ExamAce addresses their unique academic challenges. The platform is tailored to support complex problem-solving and technical questions that are prevalent in engineering disciplines.
5. **Continuous Improvement Through Feedback:** A planned feedback system enables ExamAce to refine its model outputs continually based on student reviews and interactions. This iterative process ensures that the platform evolves alongside user needs, enhancing its effectiveness over time.

III. PROPOSED METHODOLOGY

The ExamAce system architecture is designed to seamlessly integrate data preprocessing, retrieval, and generation to provide accurate and relevant answers to students' queries. The architecture comprises two key stages: Data Preprocessing and Data Retrieval and Generation, supported by advanced models and tools for efficient functioning.

(A) Architecture

1. **Preprocessing of Data:** The first stage focuses on preparing the data to ensure high-quality responses. The system utilizes a diverse range of academic resources, including reference books, lecture notes, professor presentations (PPTs), and previous year exam questions. These resources are processed through the following steps:
 - **Text Embedding:** The textual data is transformed into vector embeddings using the nomic-embed-text model, ensuring the semantic meaning of the content is captured effectively. The textual data is transformed into vector embeddings using Binary and Scalar Embedding Quantization techniques. These methods significantly improve retrieval speed and reduce computational costs by efficiently encoding the semantic meaning of the content, enabling faster and more cost-effective similarity-based searches.

- **Vector Database:** The embeddings are stored in a scalable vector database, FAISS (Facebook AI Similarity Search), which allows for efficient and accurate similarity-based retrieval during the query process.[7]
- Building upon the choice of FAISS, it's important to understand its significance in the landscape of vector databases. While cloud-based alternatives like Pinecone and MongoDB Atlas offer managed vector storage, FAISS provides a powerful, local alternative that can lead to fast retrieval. By storing FAISS databases directly within the project, network latency is eliminated, potentially accelerating query times. This makes the selection of FAISS crucial due to its efficiency, the flexibility it offers in choosing indexing algorithms, and the level of control it provides over the retrieval process. Although FAISS is particularly well-suited for local, high-performance retrieval, other methods exist, including brute-force search, alternative Approximate Nearest Neighbours (ANN) libraries such as Annoy and NMSLIB, and graph-based approaches like HNSWlib. Each of these methods presents different trade-offs between search speed and accuracy. Ultimately, FAISS stands out for its optimized performance and the customization capabilities it offers when local control and speed are paramount.
- This preprocessing ensures that the system has a robust and well-structured knowledge base, capable of retrieving relevant information for diverse academic queries.[6]

2. Data Retrieval and Generation

The second stage involves query processing and answer generation, leveraging a Retrieval-Augmented Generation (RAG) pipeline. The steps are as follows:

- **Query Embedding:** When a student submits a query, it is converted into a vector embedding using the same embedding model. This ensures consistency with the pre-processed data.
- **Relevant Data Retrieval:** The system searches the FAISS vector database for the most relevant embeddings corresponding to the input query. The retrieved data is then used to construct a meaningful prompt.
- **Multi-Agent LLM Response Generation:** The constructed prompt is processed by a multi-agent system of fine-tuned Language Learning Models (LLMs), including Llama 3.1, Llama 2, and Mistral. Each model contributes to generating a detailed and accurate response.
- **Final Answer Selection:** The outputs from the models are combined and evaluated for accuracy and relevance before delivering the final answer.

