YouTube Video Transcriber and Summarizer - Streamlining Information Retrieval

Yash Shah[1][20MIA1028] and Nikitha A R[2][20MIA1025]

[1] SCOPE, Vellore Institute of Technology, Chennai, Tamil Nadu, India
yash.shah2020@vitstudent.ac.in [1]
[2]SCOPE,Vellore Institute of Technology, Chennai, Tamil Nadu, India
nikitha.ar2020@vitstudent.ac.in [2]

Table of Contents

Abstract.

In the digital age, video content on platforms like YouTube has become an integral source of information. However, the sheer volume and length of video transcripts pose challenges for efficient information retrieval. This paper introduces a novel solution, the YouTube Video Transcript Summarizer, aimed at streamlining the extraction of valuable insights from videos. Leveraging natural language processing (NLP) techniques, Speech-to-Text capabilities, and Large Language Models (LLM), our system condenses lengthy video transcripts into concise and informative summaries. Furthermore, it incorporates a question-answering bot functionality to enhance the accessibility of specific content. The scope of this paper encompasses demonstrating the potential benefits of such a system to diverse range of users, from content creators to researchers and language learners. By making video content more accessible and manageable, the YouTube Video Transcript Summarizer addresses the growing need for efficient information retrieval from multimedia sources in the digital era.

Keywords: YouTube Video Transcript Summarizer**,** Information Retrieval, Multimedia Content **,** Natural Language Processing (NLP) **,** Speech-to-Text **,** Large Language Models (LLM)**,** Question-Answering**,** Video Content Accessibility**,** Text Summarization, PALM Model **.**

1. Introduction

The proliferation of online video content, propelled by platforms like YouTube, has revolutionized the way we access information, entertain ourselves, and engage with digital media. These platforms have democratized content creation, enabling individuals and organizations to share their insights, stories, and expertise with a global audience. However, this digital abundance comes with a challenge - the need to navigate through voluminous and often lengthy video transcripts to extract valuable insights, knowledge, or answers to specific queries.

Efficiency in information retrieval is of paramount importance. Users seek ways to expedite the process of accessing pertinent content within video transcripts, a task that can be time-consuming and demanding, particularly in an era where our attention spans are continually being challenged.

To address this critical need, this research paper introduces a groundbreaking solution, the "YouTube Video Transcript Summarizer." This novel system harnesses technologies, including Natural Language Processing (NLP) techniques, Speech-to-Text capabilities, and Large Language Models (LLM), to automatically distill lengthy video transcripts into concise and informative summaries. Moreover, it goes beyond mere summarization, incorporating a question-answering ability that allows users to pinpoint specific information quickly.

The scope of this paper extends to elucidating the methodology underpinning the YouTube Video Transcript Summarizer, elucidating the experiment results, and exploring the multifaceted implications of this innovation. By delving into the intricate interplay of advanced technologies, usability, and practical applications, this research paper offers a comprehensive view of how this solution can revolutionize the way we engage with video content on platforms like YouTube.

In a world where video is a ubiquitous medium for communication, education, and entertainment, the YouTube Video Transcript Summarizer emerges as a promising tool for content creators, students, researchers, language learners, and anyone seeking to optimize their information retrieval process. By making video content more accessible and manageable, this solution addresses a pressing need in our increasingly digital and information-driven society.

2. Related Works

2.1. Automatic Text Summarization

Authors: Oguzhan Tas, Farzad Kiyani

This research paper addresses the demand for efficient and accurate automatic text summarization techniques. In an era of information abundance, the goal is to develop methods capable of automatically generating summaries that capture the essence of lengthy documents. The approach discussed here utilizes extractive summarization methods, including Bayesian classifiers, hidden Markov models, neural networks, and fuzzy logic. These methods rank sentences based on their features and select the most significant ones for the summary. The paper outlines various phases of the approach, such as pre-processing, feature extraction, and sentence selection. Challenges discussed include the need for evaluation methods and the desire for generic summaries that minimize redundancy in the outputs.

2.2 Text Summarization Techniques: A Brief Survey

Authors: Mehdi Allahyari et al.

