VIKI : AI powered Video Insights and Knowledge Interface

Amit Kumar Singh[a], Gayatri Yendamury[b], Mukesh Sharma[c], Shubham Yadav[d], Yogananda Tanguturi[e] and Sandya K M[f]

[a,d,f]Mtech in DSBA, IISc Bangalore
[b,c]Mtech in AI, IISc Bangalore
[e]Mtech in ECE, IISc Bangalore
ORCID (Amit Kumar Singh): https://orcid.org/amitsingh@iisc.ac.in, ORCID (Gayatri Yendamury): https://orcid.org/gayatriy@iisc.ac.in, ORCID (Mukesh Sharma): https://orcid.org/mukeshsharma@iisc.ac.in, ORCID (Shubham Yadav): https://orcid.org/shubhamy@iisc.ac.in, ORCID (Yogananda Tanguturi): https://orcid.org/yoganandat@iisc.ac.in, ORCID (Sandya K M): https://orcid.org/sandhyakm@iisc.ac.in

Abstract. The exponential growth of digital video content has created significant challenges in information retrieval, assessment, and accessibility, often requiring users to invest considerable time in manual exploration. This paper presents VIKI, an interactive AI-powered platform designed to transform video content navigation and comprehension. VIKI leverages deep learning and generative AI to generate instant video summaries, enable context-aware question answering, and facilitate smart navigation through natural language queries. The platform further enhances user engagement through auto-generated quizzes. Distinct from existing approaches that rely on static metadata, VIKI provides dynamic, content-based search and seamless interaction. Deployment of VIKI shall minimize time for exploration, overcome language barriers and enable active learning. By making video knowledge instantly searchable and understandable, VIKI redefines how individuals and organizations engage with large-scale video libraries, unlocking actionable insights and setting new standards for intelligent media interaction.

1 Introduction

The unparalleled rise in digital video content across educational and professional domain is challenging due to the too much time consumed during manual exploration and limited interaction. Traditional ways of video exploration generally depend on metadata such as titles or timestamps which often fail to capture the actual essence of the video. Thus, users are required to invest significant time and effort in manually browsing through lengthy videos to locate specific information or gain understanding. Limited interaction results in low engagement and inefficient method of learning since there is no structured way to extract knowledge, test the understanding and obtain insights. The motivation behind this work emerges from the need for an intelligent, engaging and optimized video content navigation. The main goal is to increase efficiency and improve learning outcomes of the users. This is a critical requirement for systems that go beyond keyword-based search and enable context-aware, content-driven interaction between user and the video.

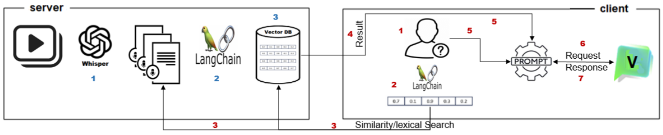
Our work presents VIKI (Video Insights and Knowledge Interface), an AI-powered platform addressing these challenges by utilizing advance deep learning and generative AI. VIKI provides a variety of functionalities such as concise video summarisation, natural language-based navigation, contextual question answering and quiz generation. By enabling dynamic, content-based search and intelligent interaction, VIKI redefines video comprehension, reduces time, bridges and enables efficient learning. This platform sets a new benchmark in how individuals and organizations can interact with video libraries to derive meaningful and actionable insights.

2 Related Work

Modern developments in deep learning and NLP have made intelligent systems for automated quiz generation, retrieval, and summarisation possible. A study presents an MCQ generator via Langchain and Gemini LLM. It is constructed with Streamlit, Langchain and google_genai. The study emphasizes cost-effective prompt engineering instead of model training[3]. Yet another study uses the BERTSUM model for extractive summarisation with RAKE for keyword extraction and WordNet to generate distractors. The research illustrates a hybrid method by integrating deep learning and linguistic resources for MCQ generation[5]. Video summarisation studies integrate transcript extraction, BERT-based summarisation, and image/audio processing. Tools such as the YouTube Transcript API and Google Translate API support multilingual capabilities[2]. For information retrieval, a Blended RAG system is used in the study which improves precision by integrating BM25, KNN and Elastic Learned Sparse Encoder (ELSER). Hybrid query methods incorporating dense, sparse and semantic search are tested on benchmarks such as SQuAD and Natural Questions resulting in enhanced retrieval accuracy in RAG pipelines[7]. In study on speech processing, a multilingual speech-to-text application employs MFCC features and classifies audio through SVM and MDC models.[1][6]. In another study, an audio summarisation system is developed using SpeechRecognition, PyAudio, and SpaCy. It applies tokenization, frequency-based analysis and sentence scoring to produce summaries. A graphical user interface is provided via Tkinter[4].