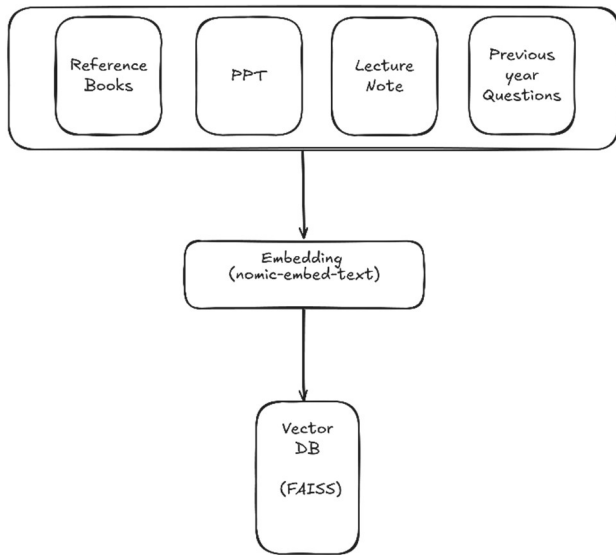


Fig 1: Preprocessing of Data

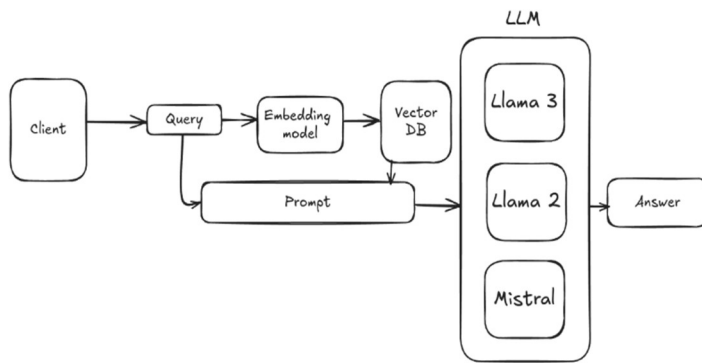


Fig 2: Data retrieval and generation

IV. RESULT

METEOR Score

- METEOR (Metric for Evaluation of Translation with Explicit Ordering) is an automatic evaluation metric designed primarily for machine translation.
- Unlike simpler metrics like BLEU, METEOR considers semantic similarity by incorporating synonyms, stemming, and word order.
- It was specifically developed to address limitations in other metrics by better correlating with human judgments of quality.
- METEOR is particularly effective for evaluating natural language generation tasks, including summarization and question-answering.
- The metric ranges from 0 to 1, where 1 indicates a perfect match between generated and reference texts.

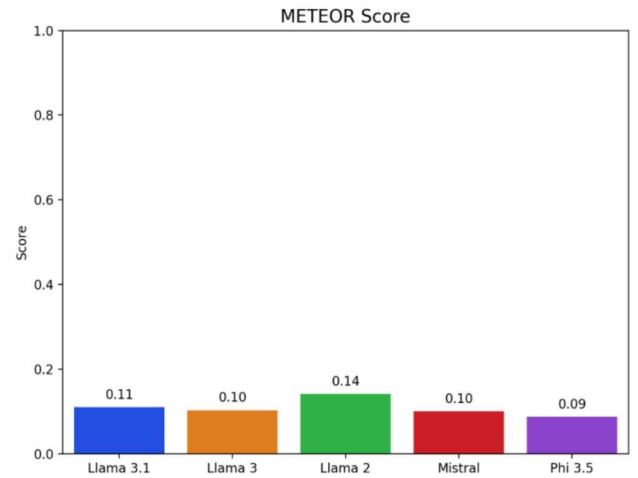


Fig 3: METEOR SCORE

ROUGE Score

- ROUGE is a set of metrics used to evaluate the quality of generated text by comparing it with reference text(s). It is widely used for assessing summarization, machine translation, and text generation tasks in large language models (LLMs).

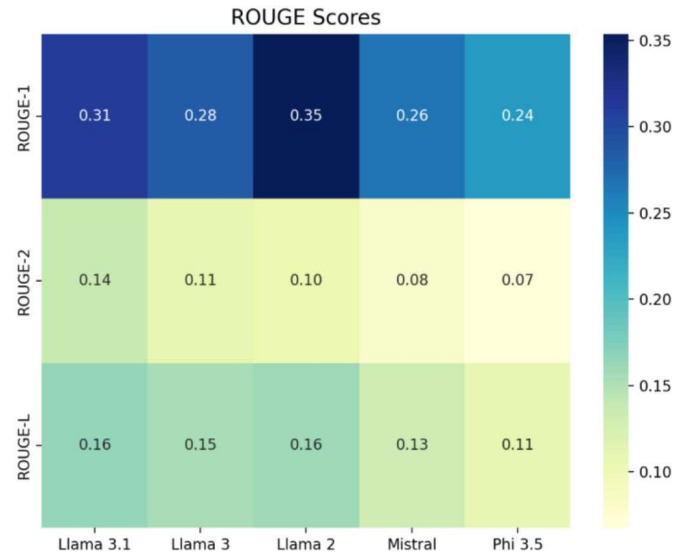


Fig 4. Rouge Score

Cosine Similarity

- Cosine Similarity is a metric used to measure the semantic similarity between two text embeddings. It is widely used in large language models (LLMs) for tasks like information retrieval, text matching, and document ranking.
- Cosine Similarity is mainly used for semantic search to rank relevant documents, helping in text clustering and paraphrase detection efficient for vectorized text comparison in high-dimensional spaces.