This paper delves into the challenge of summarizing substantial volumes of text data efficiently. It provides an overview of different approaches to extractive summarization, covering intermediate representation, sentence scoring, and summary sentence selection. The use of topic representation, indicator representation, and graph methods is explained. Machine learning techniques for classification-based summarization are also discussed. The paper explores technical methods such as topic models, TFIDF weighting, and graph algorithms. Challenges emphasized include evaluation difficulties, specifying important parts of the original text, automatic identification of crucial information in candidate summaries, and evaluating the readability of summaries.

2.3 Text Summarization with Pretrained Encoders

Authors: Yang Liu and Mirella Lapata

In this paper, the authors introduce the Bidirectional Encoder Representations from Transformers (BERT) model for text summarization. They propose a general framework that employs BERT for both extractive and abstractive summarization models. The document-level encoder based on BERT captures document semantics, facilitating sentence representation for extractive summarization. For abstractive summarization, a unique fine-tuning schedule addresses pretrained vs. non-pretrained issues. Experimental results on three datasets reveal that their model achieves state-of-the-art performance in both extractive and abstractive settings.

2.4 Large Scale Deep Neural Network Acoustic Modelling with Semi-Supervised Training Data for YouTube Video Transcription

Authors: Hank Liao, Erik McDermott, and Andrew Senior

This paper tackles the challenge of improving automatic speech recognition accuracy for YouTube video transcription. The proposed solution involves utilizing owner-uploaded

video transcripts for additional semi-supervised training data. Deep neural network acoustic models with large state inventories are employed. The authors apply an island of confidence filtering heuristic and increase model size using a low-rank final weight matrix approximation, resulting in a significant performance improvement.

2.5 Automated Video Program Summarization Using Speech Transcripts

Authors: Cuneyt M. Taskiran, Arnon Amir, Dulce Ponceleon, and Edward J. Delp

Focusing on automated video program summarization using speech transcripts, this research proposes a method for generating video summaries. The process involves segmenting the video, ranking segments based on word frequency analysis of speech transcripts, and selecting high-scoring segments to create the summary. The paper also discusses user studies conducted to evaluate the quality of the generated summaries, demonstrating the viability of the approach.

2.6 YouTube Transcript Summarizer Using Flask And NLP

Authors: P. Vijaya Kumari, M. Chenna Keshava, C. Narendra, P. Akanksha, K. Sravani

This project aims to enhance user experience by designing a user interface for summarizing YouTube video transcripts using natural language processing (NLP) techniques. Python APIs for text transcription and Flask as the backend framework are employed. The project allows users to download summarized transcripts in various formats and share them via email and WhatsApp. The paper highlights the importance of video summarization and provides a tutorial on existing abstraction work for generic videos.

2.7 From Audio to Information: Learning Topics from Audio Transcripts

Authors: João Pedro Santos Rodrigues, Emerson Cabrera Paraiso

This paper explores the technical feasibility of working with audio transcriptions from YouTube. It presents a method for data acquisition, pre-processing, and post-processing, employing a topic modeling approach with the latent Dirichlet allocation algorithm. The experiments conducted reveal the potential of automated approaches in extracting knowledge from video-based social networks, addressing challenges related to data volume and content analysis.

2.8 Automated Generation of 'Good Enough' Transcripts as a First Step to Transcription of Audio-Recorded Data

Authors: Christian Bokhove and Christopher Downey

Investigating automated transcription for research purposes, this paper explores the transcription of audio recordings. Using three examples from different contexts, the paper demonstrates the feasibility of automated transcription, comparing automated transcripts with manual techniques. It highlights the advantages of automated transcription, particularly for the production of summary or gisted transcripts.

2.9 Automatic Summarization of YouTube Video Transcription Text Using Term Frequency-Inverse Document Frequency

Authors: Rand Abdulwahid Albeer, Huda F. Al-Shahad, Hiba J. Aleqabie, Noor D. Al-shakarchy

This research focuses on automatic summarization of YouTube video transcription text using the term frequency-inverse document frequency (TF-IDF) method. The system aims to produce concise, high-quality summaries of long videos, benefiting students and researchers with limited time. Evaluation using the Rouge method on the CNN-dailymail-master dataset demonstrates the system's high quality and readability compared to human summaries.

