Advancements in Local LLMs: A Study of Llama 3.1 and 3.2 in a RAG-Based Hindu Scripture Chatbot

Dr. Snehal Rathi
Department of Computer Engineering
Vishwakarma Institute of Information Technology
Pune, India
snehal.rathi@viit.ac.in

Mr. Sumit Mule
Department of Computer Engineering
Vishwakarma Institute of Information Technology
Pune, India
sumit.22111155@viit.ac.in

Mr. Tanmay Bhansali
Department of Computer Engineering
Vishwakarma Institute of Information Technology
Pune, India
tanmay.22111091@viit.ac.in

Ms. Muskan Gourkhede
Department of Computer Engineering
Vishwakarma Institute of Information Technology
Pune, India
muskan.22110918@viit.ac.in

Mr. Aakash Das
Department of Computer Engineering
Vishwakarma Institute of Information Technology
Pune, India
aakash.22111180@viit.ac.in

***Abstract*—The exploration and understanding of ancient religious texts is important for finding insights and preserving the heritage of the culture. Mahabharata, Ramayana, and Vedas like scriptures from Hindus contain centuries of knowledge, mythological narratives, teachings related to ethics and morals that continue to resonate with people worldwide. However, the complexity of these scriptures along with languages they are written in often arises as an issue for accessibility and understanding. Recent progress in artificial intelligence and natural language processing (NLP), gives promising ways to solve this issue.**

The primary goal of this research is to compare and evaluate two LLM models and development of a chatbot based on them designed to answer questions and provide brief explanations regarding Historic scriptures. In the subsequent sections of this paper, we go into the methodology implemented for developing the chatbot and the models used and how they perform, discuss the challenges encountered, and present the outcomes of our experiments. Through this work, we see a future where AI-driven chatbots serve as valuable companions for cultural exploration, unlocking the values of ancient wisdom for coming generations.

Keywords—LLM, Machine learning, Artificial Intelligence, Models, Hindu Scripts

I. Introduction

Mahabharat and Ramayan are one of the historical scripts that have immense importance in our country. These have provided us with knowledge, moral practices, and interesting arcs of humanity insights. But the main problem that is usually faced by readers is the language in which these scripts are written, majorly Sanskrit and majority of Indians are not well versed with the complex Sanskrit language.

Ever since the use of ChatGPT became fashionable, several chatbots based on AI and Bhagavad Gita have also been developed in India. They are used by millions of people. Thanks to a true leap forward in AI and NLP, systems now are able to make sense of and even translate even very complicated texts in several languages. This study seeks to compare two Llama models and their performance by using a chatbot based on them. The LLM-based chatbot employs large language models (LLMs), which allow the chatbot to respond in English.

As this is concerned with history and believes it is important that the technology blends well without husting the emotions of readers/users. AI Bots like this have a huge concern because of its association to religions, people in the profession say that this is technology that has great risks when misused and sent to unfriendly hands [1] After all it is a system that has been trained and is still training.

The main purpose of this research was gathering an extensive dataset based on the scriptures, comparing and evaluating two LLM models and based on it we created a chatbot that would answer user queries in most appropriate and educational contents that would help users to understand the context thus bridging language gaps.

II. Literature Review

For preserving the cultural heritage it is important to study religious texts.Hindu scripts like Mahabharat, Ramayana, Bhagavad Gita etc hold an immense importance as they are built on principles which impart knowledge on ethics,philosophy. The complexity and usage of languages like Sanskrit make them difficult to access and to be understood by users.

This paper explores the intersections of Artificial Intelligence (AI) and religious traditions from the perspective of Bernard Lonergan's critical realism. It analyzes how AI transforms and contradicts conventional religious ideals while looking at how it affects religious symbols, rituals, and the development of new spiritual meanings. In order to achieve the common good, the emphasis is on striking a balance between religious life and technical progress.[16]

Religious texts like Vedas, the Upanishads, and the epics such as the Mahabharata and Ramayana, are written in languages like Sanskrit.These languages, though foundational, are no longer widely spoken or understood by the masses. Translations, while useful, often lose subtle meanings embedded in the original. Scholars such as Pollock (2006) [2] have explored the writings of Sanskrit and its interpretations.Thus increasing the gaps in understanding technology and its interaction with historical scriptures. This paper creates a visual representation of Ramayana by summarizing key topics and bridging them with characters and locations.It focuses on Hindi translation of the text using AI models trained in hindi to extract events and summaries topics.[6]

