INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY
IJARIIT
ISSN: 2454-132X
Impact Factor: 6.078
(Volume 11, Issue 5 - V11I5-1176)
Available online at: https://www.ijariit.com

Ancient Indian Scripture Based Retrieval-Augmented Systems: A Comprehensive Analysis

Pradhyumna Prakash
pradhyumna.id@gmail.com
Delhi Public School, Bangalore East, Karnataka

ABSTRACT

This paper focuses on the development and systematic comparison of Retrieval-Augmented Generation (RAG) systems, retrieval-only systems and LLM models all trained on ancient Sanskrit Scriptures. This was done in order to analyse whether RAG systems improved faithfulness in answers to reflective questions, by storing two pertinent Sanskrit scriptures: the Itihasa (including the Mahabharata and Ramayana) and the Bhagavad Gita in a FAISS index, I developed the following: a basic retrieval system from the FAISS index, a prebuilt LLM model (Qwen 2.5-3B-Instruct), an RAG system with the LLM model Qwen 2.5-3B-Instruct and an RAG system with Gemini 2.5 Flash. After development, I evaluated the four models on a list of twenty questions pertaining to philosophy, interpersonal and intrapersonal understanding, and emotional well-being. I ranked each answer on a scale from 1 to 5 on relevance, helpfulness, clarity and faithfulness. All retrieval and RAG models scored a perfect 5 in the 'faithfulness' metric in contrast to the base LLM model, which scored a 4.3. Moreover, I discovered that the use of a weaker LLM model in an RAG system can lead to worse results in the 'helpfulness' and 'clarity' metrics when compared to a regular LLM model when the retrieved verses are low. Through the methods and results of my research, I showed that RAG systems are necessary to provide specific and faithful answers from ancient Sanskrit philosophy.

Keywords: *Large Language Models, Retrieval-Augmented Generation (RAG), Embeddings, FAISS, Qwen 2.5-3B-Instruct, Gemini 2.5 Flash.*

1. INTRODUCTION

There is no doubt that the knowledge held within ancient scriptures is vast–providing details on diverse subjects such as philosophy, religion, medicine, mathematics, astronomy and arts. Nowadays, especially in India, the integration of Sanskrit scriptures within our lives has seen a drastic increase. With Ayurvedic medicine and meditation practices on a rise, the demand for the knowledge present within ancient Sanskrit scriptures is at an all time high.

Moreover, the discussion of religion and spirituality during therapy has been an important debate in Psychology. Although up to 80% of practicing psychologists say they received little to no training on addressing spiritual and religious issues during therapy, many say it's time for change. By centering the patient and their existing beliefs, psychologists can help people leverage their religious and spiritual resources as a source of strength during difficult times [1]. In the country of India, the use of ancient scriptures for small levels of therapy–particularly Sanskrit scriptures, have slowly been on the rise. For those deeply rooted in Indian traditions, scripture-based therapy feels more authentic and culturally acceptable [2].

However, easy access to the knowledge present within the scriptures is still a pertinent challenge due to linguistic (the original scriptures are in Devanagari) challenges, translation challenges and significant cultural differences between the current era and those of thousands of years ago [3].

It is for this reason that in this paper, I have developed, trained and evaluated different Large Language Model (LLM) systems that attempt to access the knowledge present within ancient Sanskrit scriptures to provide useful solutions to a person's queries and challenges.

In order to propose said systematic analysis, I first chose two pertinent Sanskrit scriptures for model training and evaluation: the Bhagavad Gita and the Itihasa. Using these scriptures, I developed four LLM systems: a FAISS retrieval, a base Large Language Model (Qwen 2.5-3B-Instruct) and two LLM based RAG systems: RAG+Qwen 2.5-3B-Instruct and RAG+Gemini 2.5 Flash. Moreover, I included two RAG+Qwen 2.5-3B-Instruct models: one which would retrieve and therefore analyze the top three verses and one which would retrieve and analyze the top five. This was done in order to evaluate how the number of retrieved verses affects model answers. The inclusion of the FAISS retrieval was to provide a baseline to evaluate the other models on.

1.1 Dataset Description

Two ancient and highly important Sanskrit scriptures have been used in this study: The Bhagavad Gita and the Itihasa. The Bhagavad Gita is widely considered to be the most prominent Sanskrit scripture.

It deals with complex topics ranging from the soul and inner peace to following karma (action) and dharma (an individual's inherent duty).