3 System Overview

Our VIKI system has a modular client-server architecture that enables smart interaction with videos through audio-to-text, semantic query, and generative AI algorithms. The system is split into two significant parts: server-side processing and client-side query Interaction.



- **server**
    1. Generate & Store transcripts using an open source AI model
    2. Splits/Embeds the transcripts (*)
    3. Store in vector database (*)
- **client**
    1. Client queries about the content of the video via the assistant
    2. Embed client query into a vector (*)
    3. Hybrid search (similarity and lexical search) over query in the static vector database (*)
    4. Fetch results of the hybrid search retriever followed by re ranker (*)
    5. Engineer prompt to include the transcript or search results(*) to the client query
    6. Request Qwen/Qwen3-30B-A3B Hugging Face model for response based on the final query (including the user's query and the transcript or search result (*))
    7. Fetch results from the model

* Required in case the video is lengthy

Figure 1. System Overview.

A. Server-Side Processing

The server handles tasks such as transcription, embedding, indexing, and generating LLM-based responses.

1. **Transcript Generation:**
   An open-source Automatic Speech Recognition (ASR) model (e.g., Whisper) parses the video to produce time-aligned transcripts. First, the audio is extracted from the video, after which it is transcribed into text.
2. **Segmentation and Embedding:**
   The transcripts are segmented into chunks and embedded into vector representations.
3. **Vector Storage:**
   The embeddings are stored in a vector database to enable efficient similarity and hybrid (semantic + lexical) search.

B. Client-Side Interaction

The client side provides an interface for users to interact with the system, handling user input and displaying the generated output.

1. **Query Input:**
   The user inputs a question pertaining to the video content.
2. **Hybrid Search and Re-ranking:**
   Lexical and semantic search is used to find relevant segments from the vector store.
   The retrieved segments are then re-ranked based on similarity and contextual appropriateness.
3. **Prompt Construction:**
   A prompt is constructed by combining the user query with the top-ranked relevant segments.
4. **Model Invocation:**
   A response is generated by an open source LLM hosted on local GPU machine using vllm which is context-aware, and this response is returned through the user interface.

4 Methodology

Our work is framed in a modular and Retrieval-Augmented Generation (RAG)-oriented architecture for intelligent video understanding. It supports functionalities such as summarisation, question answering, quiz generation, and natural language-based navigation.

1. **Transcription**
   Each video is transcribed using Whisper (an open-source ASR model) to produce reliable transcripts.MP3 audio file is extracted from the video using the FFmpeg library. These transcripts are then divided into semantically meaningful segments with corresponding timestamps. This text along with the timestamp is saved into the SQLite database that will be used for navigation in one of the modules.
2. **Embedding**
   To enable semantic search, each transcript segment is converted into a dense vector using Hugging Face's sentence-transformer model `all-MiniLM-L6-v2`. These vectors capture contextual semantics and are indexed in Milvus. This embedding and indexing operation enables similarity-based retrieval during user interaction.
3. **Context Construction via Hybrid Retrieval**
   When a user issues a natural language query—whether to summarize, ask a question, or navigate to a video segment—a hybrid retrieval method is employed. The query is embedded and compared against transcript vectors in Milvus, and the most relevant segments are retrieved.
4. **Local LLM Inference with vLLM**
   Prompt execution is handled by a self-hosted server using vLLM. Generated outputs such as summaries, Q&A responses, and quizzes are stored for future reuse. We are using the open-source LLMs self-hosted on a GPU machine using the vllm library. Vllm provides OpenAI and Langchain compatible APIs, which we use to implement modules. For longer transcripts, we break them down into sub-documents. Each sub-document is then passed as context in the summarisation prompt. For the Q&A module, relevant context is retrieved from vector storage using lexical search. The merged results are ranked using a re-ranker before being passed to the prompt for answering a specific question.
5. **Modularity and Task-Specific Engines**
   The design is structured into modular components, each responsible for a particular task. Core modules include:
   - `ContextManager`: Fetches and constructs meaningful transcript pieces to provide relevant context for downstream tasks.
   - `LangchainInterface`: Manages prompt construction and interacts with the LLM via API calls for inference.
   - `Task Engines`: Specialized modules for specific workflows,including:summarisationEngine,QAEngine,QuizGenerator, NavigationEngine

This methodology bridges the gap between unstructured video data and interactive comprehension by leveraging transcription, semantic indexing, hybrid retrieval, and LLM reasoning. Its modularity, scalability, and real-time response make it suitable.