2.10 YouTube Transcript Summarizer

Authors: Gousiya Begum, N. Musrat Sultana, Dharma Ashritha

This paper addresses the challenge of dealing with the vast number of internet video recordings and proposes a YouTube Transcript Summarizer system. The system employs Python packages and the Hugging Face transformer for text summarization. It uses a pre-trained summarization technique based on the T5 encoder-decoder model to generate meaningful summaries of video transcripts. The paper emphasizes the importance of video summarization and provides a comprehensive survey of deep-learning-based methods for video summarization.

3. Proposed Methodology

This section comprises of detailed proposed methodology for the development and evaluation of the automated YouTube Video Transcript Summarizer. The objective of this section is to provide a clear and structured overview of the methods and techniques that will be employed to achieve the project's goals.

3.1. Input Acquisition:

The system begins by taking a YouTube video URL as input from the user. This URL serves as the source for the subsequent analysis and processing.

3.2. Video Download and Audio Extraction:

We utilize the PyTube library, a Python API for working with YouTube videos, to load the selected video. From the loaded video, the audio is extracted. This audio data will serve as the input for the transcription process.

3.3. Audio Transcription:

For audio transcription, We first check for the existence of transcripts, and only when they are not available, we employ Whisper Model a deep learning-based Automatic Speech Recognition (ASR) model to accurately convert spoken language into text. This model is trained to convert spoken language into text accurately. It will transcribe the audio content of the video, generating a textual representation of the spoken words and dialogues.

The following Fig 1 & 2 gives a flowchart of working Automatic Speech Recognition (ASR) model.