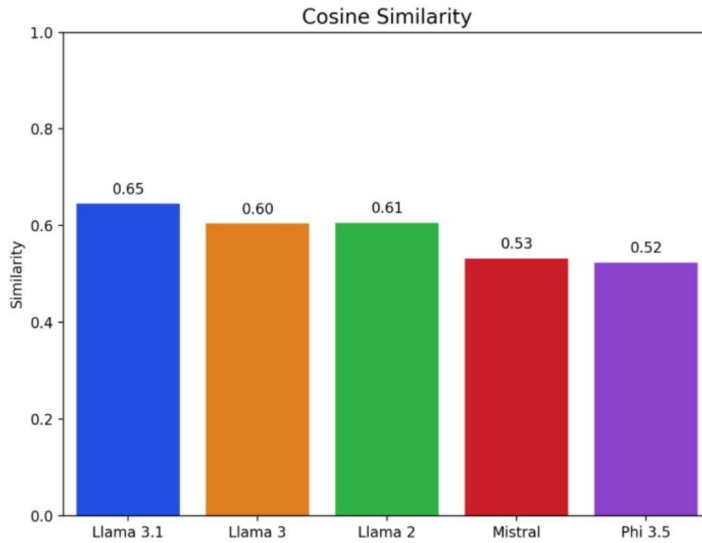


Fig 5. Cosine Similarity

Coherence Score

- Coherence Score is a metric used to evaluate the logical consistency and topic alignment of generated text. It is commonly applied in topic modeling and language model evaluation to assess whether the output text maintains a clear and meaningful structure.
- Coherence assessment is particularly vital in educational applications where the logical flow and connectedness of ideas significantly impact student comprehension and knowledge retention.
- In the context of ExamAce coherence score ensures that the answers generated by LLMs (Llama 3.1, Llama 3, Llama 2, and Mistral) maintain a consistent logical flow.
- It also enhances readability and usability of AI-generated responses for students.

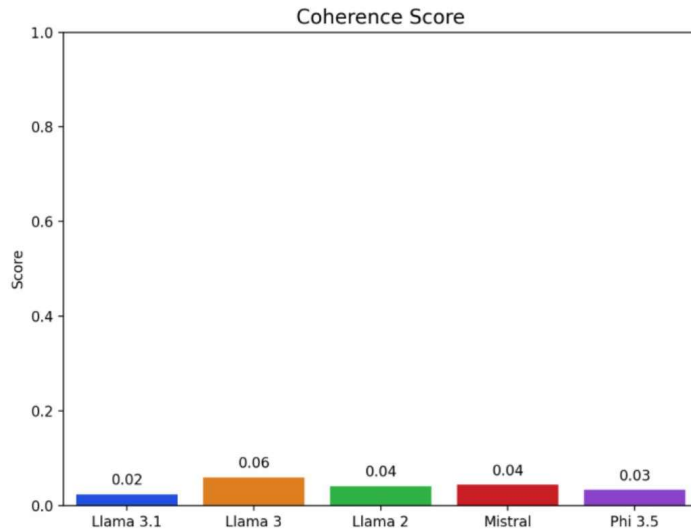


Fig 6. Coherence Score

Model Fit Score

- Model Fit Score measures how well a model's predictions match the actual outcomes, indicating the accuracy and reliability of the model. It is commonly used in regression, classification, and machine learning models to assess performance.
- Model fit in ExamAce
 - Evaluates the performance of fine-tuned LLMs (Llama 3.1, Llama 3, Llama 2, and Mistral).
 - Helps determine how well the multi-agent framework and RAG pipeline retrieve and generate relevant academic content.
 - Ensures that generated responses align closely with expected answers, improving learning outcomes for students.

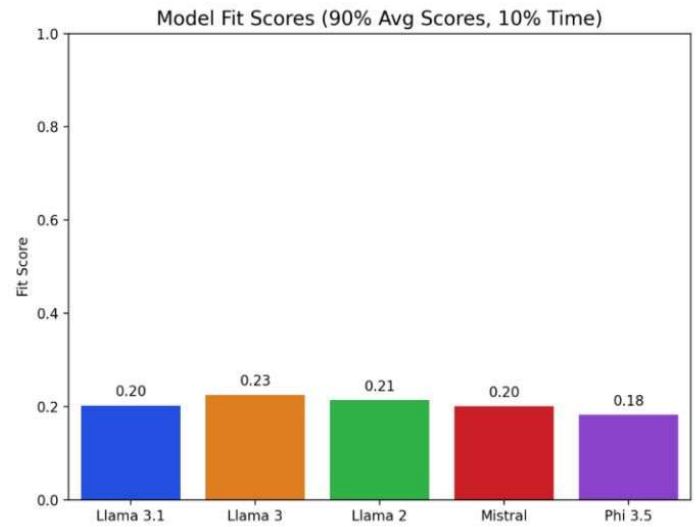


Fig 7. Model Fit Score

Accuracy Score

- Accuracy Score is a key metric used to evaluate the performance of classification models by measuring the proportion of correctly predicted instances. It is widely used in machine learning, NLP, and AI-generated content evaluation.
- Accuracy score assesses the correctness of AI-generated responses in comparison to reference answers, helps fine-tune the multi-agent LLM framework to improve output quality.
- This ensures that retrieval-augmented generation (RAG) fetches and generates relevant and precise academic content.