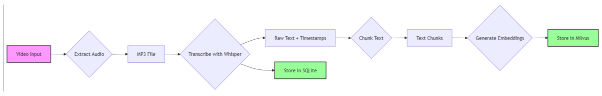


Figure 2. Data Processing Pipeline.

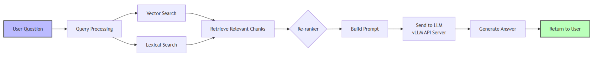


Figure 3. Q&A Module Flow.

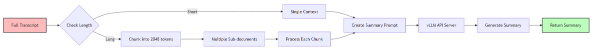


Figure 4. Summarisation Module Flow.

5 Evaluation

For the evaluation, we have selected videos from Andrej Karapathy's YouTube channel and used their transcripts as the gold standard for the original transcript. For evaluating the performance of the Whisper models, we assessed the performance of base.en, tiny.en, small.en, medium.en, large-v2 models. We have considered the following metrics:

1. **WER (Word Error Rate):** It is a useful metric for comparing texts by considering number of substitutions, deletions, hits and insertions.
2. **MER (Match Error Rate):** It is similar to WER, but it focuses more on the proportion of words that are incorrectly matched.
3. **WIL (Word Information Lost):** It measures the extent of information loss.
4. **CER (Character Error Rate):** It evaluates accuracy at the character level.
5. **BLEU (Bilingual Evaluation Understudy):** It assesses the similarity by comparing the overlapping n-grams.
6. **Precision:** It indicates the proportion of correctly identified words out of all words predicted by the model.
7. **Recall:** It identifies the proportion of correctly identified words out of all words.
8. **F1-Score:** It offers a balanced measure of false positives and false negatives.

After evaluating all the metrics, including the time taken to transcribe a video and extract the MP3 file, we decided to proceed with the base.en model [Figure 6].

A. Q&A Evaluation

We compared three open-source language models—`Qwen3-30B-A3B`, `Gemma-3-27b-it`, and `Mistral-Small-3.1-24B-Instruct`—on a question-answering (Q&A) task across five different prompts. The evaluation uses multiple metrics to assess different aspects of response quality:

- **BLEU**: Measures n-gram overlap between generated and reference responses, focusing on precision.
- **ROUGE (1, 2, L)**: Evaluates recall-oriented overlap at unigram, bigram, and longest common subsequence levels.
- **METEOR**: Considers synonyms and paraphrases beyond exact matches.
- **BERTScore**: Leverages contextual embeddings to capture semantic similarity at the token level.

The five prompts evaluate distinct Q&A capabilities:

- Synthesis and Integration (Prompt 1)

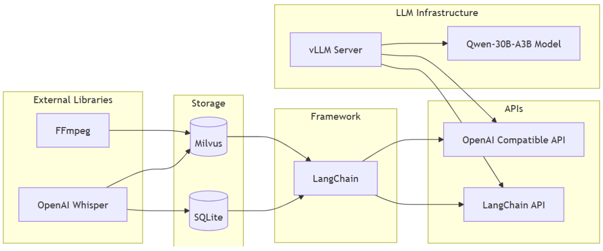


Figure 5. Components Integration Flow Module Flow.

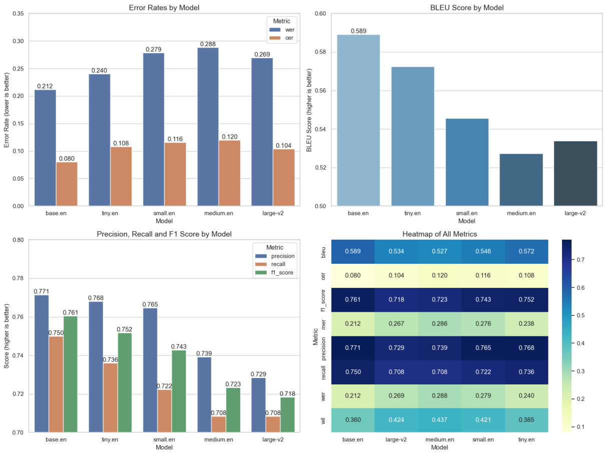


Figure 6. Whisper Models Evaluation

- Chain-of-Thought Reasoning (Prompt 2)
- Structured Data Extraction to JSON Format (Prompt 3)
- Contradiction Handling (Prompt 4)
- Comparative Analysis with Evidence Sourcing (Prompt 5)

`Mistral-Small-3.1-24B-Instruct` demonstrated the strongest overall performance, particularly excelling in Prompt 2 (BLEU = 0.3808, ROUGE-1 = 0.5666) and Prompt 3 (METEOR = 0.9815, BERTScore F1 = 0.9746). This model shows superior ability to generate responses with both high lexical overlap and semantic similarity to reference answers.