RamChat is a chatbot built by Shepherd University to help students navigate student handbook using both API-based and Local Large Language Model (LLM) for natural language processing (NLP), with a vector store system.[10]

Advances in technology like artificial intelligence and machine learning have helped to bridge the gap to a huge extent. Advances in LLM models (Large Language Models) such as Gemini, GPT 3 are capable of translating complex languages into comparatively simpler languages thus making them ideal to handle complex ancient texts.

An NLP pipeline is proposed in this research to extract computational and statistical insights from the Sanskrit-original Indian epic Mahabharata. To overcome the drawbacks of human analysis, including bias and the challenge of processing vast amounts of material, it uses the identification of patterns like word frequency, sentence structure, and lemmas.[7]

A way to generate questions based on the given context was introduced in this paper. This can be improvised in the context of scripture studies and provide users with relevant information as well as questions.[17]

This research paper explores the ability to generate further potential questions based on the user's previous questions. Further comparison of how different algorithms affect the accuracy of relevance of these questions is elaborated.[18]

Technologies in AI and ML can help in improving interactive tools for understanding complex ancient scripts by simplifying complex texts and providing explanations on user queries.AI driven chatbots help to ease off the language and make it more user accessible and encouraging diversity.

Projects like GitaGPT developed by Bengaluru-based software engineer Sukuru Sai Vineet Google's Sanskrit AI and IIT Bombay's Dharmashree [3] already developed the advances in AI and ML technologies which help them in interpreting and translating ancient hindu scripts. While these initiatives are still increasingly growing they have great potential to increase cultural preservation.

Chatbots built on using LLM APIs like Gemini helps in building and accessing hindu ancient scriptures easily. By using AI and related tools we can access wisdom along with preserving culture hence increasing diversity.

The motivation behind conducting this research stems from several compelling factors that converge at the intersection of technology, culture, and education. The development of a multilingual chatbot for Hindu scriptures is driven by the following key motivations:

A. Cultural Preservation

Hindu scriptures, including the Mahabharata, Ramayana, Bhagavad Gita, and Puranas are treasures of Indian history but they are difficult to access as these scripts are majorly composed in Sanskrit language. A chatbot is created using Artificial Intelligence and Machine learning techniques which would help users to understand the script and are easily available.

B. Educational Outreach

Due to language limitations, many academics and educational institutions majorly face an issue of including these historical scripts into their academic curriculum. This chatbot is an advanced technology to traditional teaching methods as it provides easy and intermediate interpretation of these scripts.

C. Technological Innovation

The research represents application of machine learning and AI (models of LLM) in the domain of cultural studies. By developing a chatbot capable of interpreting and responding to user questions about Hindu scriptures, we push the boundaries of AI-driven conversational agents.[9][11]

D. Cross-Disciplinary Impact

The research works on different domains (linguistic models, historical scriptures, technology). The interaction of these three domains collectively has developed this chatbot.

III. METHODOLOGY

The development of chatbot and comparison of models is done using a structured methodology that uses techniques for natural language understanding and RAG based models.

A. Architecture

The chatbot uses the Retrieval Augmented Generation (RAG) architecture, by combining the Ollama LLM models and the FAISS vector database.[10] Given architecture allows chatbot to create embeddings locally on the machine. The workflow involves getting relevant passages from the scriptures and generating a relevant response based on that context.