Formula for Accuracy Score

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

- A higher accuracy score (closer to 1 or 100%) indicates better model performance.
- Lower accuracy suggests errors in classification or content generation, requiring fine-tuning.

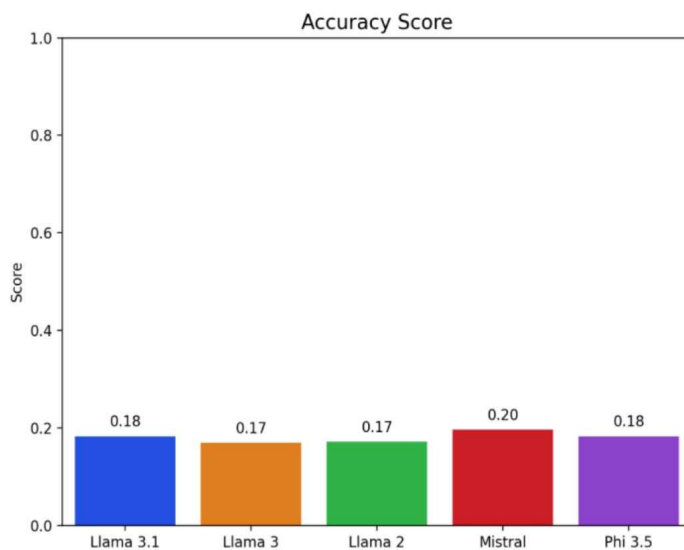


Fig 8. Accuracy Score

V. CONCLUSION AND FUTURE WORK

This paper has presented a novel approach to enhancing the reliability of question answering systems for undergraduate engineering students through the implementation of a confidence score metric that evaluates the relevance of the retrieved context. The increasing reliance on online resources and the potential for misinformation, coupled with the limitations of LLMs such as hallucinations, necessitate robust mechanisms for ensuring the accuracy and pertinence of information provided to students. Our proposed system, which leverages a vector datastore of researched reference books and an LLM for answer generation, addresses these challenges by incorporating a unique confidence score metric based on keyword matching, short-term filtering, and key term frequency analysis.

By detailing the design and functionality of our confidence score metric and comparing it with existing answer evaluation methods, we have highlighted its unique focus on context relevance assessment using a combination of techniques that offers a more direct and interpretable measure compared to many existing approaches.

Building upon the promising foundation of *ExamAce*, several enhancements and extensions can be implemented to increase its impact and usability:

Expanded Dataset Integration: Incorporate datasets from additional domains and institutions, ensuring broader applicability across diverse engineering syllabi.

Advanced Personalization: Leverage machine learning to create adaptive learning paths tailored to individual student needs and study habits.

Offline Mode: Develop an offline version of the app to make it accessible to students in areas with limited internet connectivity.

Collaborative Learning Features: Introduce features for collaborative learning, enabling students to share resources, discuss topics, and solve problems collectively within the app.

REFERENCES

- [1] Salemi, A. and Zamani, H., 2024, July. Evaluating retrieval quality in retrieval-augmented generation. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2395-2400)
- [2] Wang, B., Wang, S. and Ouyang, Q., 2024. Probabilistic Inference Layer Integration in Mistral LLM for Accurate Information Retrieval.
- [3] Smutny, P. and Schreiberova, P., 2020. Chatbots for learning: A review of educational chatbots for the Facebook Messenger. Computers & Education, 151, p.103862.
- [4] Cuconasu, F., Trappolini, G., Tonello, N. and Silvestri, F., 2024. A Tale of Trust and Accuracy: Base vs. Instruct LLMs in RAG Systems. arXiv preprint arXiv:2406.14972.
- [5] Thakur, A. and Gupta, R., 2024. Introducing Super RAGs in Mistral 8x7B-v1. arXiv preprint arXiv:2404.08940.
- [6] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T. and Riedel, S., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, pp.9459-9474.
- [7] Dakshit, S., 2024. Faculty Perspectives on the Potential of RAG in Computer Science Higher Education. arXiv preprint arXiv:2408.01462.
- [8] Modran, H., Bogdan, I.C., Ursuțiu, D., Samoilă, C. and Modran, P.L., 2024. LLM intelligent agent tutoring in higher education courses using a RAG approach.
- [9] Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P.S. and Wen, Q., 2024. Large language models for education: A survey and outlook. arXiv preprint arXiv:2403.18105.
- [10] Bernabei, M., Colabianchi, S., Falegnami, A. and Costantino, F., 2023. Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances. Computers and Education: Artificial Intelligence, 5, p.100172.
- [11] Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N. and Mian, A., 2023. A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.
- [12] Neumann, A.T., Yin, Y., Sowe, S., Decker, S. and Jarke, M., 2024. An llm-driven chatbot in higher education for databases and information systems. IEEE Transactions on Education.