Verse2Vision: A Multimodal Retrieval-Augmented Generation System for Visual and Multilingual Explanation of Indian Epics

Team ID: T19

Aruni Saxena (SID: 202418006)
Heer Chokshi (SID: 202418019)
Paarth Patel (SID: 202418045)
Yash Deshmukh (SID: 202418063)

Dhirubhai Ambani University

December 17, 2025

Abstract

Verse2Vision is a multimodal Retrieval-Augmented Generation (RAG) system developed to explain verses from the Hanuman Chalisa using text, images, and audio narration. The project focuses on cultural authenticity by ensuring that all generated outputs are grounded in a curated local knowledge base. The system supports bidirectional interaction, allowing both text-to-image generation and image-to-text explanation. Additionally, multilingual audio narration is provided using Google Text-to-Speech (gTTS) to improve accessibility.

1 Introduction

Indian epics such as the Hanuman Chalisa have traditionally been preserved through oral narration and textual study. However, modern learners often find it difficult to engage with long textual explanations, and generative AI systems may introduce inaccuracies or cultural misinterpretations.

To address these challenges, we developed Verse2Vision, a controlled AI-based system that combines Retrieval-Augmented Generation with multimodal outputs. The objective of this project is not to generate new interpretations, but to present existing authentic knowledge in a more engaging, visual, and accessible manner while remaining faithful to the original verses.

2 System Overview

Verse2Vision is a bidirectional multimodal system capable of handling multiple forms of input and output. The system supports:

- Text-to-Image generation for visual storytelling
- Image-to-Text explanation of mythological scenes
- Question Answering based on verses
- Multilingual audio narration

The application is implemented in Python and deployed through a Streamlit-based web interface.

3 Technology Stack

The major technologies used in this project are:

- **Frontend**: Streamlit
- **Embedding Method**: TF-IDF (Scikit-learn)
- **Similarity Metric**: Cosine Similarity
- **LLM and Vision Models**: Google Gemini API
- **Image Generation**: Pollinations AI
- **Text-to-Speech**: Google Text-to-Speech (gTTS)

TF-IDF was selected because it is lightweight, interpretable, and suitable for small curated datasets without requiring GPU support.

4 Knowledge Base Design

The knowledge base is stored as a structured JSON file named `kb.json`. Each entry corresponds to a verse and contains:

- Sanskrit text

- Transliteration
- Simple and detailed meanings
- Contextual story explanation
- Image prompts
- Tags and emotional attributes

This file acts as the single source of truth. All generated content is strictly derived from these entries, ensuring authenticity and traceability.

5 Retrieval-Augmented Generation (RAG)

The RAG pipeline consists of the following steps:

1. User input is converted into a TF-IDF vector
2. Cosine similarity is computed with verse vectors
3. Top-k relevant verses are retrieved
4. Only the retrieved verses are passed to the language model

This approach significantly reduces hallucination and prevents the language model from adding information not present in the knowledge base.

6 Multimodal Workflows

6.1 Text-to-Image Generation

User text queries are first used to retrieve relevant verses. These verses are transformed into structured prompts and passed to an image generation API to produce comic-style illustrations with subtitles.

6.2 Image-to-Text Explanation

Uploaded images are analyzed using a vision model to generate a descriptive caption. This caption is then used as a query in the RAG pipeline to retrieve matching verses and generate an authenticated explanation.

7 Multilingual Audio Narration

Verse2Vision uses Google Text-to-Speech (gTTS) to convert generated text into audio narration. The system automatically detects the language based on script and keywords. Supported languages include:

- English
- Hindi
- Marathi
- Bengali
- Tamil
- Telugu
- Kannada
- Malayalam

This feature improves accessibility and allows users to consume content in their preferred language.

8 Cultural Preservation Aspect

The project demonstrates how AI can be used responsibly for cultural preservation. By grounding all outputs in authentic verses and avoiding external internet data, Verse2Vision maintains cultural accuracy while presenting traditional knowledge in a modern and engaging format.

9 Conclusion

Verse2Vision successfully integrates Retrieval-Augmented Generation with multimodal AI to provide accurate, engaging, and culturally grounded explanations of the Hanuman Chalisa. The project highlights the importance of controlled generation when applying AI to culturally sensitive domains. Future improvements include expanding the knowledge base to other epics and enhancing system scalability.