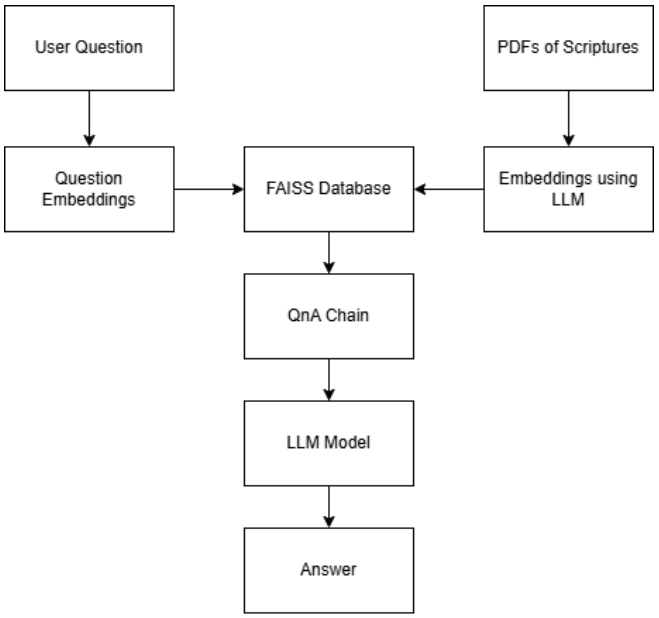

Fig. 1. Architecture of Chatbot

B. Core Functionalities

- User Input Processing: When a query is entered, the chatbot processes it to find out intent and context.
- Text Chunking: To pluck out relevant text segments from the Mahabharata and Ramayana, the chatbot uses the LangChain library's text splitter module.
- Vectorization: Drawn out text is vectorized with the help of the nomic-embed-text model and stored in the FAISS vector database for quick and efficient retrieval.
- Question Answering: The system creates a combination of retrieved text chunks to create precise final responses.

For the research we are using Ollama which is an open-source platform that lets users run AI models on their local devices without an internet connection. It's designed to simplify the process of running large language models (LLMs) locally. This allows users to focus on interacting with the LLM and exploring its capabilities, rather than managing the complexities of model weights, configurations, and dependencies.[19]

Current models used are Llama 3.1 and Llama 3.2 which are Meta's latest flagship language models. Llama 3.1 has multiple models with 8B, 70B and 405B parameter sizes. We are using a model with 8B parameters of Llama 3.1. Whereas for Llama 3.2, we are using its 3B sized model. These models excel at text summarization, text classification, sentiment analysis, and language translation requiring low-latency inferencing. Apart from using this model to ask queries on the provided context, we used nomic-embed-text model to create embeddings of the textual context that is provided to it.

The flow is as follows :

1. All the scripture related files and PDFs are put in a single folder called ‘content’.
2. The program scans that folder and retrieves raw text from those files.
3. We divide this text into smaller chunks to avoid exceeding its maximum sequence length and causing memory issues or incomplete processing.
4. These embeddings are used to create a FAISS vector database for efficient similarity searchings.
5. It provides us with context that contains content related to the use query.
6. This context is then provided to the LLMs model that is running locally with the user query to get a curated response.

Main support material used as a scripture reference here was - “The Ramayana - R. K. Narayan” and some other relevant content.

IV. RESULT AND ANALYSIS

Each of the following are the user queries provided to the model and their respective responses :

Question 1 :

“Who was Ravana ?”

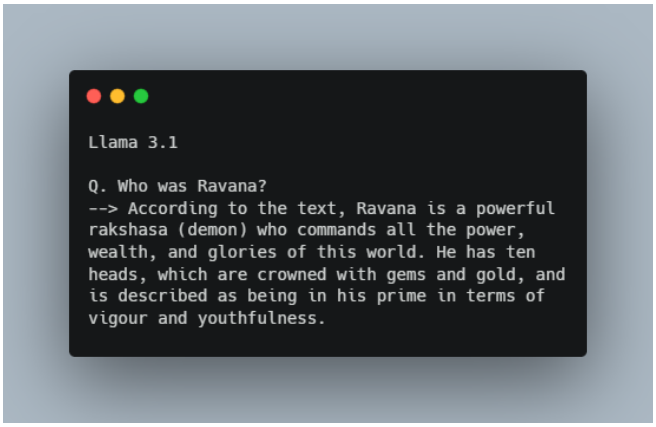

Fig. 2. Output from Llama 3.1 (Question 1)

