

Vivechan AI: Extracting Wisdom from Ancient Indian Texts Through LLM

Om Soni^a, Jatayu Baxi^b, Bhavika Gambhava^c and Brijesh Bhatt^d

Department of Computer Engineering, Dharmsinh Desai University, Nadiad, India

Keywords: LLM, FAISS, QA, NLP, ML.

Abstract: In a time when technology and human efforts coexist together, the preservation and sharing of ancient wisdom becomes increasingly important. In this context, “Vivechan” provides a unique solution that uses Artificial Intelligence (AI) to extract insights from ancient Indian scriptures. This paper presents the implementation of Vivechan, an AI question-answering tool that uses Large Language Models (LLM) to explore the depths of great texts such as Shiv Puran, Ramayan, Bhagavad Gita etc. Vivechan's standard data preparation and indexing modules translate the huge repository of ancient knowledge into an organized dataset, which allows rapid retrieval. Vivechan effectively links user queries with relevant passages from old texts using state-of-the-art technologies like Facebook AI Similarity Search. These relevant passages provide the LLM necessary background to provide clear and meaningful answers. This paper describes the Vivechan project's architecture, challenges, methods, and outcomes, providing details on its ability to bridge the gap between ancient knowledge and modern technology.


1 INTRODUCTION


Indian Vedic literature religious texts¹ and epic poems hold significant historical, cultural and spiritual importance and serve various purposes. Many religious texts, especially ancient scriptures, are written in archaic or poetic language that can be difficult for modern readers to understand without proper guidance or interpretation. Religious texts often contain complex concepts, allegories, and symbolism that may be difficult to grasp on without expertise.


The Ramayana² is an epic poem of ancient Indian literature. The Bhagavad Gita³ is highly revered in Hinduism and is considered one of the most important spiritual classics in the world. It addresses profound philosophical and ethical questions and covers various aspects of life, duty, morality, and the nature


of existence. The teachings of the Bhagavad Gita have also influenced many other religious and philosophical traditions globally. The Shiv Purana⁴ is one of the eighteen major Puranas in Hinduism and holds significant importance for devotees and scholars alike.

The focus of previous machine learning-related research was on extracting information from individual ancient texts, such as the Ramayana [1] the Mahabharat [2] the Bhagavad Gita [3] Shiv Puran. Our work takes into consideration multiple prominent texts rather than a single resource. Each religious text provides unique perspectives, teachings, and insights into spirituality, morality, and the nature of existence. By studying multiple texts, individuals can gain a more comprehensive and nuanced understanding of religious beliefs and practices, allowing them to appreciate the diversity and richness of religious traditions. But it is very difficult for a user to do a

^a  <https://orcid.org/0009-0003-3164-2364>

^b  <https://orcid.org/0000-0001-5377-7161>

^c  <https://orcid.org/0000-0001-9237-9035>

^d  <https://orcid.org/0000-0002-7934-7992>

¹ https://en.wikipedia.org/wiki/Hindu_texts

² <https://en.wikipedia.org/wiki/Ramayana>

³ https://en.wikipedia.org/wiki/Bhagavad_Gita

⁴ https://en.wikipedia.org/wiki/Shiva_Purana

detailed of all these texts considering time constraints, complexity and language barrier.

Our question-answer system “Vivechan AI” provides clarity by addressing specific queries and explaining key aspects of the text, enhancing understanding for readers. Interacting with Vivechan AI encourages active engagement with religious texts. By asking questions and receiving answers, readers can explore different perspectives, and interact with the material in a more meaningful way. Every reader approaches religious texts with their own background, beliefs, and questions. Vivechan AI allows for personalized learning experiences, where readers can seek answers to their specific inquiries and build an understanding of the text according to their interests and needs.

The introduction of Large Language Model (LLM) has revolutionized the landscape of natural language processing [4,5,6]. It has enabled AI systems with an ability to understand and generate human-like text. LLMs are built upon vast amounts of textual data and trained using advanced deep learning techniques. By using the power of LLMs, question-answering (QA) systems have emerged as an important application which allows users to pose natural language questions and receive accurate, contextually relevant answers. By employing LLMs for QA tasks, these systems can effectively interpret the intent behind user queries and generate informative responses.

