# Multi-Modal Approach for Summarization

**Popat Amee Dipak**
Department of Computer Science and Engineering (AI and ML), Dayananda Sagar University, Bangalore, Karnataka, India

**Kishan C**
Department of Computer Science and Engineering (AI and ML), Dayananda Sagar University, Bangalore, Karnataka, India

**K Himasree**
Department of Computer Science and Engineering (AI and ML), Dayananda Sagar University, Bangalore, Karnataka, India

**Neha M**
Department of Computer Science and Engineering (AI and ML), Dayananda Sagar University, Bangalore, Karnataka, India

## Abstract

This paper presents a platform for document intelligence that uses AI to process text, audio, and video documents through a unified Retrieval-Augmented Generation (RAG) architecture. The system addresses key issues found in existing solutions that handle only one type of data, lack contextual understanding, and offer basic analysis without proper source attribution. Our method combines Natural Language Processing, hierarchical summarization, and semantic search using vector embeddings to create context-aware responses with confidence scores. The system can process over 15 file formats, including PDFs, spreadsheets, images, audio, and video, using specialized preprocessing pipelines. A significant innovation is hierarchical chunking based on document complexity. This feature automatically adjusts chunk sizes, ranging from 800 to 3000 tokens, and determines overlap ratios to keep the meaning clear. For multilingual support, we use Unicode-aware text-to-speech engines for English, Hindi, and Kannada. We also include automated lip-sync animation and subtitle rendering. Evaluation with various datasets shows that we achieve 93-96% OCR accuracy on scanned documents, 97% recall in semantic retrieval, and a 4.7% Word Error Rate in audio transcription, outperforming open-source benchmarks. Human evaluation indicates that 92% of the generated responses are fully backed by retrieved evidence. The system uniquely merges these features into a platform that is ready for production, with scalable vector database management, real-time performance monitoring, and analytics-driven optimization. Its applications include enterprise document processing, educational content synthesis, and services for non-English speakers.

## Keywords

Audio Transcription, Document Intelligence, Multi-Modal Processing, Natural Language Processing, Retrieval-Augmented Generation, Vector Embeddings.