While `Gemma-3-27b-it` shows competitive performance across most metrics, particularly in semantic similarity measures (BERTScore), it demonstrates more balanced performance across prompts. `Qwen3-30B-A3B`, despite having the largest parameter count, shows comparatively lower scores in lexical overlap metrics (BLEU, ROUGE) but maintains reasonable semantic similarity scores.

B. Summarisation Evaluation

The five prompts evaluate distinct Summarisation capabilities:

- Conciseness and Brevity (Prompt 1)
- Structured summarisation (Prompt 2)
- Audience-Specific summarisation (Prompt 3)
- Abstractive vs. Extractive Ability (Prompt 4)
- Comparative summarisation (Prompt 5)

The evaluation reveals distinct performance patterns across the three models. `Mistral-Small-3.1-24B-Instruct` demonstrates superior performance in traditional n-gram metrics, particularly excelling in ROUGE scores across most prompts, with

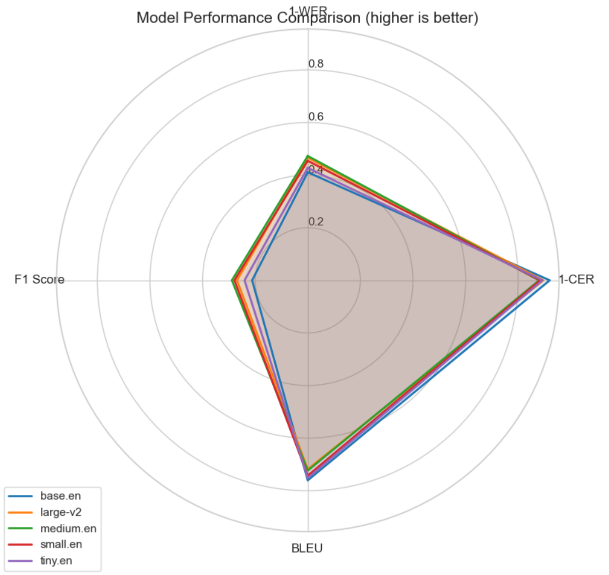


Figure 7. Evaluation Summary

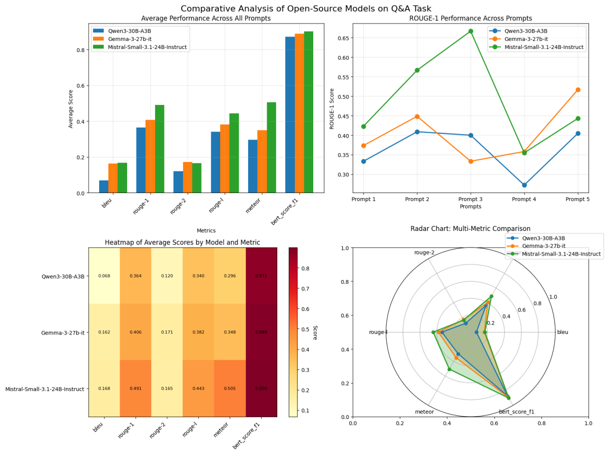


Figure 8. Comparative Analysis of Open-Source Models on Q&A Task

notable strength in Prompt 1 (ROUGE-1: 0.2593) and Prompt 3 (ROUGE-1: 0.2661). `Qwen3-30B-A3B` shows competitive performance with particularly strong results in Prompt 2, achieving the highest ROUGE-1 score (0.4333) and demonstrating better consistency in BERTScore metrics. `Gemma-3-27b-it`, while showing lower scores in traditional metrics, maintains competitive BERTScore values, suggesting it may produce semantically meaningful summaries despite lower lexical overlap. The high BERTScore values across all models (generally above 0.85) indicate that all three models produce semantically coherent summaries, though with varying degrees of extractive precision.