LLM-based information retrieval has been gaining attention in the current times [7,8]. Most of the existing work in the field of QA using LLM is focused on general knowledge questions and reasoning tasks [9,10,11]. Despite the advances in QA systems and LLMs, most of the current implementations are general-purpose solutions that cover a broad spectrum of subjects and domains. However, there are surprisingly few QA systems designed for spiritual questions, especially in the context of ancient Indian texts. Recognizing this gap, we developed Vivechan to unlock the profound insights contained within these texts.

2 SYSTEM ARCHITECTURE

Figure 1 shows the architecture of Vivechan system. Vivechan works by first encoding user queries and then building indexes to make it easier to find relevant information quickly inside the indexed dataset. Based on the encoded query, a similarity search is then carried out to find contextually relevant content. After

the context has been obtained, it is used to create a thorough understanding of the user's question. This allows the system to use LLMs to produce precise and contextually relevant answers. The source code of the system is publicly available on GitHub⁵.

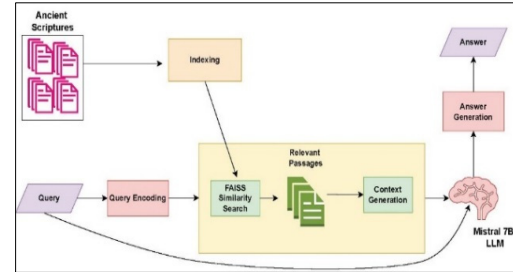


Fig. 1. Architecture of the System.

3 DATA PREPERATION

The first step in preparing the dataset for the Vivechan project is to acquire spiritual literature such as the Shiv Puran, Ramayan, and Bhagavad Gita largely available in PDF format. Using PDF parsing techniques, the contents of these documents are carefully retrieved and translated into machine-readable plain text.

The extracted text file is then segmented, which separates it into discrete lines, each representing a different portion of the text. This segmentation step helps to organize and parse the text, setting the framework for further processing stages. Following line segmentation, the next important step is to organize the split lines into unified paragraphs. This method ensures grouping of related lines into contiguous blocks of text, resulting in cohesive paragraphs. To ensure consistency and readability within the dataset, each paragraph is limited to a maximum of 256 words.

Once paragraphs have been created, they are annotated with metadata that provides contextual information about their source. Each annotated paragraph includes information such as the title and chapter of the originating book, making it easier to trace characters of any form or language are allowed in the title and chapter of the originating book, making it easier to trace and refer back to the original materials. Table 1 shows the details of the data sources along with number of pages. The dataset contains total 80,780 examples.

⁵ <https://github.com/om-ashish-soni/vivechan-ai-v3>

The Vivechan project's dataset preparation procedure allows the creation of a well-structured library of ancient Indian knowledge. The dataset is contributed to Hugging Face and is now publicly available at <https://huggingface.co/datasets/om-ashish-soni/vivechan-spiritual-text-dataset-v3>. This accessibility assures that the researchers can access and use this large dataset for future research and development in the field of spiritual text processing and AI applications. Figure 2 shows the snapshot of the dataset from HuggingFace.

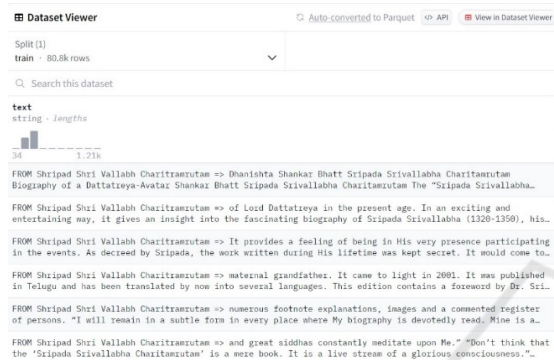


Fig. 2. Snapshot of the Dataset from HuggingFace.