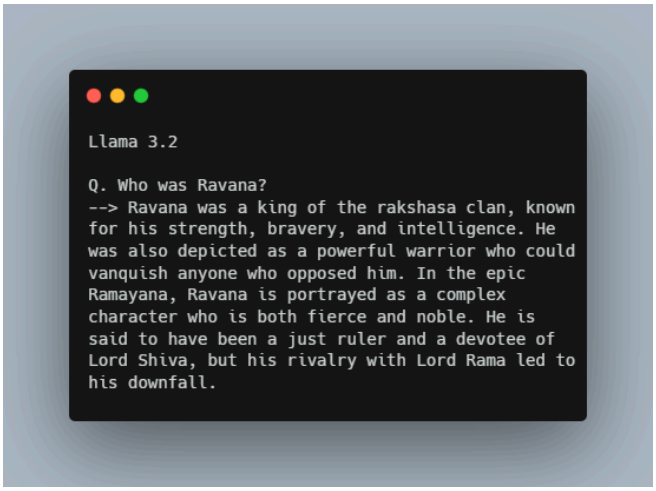

Fig. 3. Output from Llama 3.2 (Question 1)

Question 2 :

"Who was Hanuman ?"

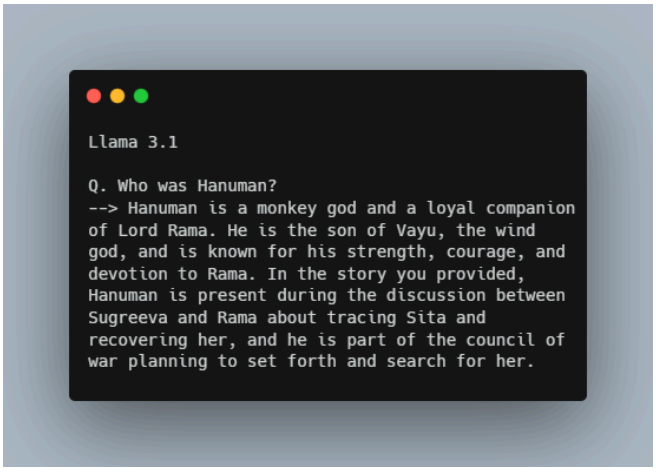

Fig. 4. Output from Llama 3.1 (Question 2)

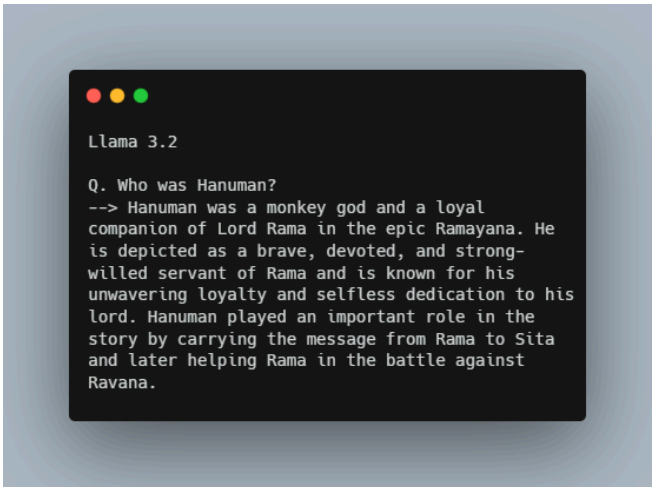

Fig. 5. Output from Llama 3.2 (Question 2)

These are the responses from the Llama 3.1 (8b) model. Studies show that the parameter size of the model plays a vital role in generating an accurate response for the provided context.

TABLE I. ANALYZED DIFFERENCE

Criteria	**Model**	
	Table column subhead	***Subhead***
Tone & style	Straightforward and basic content	Nuanced and versatile.
Speed of response	Slower on hardware with less power due to its size.	Fast and efficient on laptops with good hardware.
Parameters	8B	3B
Content Quality	Reliable but can be too detailed.	Well structured, high quality content.

The size of the Llama 3.1 model used here is 8B parameters whereas the Llama 3.2 model has a 3B parameter size. However, using a model with larger size doesn't necessarily mean better for faster results. The smaller model was seen giving similar results on the same machine using fewer resources.

Rather using cloud based models can provide response faster and in a detailed manner because of their larger sequence length size and more parameters.

V. CONCLUSION

In conclusion, our development of a chatbot for Hindu scriptures showcases the transformative potential of AI, particularly LLM models, in cultural preservation and educational outreach. Also using

larger cloud based models can provide better results for a provided context as compared to models that are running locally.