It consists of 700-701 verses spread out over 18 chapters [4]. On the other hand, the Itihasa is an ancient Sanskrit scripture that particularly deals with the events of the Mahabharata and the Ramayana (two of Hinduism's greatest epics) [5]. The total number of verses contained within the Itihasa is not known.

1.2 System Architectures

Facebook AI Similarity Search (FAISS):

FAISS is an open-source library built by Meta-AI for similarity searches between vectors by comparing their distances (using Euclidean distance or Cosine similarity).

The basic structure of Faiss is an index that can have multiple implementations. It can store a large number of database vectors. When the index receives an input vector, the FAISS index returns the database vector that is closest to the input vector (here closest refers to Euclidean distance) [6].

The indexing algorithm can be split into three steps [7]:

i. K-means clustering: Breaks the data into clusters narrowing the search space needed
ii. Product Quantization: Product Quantization compresses vectors into shorter codes, drastically reducing memory usage and efficiency
iii. Optimized Product Quantization (OPQ): OPQ is an enhanced version of Product Quantization that rotates the clustered data to better fit the quantization grid.

Base retrieval systems that use a FAISS index simply retrieve the FAISS index vector that is most similar to the input text's vector. In this study, I have chosen a base retrieval system (that only retrieves and outputs the most similar vector in the FAISS index) in order to create a baseline for the subsequent models.

Retrieval-Augmented Generation:

Retrieval augmented generation (RAG) is an architecture that improves the results of a Large Language model by connecting it with external knowledge bases. RAG systems provide a cost-efficient implementation and lower the risk of hallucinations. The workflow of an RAG system is given below:

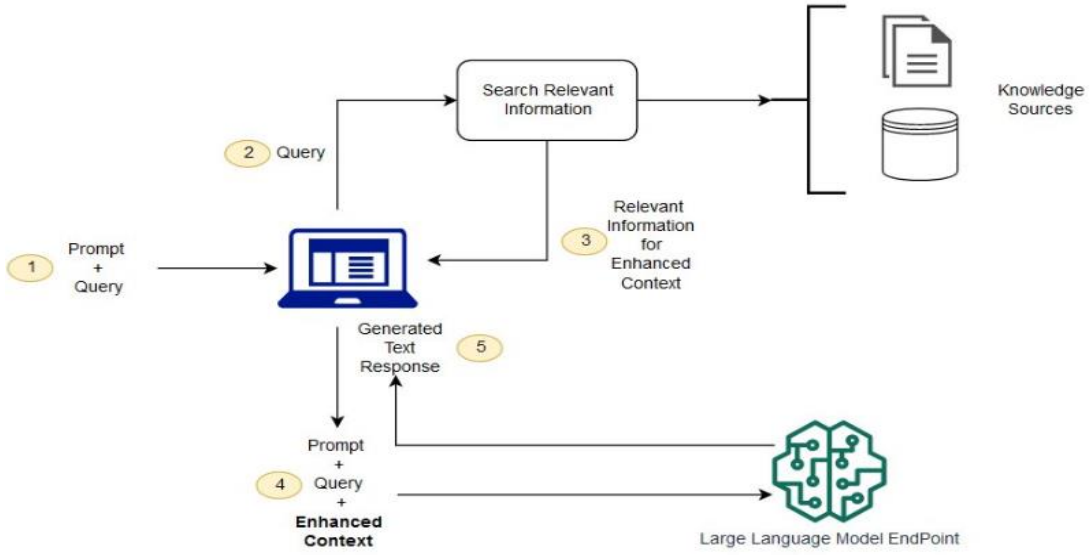

Figure 1: Conceptual flow of an RAG system using LLMs [8]

Any RAG system always follows a 5-step process [9]:

i. The system first receives the input text from the user
ii. It then identifies relevant data from a connected database
iii. This relevant information is then returned to the integration layer of the system
iv. The RAG system, using this relevant information, creates a specific prompt for the connected LLM model
v. The LLM model receives this prompt, answers it and then returns it back to the user

Model Selections:

After rigorous testing, I came to the conclusion that the large language model Qwen 2.5-3B-Instrust, developed by Alibaba Cloud, was the most efficient, non-gated architecture that could present proper answers with resource constraints. The '3B' in the name represents 3 billion parameters used by the model.

Furthermore, I selected Google Gemini 2.5 Flash to be the topline LLM model in this study due to it being resource-efficient and free of cost while still be a flagship model.