Table 1. Details of the Data Sources.

Sr No	Spiritual Text	# of Pages
1	Shrimad Bhagwat Mahapurana	1284
2	Shripad Shri Vallabha Charitramrutam	594
3	Shiv Mahapurana Sankshipt	171
4	Valmiki Ramayan	1709
5	Vachanamrutam	928
6	Shikshapatri	112
7	Shree Sai Charitra	830
8	Devi Mahatmaya (Chandipath)	118
9	Eknathi Bhagwat	1635
10	Shri Dattapurana	767
11	Shri Gurucharitra	89
12	Shrimad Bhagwad Gita	224

4 DATA PREPERATION

For Vivechan project, an important component of the question-answering pipeline involves performing similarity search operations using the FAISS technique. In this section, we provide details of the

similarity search process implemented in the Vivechan System.

After the dataset is created, we need to generate indices. The basic purpose of indexing is to make search operation faster. In our system, we make use of the Facebook AI Similarity Search (FAISS) framework [12] for this purpose. It uses the concept of grouping dense vectors for similarity search purposes. It supports large-scale datasets and high-dimensional vectors, which are often encountered in machine learning applications. The indexing structures and search algorithms of FAISS result in rapid and accurate similarity searches. In the current section, we show the process of converting our dataset into the .faiss file format.

Since FAISS processes numerical vectors, the first step is to convert text data in the dataset into numerical vectors. This process is known as vectorization. After this step, FAISS indices are constructed on the high-dimensional vectors for the similarity search operation. These indices capture the underlying data distribution to improve search performance. After this indexing process, the FAISS index is saved in a compressed format. For similarity search purposes, this index file is used. It captures all the information such as indexing structure, parameters, and metadata. The overview of the similarity search process is as follows:

- The vector representation of the query is stored in the encoder object. The vivechan system is capable of processing question of any length.
- The encoded query vector is utilized to perform a similarity search against the FAISS index, known as VectorIndex, which contains vectorized representations of paragraphs from the indexed spiritual texts.
- The search returns the distances between the query vector and the k most similar vectors in the index and the positions of these k most similar vectors in the dataset.
- The distances obtained from the search are normalized to calculate the similarity scores. Similarity of paragraph i is calculated using the below formula:

$$Similarity[i] = 1 - \frac{distance[i] - max_distance}{max_distance - min_distance} \quad (1)$$

Where $distance[i]$ = Distance of paragraph i from the query vector, $max_distance$ = overall maximum distance and $min_distance$ = overall minimum distance.

- We define *matching_threshold* parameter. All positions with similarity scores above this threshold are treated as better positions which contains relevant context for the answer generation.
- We use *generate_context()* function to extract the textual context based on better positions. This context will be used in the later stages to generate the actual answer.

5 ANSWER GENERATION

Vivechan's primary application is to generate insightful answers based on the ancient Indian scriptures satisfying user queries. In this section, we describe the use of LLMs for this answer generation purpose. LLMs use user queries along with the generated context for this purpose. The context is basically part of the ancient text in which the answer to the user query can be found. Vivechan makes use of LLMs for the evaluation of user queries, uses the context generated by the similarity search module, and creates informative natural language responses.

In Vivechan system, we use Mistral 7B LLM for generating answers. Mistral 7B is a 7-billion-parameter language model developed by Mistral AI [13]. Mistral 7B is a carefully constructed language model that combines efficiency and high performance to enable real-world applications. It is appropriate for real-time applications that require speedy answers. At the time of its release, Mistral 7B outperformed the top open source 13B model (Llama 2 [14]) in all benchmark tests. Below are the steps for the answer generation process:

- **Query Encoding:** The query entered by the user is encoded into a vector representation of numbers. The query's semantic meaning is captured during this encoding phase.
- **Context Retrieval:** Relevant passages from the indexed dataset are retrieved using a similarity search against the FAISS index using the encoded query vector. These passages provide background knowledge for the answer generation.
- **Answer Generation with LLM:** The retrieved context, along with the translated query, is provided as input to the LLM. The LLM processes this input and generates a response based on its understanding of natural language and the contextual information provided.
- **Output and Presentation:** The generated answer is presented to the user through the Vivechan interface.

