LLaMA vs Mistral: A Comparative Evaluation for a University Fresher Information Chatbot

Thota Ramathulasi[1] – ramathulasi.t@vitap.ac.in
Pula Aquib Younis[1] – aquibyounis1@gmail.com
Yettapu Jyothsna[1] – yettapujyothsnareddy@gmail.com
Punna Sudha Kalyan[1] – punnasudhakalyan@gmail.com
Kogatam Kousik[1] – kogatamkousik55@gmail.com

Department of Data Science & Engineering,
School of Computer Science & Engineering,
Vellore Institute of Technology–AP, India
ramathulasi.t@vitap.ac.in

Abstract. Conversational agents are being used on college campuses to help new and returning students with academic advice, administrative procedures, and general campus questions. Retrieval augmented generation (RAG) is now a feasible method for anchoring chatbot responses in trustworthy external knowledge because to developments in large language models (LLMs). However, selecting the best LLM is not simple because it requires weighing response quality against computing cost and inference speed.

In this research, two open-source LLMs Mistral and Llama 3.2 integrated separately as model within a single RAG-based college information conversation application are compared and evaluated. Both models were tested in identical environments with a fixed context frame of 2000 tokens, same database, and the same retriever setup. Quantitative result like response time and output length were paired with automated qualitative indicators like completeness, fluency, and relevance. The findings indicate that Mistral generates more tokens and comprehensive responses, LLaMA 3.2 produces quicker and more concise responses. These results offer useful insights for choosing LLMs for conversational systems focused on campuses.

Keywords: Conversational Agents · Large Language Models · Retrieval-Augmented Generation · Mistral

1 Introduction

Conversational AI systems have become essential applications for information access in domains such as education, healthcare, and customer support in majority peoples life. In universities, students frequently rely on college websites and help desks to obtain information such as academics, hostels, placements,

and procedures. These systems are often static, difficult to navigate multiple platforms, outdated.

The conversational capabilities of large language models (LLMs) are constantly evolving, leveraging historical information to support reasoning and extended multi-turn interactions. When added to retrieval-augmented generation (RAG) as model, these can clearly respond from campus database, reducing hallucinations while containing conversational flexibility. Despite these advantages, real deployment requires careful consideration of inference latency, response clarity, and overall user experience.

This study presents a controlled comparison of LLaMA 3.2 and Mistral, two widely used open-source LLM's to select an LLM for university fresher information chatbot. The objective is to practically evaluate trade-offs between response efficiency and response richness under identical environment.

1.1 Contributions

The primary contributions of this work are:

- A controlled comparison between Mistral and LLaMA 3.2 in a similar RAG-based campus data chatbot under the same conditions.
- An analytical evaluation which is based on qualitative response quality metrics lower latency and output length.
- Practical insights for selecting LLMs based on deployment constraints in institutional environments.

1.2 Motivation and Problem Statement

New students often ask questions about campus services, academic steps, placement processes, and dormitory rules. Static FAQ sites frequently fail to offer useful clarification or contextual follow-up, which results in more administrative effort and recurring enquiries.

While LLM-powered chatbots can address this gap, institutions must select models that balance responsiveness with answer quality. Benchmark-based evaluations alone are insufficient for this purpose, as they often ignore deployment-specific constraints. Therefore, this work evaluates LLaMA 3.2 and Mistral within a realistic RAG-based deployment tailored to university environments.

2 Related Work

The scalability and flexibility of early conversational systems were constrained by their reliance on rule-based techniques along with matching keywords. Although neural sequence-to-sequence models were more flexible, they had trouble retaining context over time. Transformer architectures, which rely on self-attention to effectively model long-range dependencies and enable scalable and coherent dialogue generation, were able to mitigate these limitations.

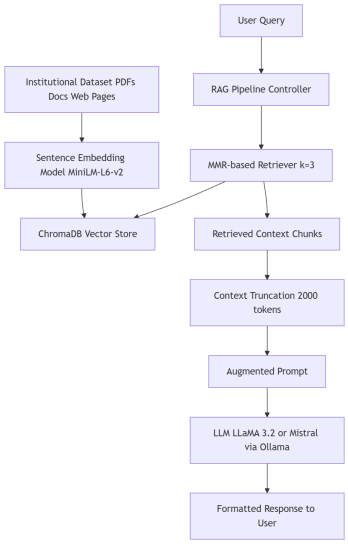

Fig. 1. Retrieval-augmented generation (RAG) architecture employed in the proposed university chatbot. All components except the language model are held constant to ensure a fair comparison between LLaMA 3.2 and Mistral.

Trade-offs between inference speed and response quality have been highlighted in recent studies that have focused on efficiency-aware evaluation of LLMs. There are still few controlled comparisons in domain-specific university chatbot settings, despite the fact that LLaMA and Mistral have each been assessed separately in a variety of tasks. By concentrating on a practical deployment scenario, this work fills this gap.

3 System Architecture

A retrieval-augmented generation (RAG) pipeline is introduced. A sentence-transformer model is used to embed user queries, which are then compared to a persistent vector database that holds university data. Maximal Marginal Relevance(MMR) is used to retrieve pertinent documents while balancing diversity and relativeness.