By leveraging advanced LLM technologies and intuitive design, our chatbot breaks down language barriers and provides insightful answers to users.

VI. References

[1] Nadia Nooreyezdan (2023) India's religious AI chatbots are speaking in the voice of god — and condoning violence. Retrieved from : https://restofworld.org/2023/chatgpt-religious-chatbots-india-gitagpt-krishna/

[2] Pollock, S. (2006). The Language of the Gods in the World of Men: Sanskrit, Culture, and Power in Premodern India. University of California Press.

[3] Dharmashree Project, IIT Bombay.

[4] Borek, Cecylia. "COMPARATIVE EVALUATION OF LLM-BASED APPROACHES TO CHATBOT CREATION." (2024).

[5] Chiang, Wei-Lin, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez and Ion Stoica. "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference." ArXiv abs/2403.04132 (2024): n. pag.

[6] Thottempudi, Sree Ganesh. "A visual narrative of ramayana using extractive summarization topic modeling and named entity recognition." CEUR Workshop Proc.. Vol. 2823. 2021.

[7] Gadesha, Vrunda, Keyur Joshi, and Shefali Naik. "Estimating related words computationally using language model from the mahabharata an indian epic." ICT Analysis and Applications: Proceedings of ICT4SD 2022. Singapore: Springer Nature Singapore, 2022. 627-638.

[8] Reed, Randall. "AI in Religion, AI for Religion, AI and Religion: Towards a theory of religious studies and artificial intelligence." Religions 12.6 (2021): 401

[9] Tamrakar, Rohit & Wani, Niraj. (2021). Design and Development of CHATBOT: A Review.

[10] Burgan, Cara & Kowalski, Josiah & Liao, Weidong. (2024). Developing a Retrieval Augmented Generation (RAG) Chatbot App Using Adaptive Large Language Models (LLM) and LangChain Framework. Proceedings of the West Virginia Academy of Science. 96. 10.55632/pwvas.v96i1.1068.

[11] Lin, Chien-Chang & Huang, Anna & Yang, Stephen. (2023). A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022). Sustainability. 15. 4012. 10.339

[12] Sharma, Sachin, Yunika Kadayat, and Ranu Tyagi. "Artificial Intelligence Enabled Sustainable Life Cycle System Using Vedic Scripture and Quantum Computing." 2023 3rd International Conference on Intelligent Technologies (CONIT). IEEE, 2023.

[13] Jacoba, Rico C. "Exploring the Role of Artificial Intelligence in Interreligious Discourse." RELIGION AND SOCIAL COMMUNICATION Journal of the (2023): 375.

[14] Youvan, Douglas. (2024). GPT-4's Religious Training by Red Teams. 10.13140/RG.2.2.29200.42241.

[15] N. T. K. Le, N. Hadiprodjo, H. El-Alfy, A. Kerimzhanov and A. Teshebaev, "The Recent Large Language Models in NLP," 2023 22nd International Symposium on Communications and Information Technologies (ISCIT), Sydney, Australia, 2023, pp. 1-6, doi: 10.1109/ISCIT57293.2023.10376050.

[16] Umbrello, Steven. (2023). The Intersection of Bernard Lonergan's Critical Realism, the Common Good, and Artificial Intelligence in Modern Religious Practices. Religions. 14. 1536. 10.3390/rel14121536.

[17] 13. S. Rathi, P. Wawage and A. Kulkarni, "Automatic Question Generation from Textual data using NLP techniques," 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2023, pp. 1-7, doi: 10.1109/ESCI56872.2023.10100278.

[18] 20. S. Rathi, S. Mule, S. Nichat, J. Mulla and S. Neware, "Analysis and Implementation of Machine Learning Approaches for Adaptive Learning Using Personalized Assessment," 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 2024, pp. 1533-1541, doi: 10.1109/IDCIoT59759.2024.10468025.

[19] "Exploring the Impact of Different Interaction Design Principles on User Experience: Case Studies and Experiments", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.11, Issue 5, page no.d92-d97, May-2024, Available :http://www.jetir.org/papers/JETIR2405308.pdf