6 MULTILINGUAL SUPPORT

Vivechan provides multilingual support for people from different language backgrounds. User can enter the query in different languages. Currently, the system supports English, Hindi, Gujarati, Marathi, Tamil, Telugu, Kannada, and Bengali languages. Vivechan uses the Google Translate API to convert user-submitted queries into English when they are submitted in a language other than English. Following the translation of the user query into English, the answer is generated. Once the English answer is generated, the Google Translate API is once again used to translate the answer back into the original language of the user query. Due to the multilingual support, users can communicate with the tool in their own tongue, improving accessibility and inclusivity. This strategy increases Vivechan's global reach and also helps people to understand the spiritual teachings on a deeper level regardless of language obstacles.

7 OBSERVATION AND EVALUATION

7.1 Evaluation

In order to evaluate the performance of the Vivechan question-answering system, we conducted the evaluation using a set of 100 gold standard question-and-answer pairs. These pairs are selected to represent wide range of topics found within the ancient Indian texts. The primary aim of this evaluation is to assess the system's ability to generate grammatically correct sentences and also to ensure that the answers were contextually relevant to the questions posed. The evaluation parameters are as follows:

- **Answer Sentence Quality:** This criterion focuses on the grammatical correctness and clarity of the answers generated by Vivechan.
- **Relevance of the Answer with Respect to the Question:** This criterion evaluates the contextual relevance and accuracy of the answers. It involves checking whether the answers address the questions' core elements or not.

The evaluation is carried out by a panel of human evaluators who were well-versed in the subject matter of the ancient texts. Each evaluator independently rated the answers on a scale of 1 to 10 for both

sentence quality and relevance, with 1 being the lowest and 10 the highest. Table 2 shows the sample questions and their corresponding evaluation. We observe that the average rating for the answer sentence quality is **8.66** and for the relevance is **8.59**. The evaluation results presented suggest that the Vivechan system is performing effectively in generating responses that are well-composed but also highly relevant to the queried topics from the ancient texts.

Question	Part of Answer generated by Vivechan	Sentence Quality	Answer Relevance
What is the essence of the teachings of the Bhagavad Gita?	The Bhagavad Gita, a sacred Hindu scripture, is a conversation between Lord Krishna and Arjuna, where Lord Krishna imparts spiritual knowledge and guidance to Arjuna. The essence of Bhagavad Gita's teachings, as stated in the text, is encapsulated in the verse 12 and 13. These verses emphasize the importance of humility, pridelessness, nonviolence, tolerance, simplicity, approaching a spiritual master, cleanliness, steadiness, self-control, renunciation of sense gratification, absence of false ego, and the perception of the evil of birth, death, old age, and disease.	8	9
What is the story behind the golden	The golden deer in the Ramayana is a significant	8	8

deer in the Ramayana, and how does it lead to Sita's abduction?	character in the epic, and its story leads to Sita's abduction. The golden deer first appears in the Aranya Kanda (Book 5) of the Ramayana, where Sita, the wife of Lord Rama, is enamored by its beauty while she is in exile in the forest.		
Describe the character of Ashwatthama.	Ashwatthama, in Hindu mythology, is a significant character mentioned in various Hindu scriptures, particularly in the Mahabharata and Eknathi Bhagwat. Ashwatthama is considered to be the son of Bhishma Pitamaha, the grandsire of the Kuru dynasty	7	4

7.2 Challenges and Limitations

In developing Vivechan, we faced multiple challenges that required careful considerations. These challenges involve language barriers and data quality issues to technical complexities in implementing effective answer generation. Following are the challenges along with the details of methodology to tackle them during the development of the system.