3.1 Retrieval-Augmented Generation Pipeline

Prior to prompt construction, retrieved contents are concatenated and truncated to a fixed context window. While avoiding excessive context expansion that might skew latency or output length, this is designed model fairness. The same embeddings, retriever parameters, prompt templates, and a fixed context window of 2000 tokens were used to integrate LLaMA 3.2 and Mistral.

4 Evaluation Methodology

4.1 Dataset

The knowledge base comprises curated institutional information relevant to university environments, including academic programs, hostel facilities, campus activities, placement procedures, and administrative guidelines. The evaluation themes were designed to reflect the kinds of conversations and information gathering scenarios that students might encounter on a daily basis. For a fair comparison, the same knowledge data and question were used to evaluate both models.

4.2 Experimental Configuration

Both models were evaluated using the same vector database, MiniLM-L6-v2 embedding model, and Maximal Marginal Relevance (MMR) retrieval strategy with $k = 3$. Both models were subject to a restricted context window of 2000 tokens to maintain fairness. The models were tested in a zero-shot scenario after being locally deployed using the Ollama runtime. The End-to-end execution time, including the document extraction and response generation, is reflected in latency measurements.

4.3 Reproducibility and Fairness Controls

Differences in context size and retrieval configurations can introduce bias when comparing large language models. To reduce this bias and ensure a fair comparison, all components of the system, including the embedding model, the retriever strategy, prompt structure, and context window, were held constant with only the underlying language model varied across experiments.

4.4 Evaluation Metrics

- **Relevance**: Measures how accurately the response addresses the user's query.
- **Fluency**: Evaluates the readability and naturalness of language.
- **Completeness**: Evaluates the model if the answer has all the necessary information required for the query.
- **Latency**: Measures end-to-end response time per query.
- **Output Length**: Captures the average amount of tokens generated per response.

4.5 Automated Quality Evaluation

An automated method based on LLM, or the LLM-as-a-judge strategy, was used to assess the standard of qualitative responses. The evaluator used a uniform making prompt that was applied consistently to both models to score each response according to its completeness, fluency, and relevance.

Table 1. Average quantitative and qualitative evaluation results

Metric	LLaMA 3.2	Mistral
Average Latency (s)	13.7	26.1
Average Output Tokens	70	83
Relevance Score	3.7	4.6
Fluency Score	3.4	4.0
Completeness Score	3.9	4.8

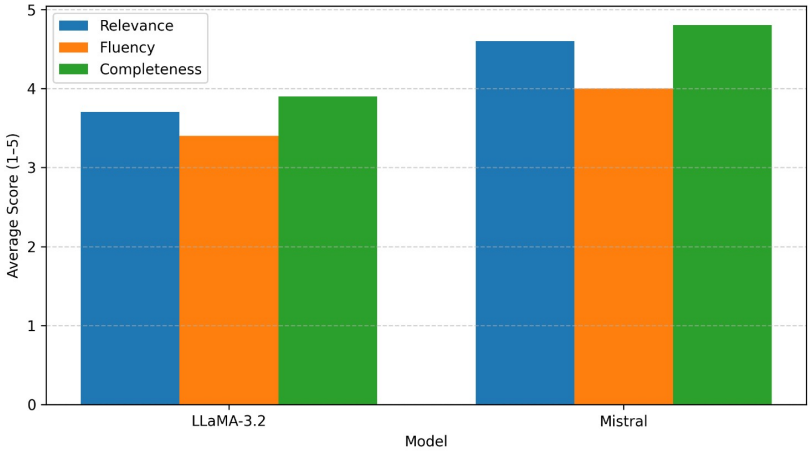

Fig. 2. Qualitative evaluation comparison of LLaMA 3.2 and Mistral.

4.6 Query Design and Evaluation Protocol

Evaluation queries comprise of administrative procedures, academic processes, campus services, and casual conversational input, all framed from a student centric perspective.

5 Results and Discussion

Table 1 summarizes the comparative quantitative and qualitative performance of both models in all evaluation metrics.

5.1 Qualitative Evaluation

Figure 2 shows the comparative qualitative analysis of LLaMA 3.2 and Mistral on the relevance, fluency, and completeness metrics.

5.2 Quantitative Evaluation

Figures 3 and 4 present the quantitative comparison of both models in terms of response latency and average output length, respectively.

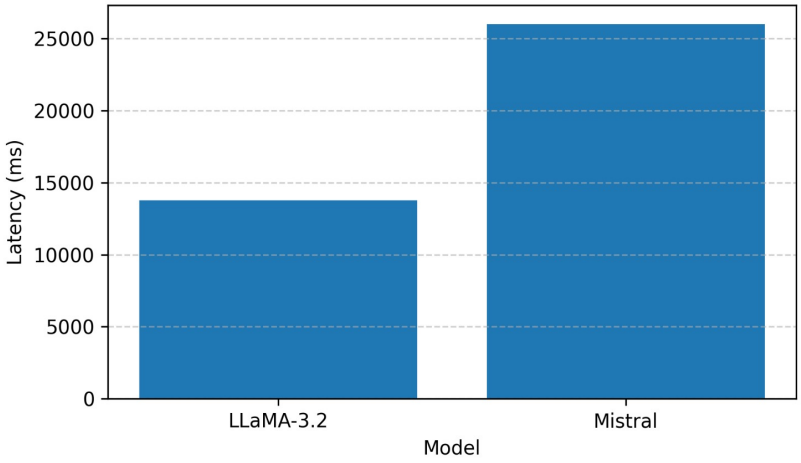

Fig. 3. Average response latency comparison.