1.3 Previous Studies

Two studies incorporating artificial intelligence within the study of Sanskrit scriptures and philosophy have been used in the making of this study.

N Biraja Isac and Himansu Das developed a comprehensive framework (SLIP - Sanskrit Linguistic Intelligence Pipeline) that has the ability to enhance translation quality for classical texts.

The framework uses a modular architecture with three primary components [10]:

i. The Sanskrit Linguistic Analyzer for morpho-phonological feature extraction
ii. The Sanskrit preprocessor for linguistically-informed text preparation
iii. The Sanskrit postprocessor for translation and quality refinement

Research undertaken by Priyanka Mandikal et al., developed an RAG model on the VedantaNY-10M dataset in order to explore the effectiveness of RAG systems for Ancient Indian Philosophy.

Using human evaluation, they found that the RAG model significantly outperforms the standard model in producing factual, comprehensive responses having fewer hallucinations. In addition, they found that a keyword-based hybrid retriever that focuses on unique low-frequency words further improves results [11].

1.4 Objectives

The objectives of this paper are as follows:

i. To develop and systematically evaluate FAISS Retrieval systems, base LLMs and RAG+LLM systems in answering a person's query based on faithfulness to Sanskrit scriptures, relevance, helpfulness and clarity
ii. To evaluate each system's likelihood of hallucinating its verses, teachings and events from the Bhagavad Gita and the Itihasa
iii. To evaluate the effect of a larger number of retrieved verses on the RAG system's answers
iv. To evaluate the effect of including a more powerful model on the RAG system's answers

2. METHODS

2.1 Data Preprocessing

I began my research by collecting, preprocessing and formatting datasets containing the Itihasa and Bhagavad Gita. I collected the Itihasa dataset from Rahular's Github repository [12]. It contains 93,031 verses of the Itihasa (extracted from M.N Dutt's seminal works on the Ramayana and the Mahabharata) with both the Sanskrit verse and English translation on each line. Moreover, the dataset had been split into training, validation and testing data (80.79% training, 12.6% testing and 6.61% validation).

Similarly, I collected the Bhagavad Gita dataset from Aman Kumar on Kaggle [13]. It contained 701 rows of verses along with both the Sanskrit verse and English translation. On loading both in Google Colab, I separated the English translation from the Sanskrit verse and combined both datasets into one JSONL file.

2.2 FAISS Index

I decided to use all-MiniLM-L6-v2 embeddings as it was free, open-source and efficient on Google Colab. After applying the embeddings over the JSONL file and storing only the english translations in an FAISS index (the inclusion of sanskrit translations could lead to a difference in vectors and therefore be stored in the wrong index), I now had 93,301 english vectors in the index. Moreover, I developed a mapping ID between each Sanskrit verse and English translation along with its index, so that each model could retrieve the related Sanskrit verse (which will be used in evaluation). In order to maintain balance between the systems, I initially set the total number of retrieval verses (k) to three. Therefore, the FAISS index will only return the top three vectors most closely related to the input text. Moreover, I added an RAG system with 'k' set to 5 (would retrieve the top 5 most related vectors) in order to compare the effectiveness of increasing 'k' to improve an RAG system's results.

2.3 Model Development

I used three categories of models: Retrieval-based, sole LLM-based and an RAG with an LLM. Furthermore, I developed two RAG+Qwen 2.5-3B-Instruct models, one that would retrieve the top 3 verses (k = 3) and one that would retrieve the top 5 verses (k = 5) from the FAISS index. This was done in order to evaluate how increasing the number of verses retrieved affects model results. Each model was specifically prompted to answer the user's query using the Bhagavad Gita and Itihasa and the retrieval-based and RAG based models were then connected to the FAISS database.

3. RESULTS

3.1 Evaluation Metrics

In order to evaluate all four systems, I carefully crafted 20 questions falling into four categories: Philosophy and Metaphysics, Emotion-Based, Interpersonal-based and Therapy & Self-Growth. Below are the 20 Questions:

Philosophy and Metaphysics:

i. What happens to the soul after death?
ii. Why should I follow my dharma?
iii. How can I practice detachment in daily life?
iv. What does karma really mean for my future?
v. How can knowledge help me reach peace?

Emotion and Self-Mastery

i. How do I control my anger?
ii. What should I do when I feel afraid?
iii. How do I deal with grief after losing someone?
iv. Why do I feel so much hesitation before making a big decision?
v. How can I overcome desire and temptation?