- **Contextual Nuances:** Ancient texts often carry varied meanings and interpretations depending on the context. To address this challenge, we implement a comprehensive approach. Instead of relying on a single search result, we utilize the k most similar matching paragraphs. Additionally, for each of these k paragraphs, we incorporate their

surrounding context into the answer generation process.

- **Language Barrier:** Even though LLMs are fine-tuned on multiple languages, they give their best performance on English language. We Solved this issue by using Google Translate to translate from Indic languages into English and then generating the answer from LLM.
- **Hallucination:** LLM's native behavior is to synthesis the answer if the answer is not known to it, which produces wrong information. We are therefore using Retrieval Augmented Generation (RAG) technique, to search answers first in the VectorStore of ancient text and we generate that context and then we give Query and Context to generate answers.
- **Proof of Correctness:** To ensure the correctness of the answers, we have labelled the paragraphs of ancient texts. This labelling facilitates the effective retrieval of answers and allows us to display the source of the information.

While the Vivechan system has shown significant capabilities responding to queries related to ancient Hindu scriptures, there are some limitations of the system as per our observations.

- **Language Barrier:** The translation of Hindu scriptures from Indic languages to English creates limitation. Many of these texts carry nuanced meanings that are deeply embedded in the cultural contexts of their original languages. During translation, some of these nuances may not be fully captured leading to inaccurate answers.
- **Spelling errors within the queries can significantly impact its performance.**
- **The effectiveness of Vivechan is highly dependent on several hyperparameters which include the choice of tokenizer, the specific LLM used for generating answers, the library utilized for searching context, and the size of the context window.** For example, consider third example in Table 2. Due to incorrect context retrieval, the system gives incorrect answer. It identifies Ashwatthama as son of Bhishma Pitamah which is incorrect. As per Mahabharat, Ashwatthama was son of Drona.

8 CONCLUSIONS

The Vivechan project represents a significant advancement in the domain of AI-driven question-answering systems in the exploration of ancient Indian texts and spiritual wisdom. Through the integration of technologies such as LLMs, FAISS and multilingual support, Vivechan offers users a powerful tool for exploring the depths of spiritual knowledge. Vivechan shows AI's ability to bridge the knowledge gap between traditional wisdom and modern technology, allowing a more appreciation of spiritual teachings in the digital era. The system is deployed and available for the public use⁶.

Disclosure of Interests. The authors declare that there is no conflict of interests.

ACKNOWLEDGEMENTS

We acknowledge Google Colab for providing the necessary GPU support for our model training and experimentation processes. Special thanks to various web resources and digital libraries that have made Indian texts such as Shiv Puran, Ramayan, and Bhagavad Gita accessible online. These resources have been very useful for the preparation and enrichment of our dataset.

REFERENCES

1. "Ramayangpt." <https://github.com/abhiyamishra/RamayanaGPT>.
2. H. Thapliyal, "Unveiling the past: Ai-powered historical book question answering," *Global journal of Business and Integral Security*, 2016.
3. A. Karekar, S. Limaye, A. Nara, and S. Panchal, "Bhagavad geeta based chatbot," in *2023 3rd International Conference on Intelligent Technologies (CONIT)*, pp. 1– 6, 2023.
4. P. P. Ray, "Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet of Things and CyberPhysical Systems*, vol. 3, pp. 121–154, 2023.
5. A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

⁶ Online link: <https://vivechan.streamlit.app/>

6. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
7. O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham, "In-context retrieval-augmented language models," 2023.
8. E. Kamalloo, N. Dziri, C. L. A. Clarke, and D. Rafiei, "Evaluating open-domain question answering in the era of large language models," 2023.
9. N. Bian, X. Han, L. Sun, H. Lin, Y. Lu, B. He, S. Jiang, and B. Dong, "Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models," 2024.
10. C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is chatgpt a general-purpose natural language processing task solver?," 2023.
11. H. Pandya and B. Bhatt, "Enhancing low-resource question-answering performance through word seeding and customized refinement," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 3, 2024.
12. M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The faiss library," 2024.
13. A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023.
14. H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.