5.3 Latency–Quality Trade-off Analysis

The results show a clean trade-off between efficiency and response outputs. LLaMA 3.2 consistently achieves lower response latency and generates minimal outputs, whereas Mistral produces more fluent and lengthy responses at the cost of higher inference time. This highlights the need to consider both response quality and system efficiency when evaluating retrieval-augmented conversational systems intended for real-world deployment.

5.4 Implications for University Deployment

Although the Mistral model achieves higher fluency and completeness, the LLaMA 3.2 model delivers faster and more concise responses. In student university chat applications, lower latency and concise responses are often more valuable than extended explanations. As such, LLaMA 3.2 is better suited towards real-time campus information systems, and Mistral may be better adapted towards advisory applications that necessitate a higher richness in response behavior.

6 Conclusion and Future Work

This paper set out to conduct a comparison and analysis between LLaMA 3.2 and Mistral within the context of a RAG chatbot supporting fresher university information. The study clearly indicates that although Mistral is far superior in terms of its response fluency and completeness, LLaMA 3.2 is far better adapted towards response efficiency and succinctness, thus making it better adapted towards deadlines-based student interaction applications.

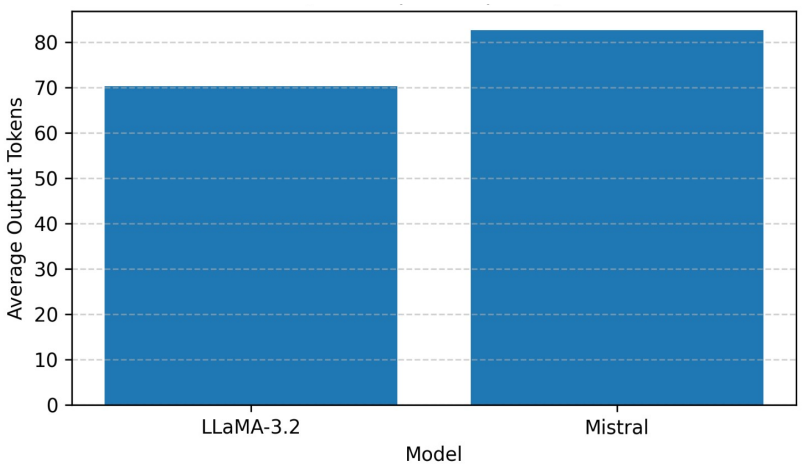

Fig. 4. Average output length comparison.

6.1 Limitations

The experiment is run on a single institutional knowledge base that covers usual and expected academic, administrative, and campus-related information that would normally be encountered within university structures. Although this is amply adequate towards comparing the behavior between the models within controlled parameters, other institutional bodies may differ within their respective datasets.

6.2 Threats to Validity

Automated experimentation may not entirely measure subjective user satisfaction or contextual preferences and biasing towards certain response qualities may also bias towards individual model behavior among LLM-based evaluators and may thus differ on quantitative assessment criteria.

6.3 Future Research Directions

Future work will incorporate human-centred evaluation by collecting feedback from student participants on the clarity, usefulness, and preferred response style of chatbot outputs. Such user-level assessments can complement automated metrics and provide stronger evidence for selecting the most suitable LLM for retrieval-augmented university information systems. Additionally, domain-specific fine-tuning may be explored to further align model behaviour with institutional communication requirements.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30 (2017).
2. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9475 (2020).
3. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997* (2023).
4. Dubey, A., Jauhri, A., Pandey, A., et al. The LLaMA 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024).
5. Touvron, H., Martin, L., Stone, K., et al. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288* (2023).
6. Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).
7. Es, S., James, J., Espinosa-Anke, L., Schockaert, S. RAGAS: Automated Evaluation of Retrieval-Augmented Generation. *arXiv preprint arXiv:2309.15217* (2023).
8. Huang, Y., Li, S., Zhang, W., Chen, X., Liu, Z. Evaluation of Retrieval-Augmented Generation: A Survey. *arXiv preprint arXiv:2405.07437* (2024).
9. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Sailer, M., Schmidt-Thieme, L., Schneider, M., Theis, L., Weidinger, L., Welzel, J., Weller, A. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, 103 (2023).
10. Yan, L., Whitelock-Wainwright, A., Guan, Q., Wen, G., Gašević, D., Chen, G. Practical and Ethical Challenges of Large Language Models in Education: A Systematic Review. *British Journal of Educational Technology*, 55(1), 90–112 (2024).
11. Zhang, Z., Chen, Q., Li, Y., Xu, Y., Wang, Y. Efficient Inference of Large Language Models: A Survey. *ACM Computing Surveys* (2024).