Interpersonal & Leadership

i. How should a good leader behave?
ii. What is the right way to treat my teacher or mentor?
iii. How can I show respect to someone I look up to?
iv. How should I stay loyal to my friends?
v. How can I balance my family responsibilities with my work?

Therapy and Self-Growth

i. How do I handle stress when life feels overwhelming?
ii. Can meditation help me control my emotions?
iii. How do I let go of attachments that hurt me?
iv. What should I do when I feel stuck in a personal crisis?
v. How can I become more resilient after failure?

After receiving each model's output, I graded each answer on four metrics: Relevance (How relevant is the model's answer to the given question. 1 = Not relevant at all, 5 = Very relevant), Helpfulness (How helpful is the answer to the given question. 1 = Not at all helpful, 5 = Very helpful), Clarity (How easy is it to understand the model's answer. 1 = Not easy at all, 5 = Extremely easy) and Faithfulness (How scripture-grounded the answers are. 1 = Does not use the scriptures at all (maximum hallucinations), 5 = Bases its answer off of the scriptures and provides specific details from verses with no hallucinations).

3.2 Human Evaluation for Faithfulness:

In order to properly evaluate each model's evaluation (especially for faithfulness) and remove bias, I instructed each model system to retrieve the specific Sanskrit verse they based their answer on. Each model's answers were then rigorously compared with the specific verses they retrieved to evaluate how scripture-based they were. Furthermore, all model answers to each question, along with my grading, have been provided in the Github repository found in the "references" section of this paper [14].

3.3 Evaluation Results

The following are the results of my research. I have provided each evaluation metrics for the model's overall and category-wise answers.

Table-1: Total Model Average

Model	Relevance	Helpfulness	Clarity	Faithfulness	Overall
Retrieval (k=3)	2.58	1.77	2.08	5	2.86
Qwen 2.5-3B-Instruct	5	4.8	4.95	4.3	4.76
RAG + Qwen 2.5-3B-Instruct (k=3)	4.95	4.7	4.75	5	4.85
RAG + Qwen 2.5-3B-Instruct (k=5)	5	4.7	4.9	5	4.90
RAG + Gemini 2.5 Flash (k=3)	5	5	5	5	5

Table-2: Philosophy and Metaphysics

Model	Relevance	Helpfulness	Clarity	Faithfulness	Overall
Retrieval (k=3)	2.93	2.0	2.2	5	3.03
Qwen 2.5-3B-Instruct	5	5	5	4.4	4.85
RAG + Qwen 2.5-3B-Instruct (k=3)	4.8	5	4.8	5	4.9
RAG + Qwen 2.5-3B-Instruct (k=5)	5	5	5	5	5
RAG + Gemini 2.5 Flash (k=3)	5	5	5	5	5

Table-3: Emotion and Self-Mastery

Model	Relevance	Helpfulness	Clarity	Faithfulness	Overall
Retrieval (k=3)	2.4	1.67	1.93	5	2.75
Qwen 2.5-3B-Instruct	5	4.6	5	4	4.65
RAG + Qwen 2.5-3B-Instruct (k=3)	5	4.6	4.4	5	4.7
RAG + Qwen 2.5-3B-Instruct (k=5)	5	4.8	5	5	4.95
RAG + Gemini 2.5 Flash (k=3)	5	5	5	5	5

Table-4: Interpersonal & Leadership

Model	Relevance	Helpfulness	Clarity	Faithfulness	Overall
Retrieval (k=3)	**2.4**	**1.6**	**2.13**	**5**	**2.78**
Qwen 2.5-3B-Instruct	**5**	**4.8**	**5**	**4.4**	**4.8**
RAG + Qwen 2.5-3B-Instruct (k=3)	**5**	**4.8**	**4.8**	**5**	**4.9**
RAG + Qwen 2.5-3B-Instruct (k=5)	**5**	**4.2**	**4.8**	**5**	**4.75**
RAG + Gemini 2.5 Flash (k=3)	**5**	**5**	**5**	**5**	**5**