6 Conclusion and Future Scope

This work presents AI powered video insights and knowledge , a Video understanding platform that combines transcription, semantic indexing and retrieval-augmented generation (RAG) with locally hosted LLMs to enable intelligent video interaction. Key functionalities include summarization, question answering, quiz generation and natural language based query navigation. The system employs Whisper for ASR, Hugging face sentence-transformers for embeddings, Milvus for vector retrieval and locally hosted vLLM for low-latency local inference. Its modular architecture supports extensibility and efficient task execution.

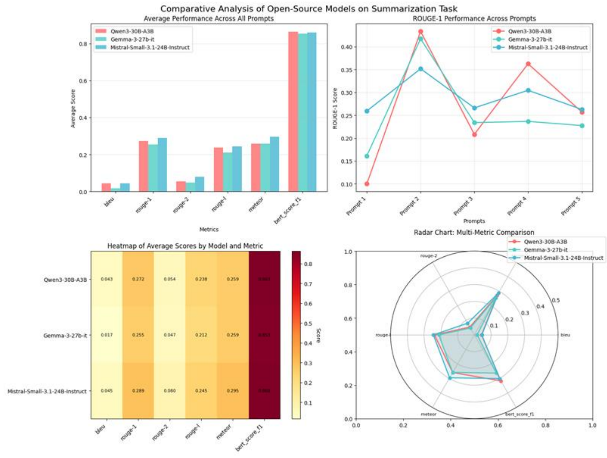


Figure 9. Comparative Analysis of Open-Source Models on Summarisation Task

Future scope include multimodal context integration,real-time streaming support, multilingual accessibility. Improved retrieval strategies can also be incorporated to expand the system's adaptability and usability with large-scale video libraries to enable seamless interaction.

7 Roles and Responsibilities

Stage	**Description**	**Responsibility**
Problem Definition	Use Case Identification and Planning	Team
Data Preprocessing	1) Video → Audio → Text (Transcribing) 2) Embedding transcript chunks	Gayatri Yendamury
Module Development	1) Established the infrastructure to host an open-source LLM on local GPU machine for downstream tasks 2) Context Manager Implementation (RAG)	Amit Kumar Singh
	QAEngine	Yogananda Tanguturi
	SummarisationEngine	Mukesh Sharma
	QuizGenerator	Shubham Yadav
	NavigationEngine	Sandya K M
Orchestration	Integration of all modules	Amit Kumar Singh
Documentation	Preparation of Report	Team

Table 1. Development Stages, Descriptions, and Assigned Team Members

References

[1] Y. H. Ghadage and S. D. Shelke. Speech to text conversion for multilingual languages. In *2016 International Conference on Communi-*

cation and Signal Processing (ICCSP), pages 0236–0240, 2016. doi: 10.1109/ICCSP.2016.7754130.

[2] I. P, N. D, T. A., M. M, M. S, N. A.S, and I. I. Video transcript summarizer. *E3S Web of Conferences*, 399, 07 2023. doi: 10.1051/e3sconf/202339904015.

[3] P. Pawar, R. Dube, A. Joshi, Z. Gulhane, and R. Patil. Automated generation and evaluation of multiplechoice quizzes using langchain and gemini llm. In *2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT)*, volume 1, pages 1–7, 2024. doi: 10.1109/ICEECT61758.2024.10739326.

[4] A. K. R. P. S. T. Pravin K, Sanket G. Audio data summarization system using natural language processing. *International Research Journal of Engineering and Technology (IRJET)*, 6(9):1770–1773, 2019. URL https://www.irjet.net/archives/V6/i9/IRJET-V6I957.pdf. IRJET, Volume 6, Issue 9, September 2019.

[5] C. M. D. C. M. R. Pritam Kumar M, Prachi J. Automated mcq generator using natural language processing. *International Research Journal of Engineering and Technology (IRJET)*, 8(5):2237–2240, 2021. URL https://www.irjet.net/archives/V8/i5/IRJET-V8I5497.pdf. IRJET, Volume 8, Issue 5, May 2021.

[6] P. D. Reddy, C. Rudresh, and A. S. Adithya. Multilingual speech to text using deep learning based on mfcc features. *Machine Learning and Applications: An International Journal (MLAIJ)*, 9(2):21–29.

[7] K. Sawarkar, A. Mangal, and S. R. Solanki. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 155–161, 2024. doi: 10.1109/MIPR62202.2024.00031.

8 Appendix

GitHub Code Respository is provided in the below link: https://github.com/amitsingh2409/DA225o-Project-AI-Powered-Video-Analyzer