Table-5: Therapy and Self-Growth

Model	Relevance	Helpfulness	Clarity	Faithfulness	Overall
Retrieval (k=3)	**2.6**	**1.8**	**2.07**	**5**	**2.87**
Qwen 2.5-3B-Instruct	**5**	**4.8**	**4.8**	**4.4**	**4.75**
RAG + Qwen 2.5-3B-Instruct (k=3)	**5**	**4.6**	**5**	**5**	**4.9**
RAG + Qwen 2.5-3B-Instruct (k = 5)	**5**	**4.8**	**4.8**	**5**	**4.9**
RAG + Gemini 2.5 Flash (k = 3)	**5**	**5**	**5**	**5**	**5**

4. DISCUSSION

4.1 Total Model Metrics

On analyzing the overall scores for each model, we can clearly see that RAG+Gemini 2.5 Flash provides the best answers–earning a '5' on all 4 categories. Moreover, all RAG systems and retrieval systems earned a '5' as they were prompted to base their answers off of the retrieved verse–and thus generated responses that were always relevant. On the other hand, the base Qwen 2.5-3B-Instruct model, although instructed to base its answers off of the Bhagavad Gita and the Itihasa, often made up stories and verse numbers. And although it did answer most questions properly (evident by its high "helpfulness" score), its hallucinations brought its "faithfulness" score down. Also, RAG+Qwen 2.5-3B-Instruct (k=5) did perform slightly better overall compared to RAG+Qwen 2.5-3B-Instruct (k=3). The base retrieval system consistently performed worse over the twenty questions, especially in 'Helpfulness' and 'Clarity' as the retrieved verses were often cryptic and difficult to understand without an external LLM.

4.2 Category Wise Metrics

On analyzing the model scores of the "Philosophy and Metaphysics" questions, it's evident that all RAG models perform well, especially RAG+Qwen 2.5-3B-Instruct (k=5) and RAG+Gemini, which both got perfect scores. The questions in this section are more direct and therefore prioritize retrieval systems. A similar story is found within the next section of questions: "Emotion and Self-mastery."

An interesting observation lies in the next section, "Interpersonal and Leadership." The base Qwen 2.5-3B-Instruct LLM model yet again performs well in the Relevance, Helpfulness and Clarity categories, but continues to hallucinate its responses–leading to a low "faithfulness" score as compared to the other models. Furthermore, the RAG+Qwen 2.5-3B-Instruct (k=5) performs slightly worse than both the base Qwen 2.5-3B-Instruct model and the RAG+Qwen 2.5-3B-Instruct (k=3) system–due to its low "helpfulness" score. This is primarily due to its answer to the question: "How should I stay loyal to my friends?" Unlike the other models, which recommended building strong bonds with friends and actively behaving like a true friend, this system took a much more negative approach, warning the user to not become too devoted to a friendship for it may break at any time. This was due to the last two verses retrieved by the RAG system, both of which took a more negative aspect to the question. The RAG+Qwen 2.5-3B-Instruct system was not strong enough, possibly due to the weaker LLM model, to understand that the incorporation the last two verses should have been accompanied by the first three verses to provide a more balanced answer–leading to it getting a '3' for helpfulness in that question and ultimately a '4.2' overall.

In the final section, "Therapy and Self-growth", the usual trend is observed again with the RAG models consistently getting the highest score in faithfulness.

One interesting limitation to RAG systems, particularly ones with a low 'k' value, is that sometimes, the perfect answer to a given question may not be present within the retrieved verses. This was the case for certain questions in the "Therapy and Self-Growth" category of questions, such as "How do I handle stress when life feels overwhelming?" Only RAG+Gemini 2.5 Flash was capable of realising that the answer to the questions was not present within the retrieved verses. For RAG systems with weaker models and constrained to only answer off of retrieved verses, this could lead to problems.

To summarize the results of this study:

i. RAG systems are required for faithful answers and to prevent hallucinations
ii. The stronger the LLM model connected to the RAG system, the more concise and helpful answers are
iii. Incorporation of weaker LLM models with RAG systems, particularly with a high retrieval value, may have unintended consequences (as found with the LLM+Qwen 2.5-3B-Instruct (k=5) model). The model may get confused with the large amounts of cryptic data and therefore may produce flawed answers
iv. Overall, one can say that a higher retrieval value leads to better results if the LLM model is powerful

5. FUTURE WORK

i. Due to resource constraints, a larger set of models and questions could not have been developed and tests with a larger number of retrieved verses could not have been evaluated. Future studies should include different RAG systems and LLM models as well as further enhance the study towards the effects of the number of retrieved verses in an RAG system's answers
ii. The models were developed on translated scriptures rather than the original scriptures. Future research should look towards developing models that are capable of accessing and analyzing the original scriptures leading to a more robust model.
iii. Future work should prioritize the inclusion of a larger set of evaluators and more standardized testing metrics
iv. Developing an LLM model that is finetuned to analyze ancient Sanskrit scriptures, rather than depending on an RAG system, could be an interesting direction.

6. CONCLUSION

To conclude, the results of this study could further propel research towards the integration of LLM systems in accessing and analyzing the knowledge of ancient Indian philosophy. Through systematic analysis, I found that RAG systems are necessary in order to provide factual information and teachings from Sanskrit scriptures and prevent hallucinations.

Future research should prioritize the incorporation of larger FAISS indexes and more robust models to further evaluate the effectiveness of RAG systems. Furthermore, developing models and systems that can analyze the direct Sanskrit verse, rather than relying on translations, may improve performance.

Ultimately, this study and the future research it catalyzes will help strengthen our knowledge of ancient Indian scriptures through the use of LLM models.

7. REFERENCES

[1] *Abrams, Z. Can religion and spirituality have a place in therapy? Experts say yes*. (2022). American Psychological Association. https://www.apa.org/monitor/2023/11/incorporating-religion-spirituality-therapy
[2] Bhati, R., Mandal, M., & Singh, T. (2025). Ancient Indian perspectives and practices of mental well-being. *Frontiers in Psychology*, *16*. https://doi.org/10.3389/fpsyg.2025.1616802
[3] Narayanrao Jadhav, K., Dr. (2018). Need and challenges of the Sanskrit studies. In *National Journal of Hindi & Sanskrit Research* (Vols. 1–16, pp. 57–60) [Journal-article]. https://sanskritarticle.com/wp-content/uploads/21-16-Kaveri.Jadhav.pdf
[4] *Doniger, W. Bhagavad Gita*. (2023). Britannica. https://www.britannica.com/topic/Bhagavad-Gita
[5] Aralikatte, R., Lhoneux D.M., Kunchukuttan, A., Sogaard, A. A Critical Note on the Evaluation of Clustering Algorithms. (n.d.). *Association for the Advancement of Artificial Intelligence*. https://arxiv.org/pdf/1908.03782
[6] *Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazare, P.E., Lomeli, M., Hosseini, L., Jegou, H. The Faiss Library*. (n.d.). arXiv. https://arxiv.org/pdf/2401.08281
[7] *Ong, R. What Is Faiss (Facebook AI Similarity Search)?* (2024, July 3). Datacamp. https://www.datacamp.com/blog/faiss-facebook-ai-similarity-search
[8] *What is RAG? - Retrieval-Augmented Generation AI Explained - AWS*. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/what-is/retrieval-augmented-generation/
[9] Belcic, I. (2025, May 28). What is RAG (retrieval augmented generation)? *IBM*. https://www.ibm.com/think/topics/retrieval-augmented-generation
[10] Isac, N.B., Das, H. SLIP: A Sanskrit Linguistic Intelligence Pipeline for Enhanced Neural Machine Translation of Classical Texts: ResearchGate. (2025). *ResearchGate*. https://www.researchgate.net/publication/395774866_SLIP_A_Sanskrit_Linguistic_Intelligence_Pipeline_for_Enhanced_Neural_Machine_Translation_of_Classical_Texts
[11] Mandikal, P. & Department of Computer Science, UT Austin. (2024). Ancient Wisdom, Modern Tools: Exploring Retrieval-Augmented LLMs for Ancient Indian Philosophy. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)* (pp. 224–250). Association for Computational Linguistics. https://aclanthology.org/2024.ml4al-1.23.pdf
[12] Rahular. (n.d.). *GitHub - rahular/itihasa: A large scale Sanskrit-English translation dataset*. GitHub. https://github.com/rahular/itihasa
[13] *Bhagavad Gita Dataset*. (2022, December 22). Kaggle. https://www.kaggle.com/datasets/a2m2a2n2/bhagwad-gita-dataset
[14] PradhyumnaPrakash. (n.d.). *GitHub - PradhyumnaPrakash/Ancient-Indian-Philosophy-Based-RAG-Systems: This repository holds the source code for RAG systems built on the Bhagavad Gita and Itihasa.* GitHub. https://github.com/PradhyumnaPrakash/Ancient-Indian-Philosophy-Based-RAG-Systems