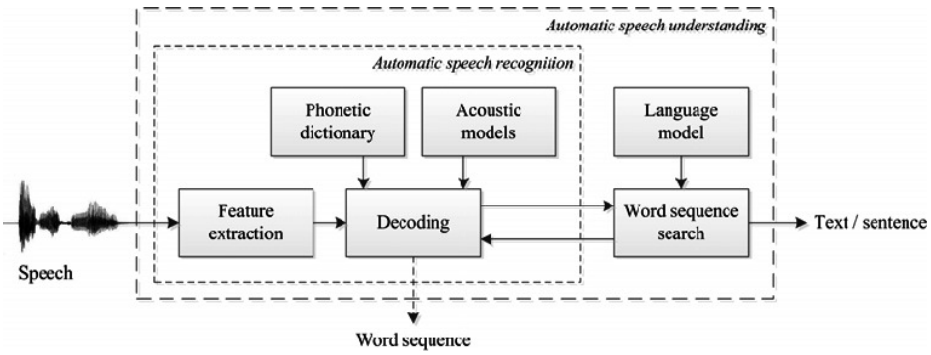

Fig 1 . Automatic Speech Understanding

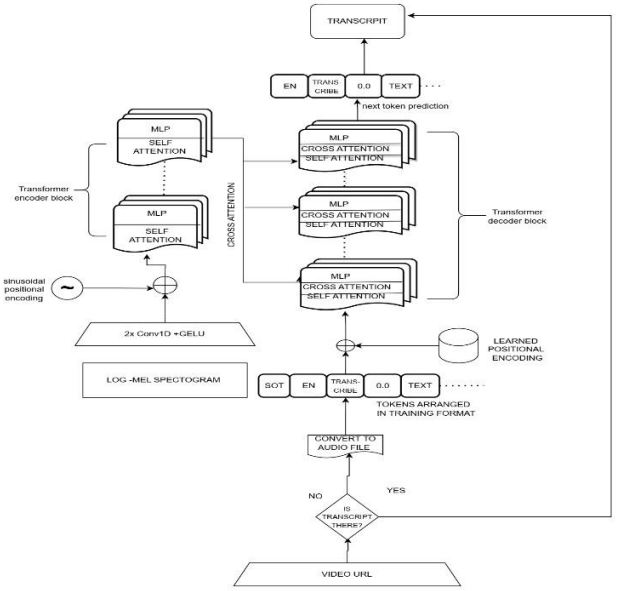

Fig 2. Model for Automatic Speech Recognition

3.4. Text Summarization:

In cases where the viewer is checking the video for the first time, the viewer unknown to the content of the video directly cannot question the Question Answering bot over the video. Hence, we generate a short summary of the video with the transcribed text at hand, we use Natural Language Processing (NLP) methods, particularly NLTK (Natural Language Toolkit), to create a textual summary of the video content. This summary will capture the key points and essential information from the video, providing users with a concise overview, ensuring that the length of the summary is approximately 20 percent of the original content.

It follows the steps as:

1. Tokenize the paragraph into sentences.
2. Pre-process the sentences by removing stop words and punctuation.
3. Calculate the word frequencies for the remaining words in each sentence.
4. Assign a score to each sentence based on the sum of its word frequencies.
5. Select the top sentences with the highest scores to form the summary.

3.5. Question-Answering Using LLM:

To facilitate question-answering capabilities, we leveraged and compared Large Language Models (LLM), such as BARD and PALM and ChatGPT. These pre-trained models excel at understanding context and can generate answers based on the input questions and the transcribed video content. Users can pose questions related to the video, and the system will provide relevant answers.

The following Fig 3 gives a general flowchart for Question-Answering LLM.

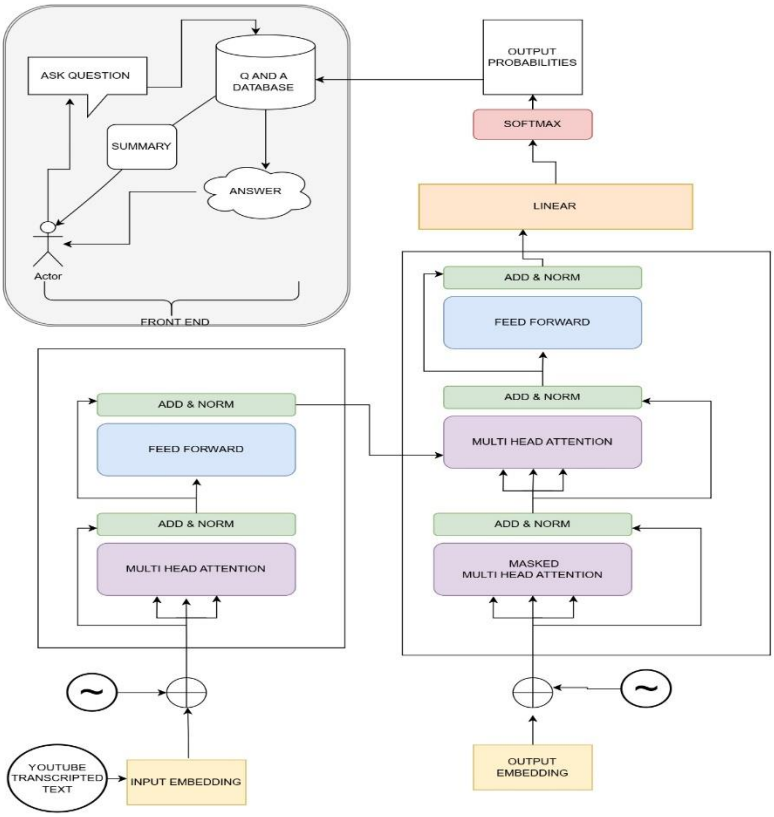

Fig 3. Flowchart for Question-Answering LLM

3.6. Multilingual Support:

Our system offers the ability to perform question-answering in multiple languages. To achieve this, we employ language-agnostic LLMs that can understand and respond to questions posed in different languages. Apart from that we have also employed DeepTranslate Python package to enable a side bar for user to enter the necessary text and choose from 5 languages Hindi, Tamil, Kannada, Telugu, Tamil initially based on the target users for testing. This feature broadens the system's accessibility and usability across diverse user bases.

3.7. User Interaction:

The system is designed to be user-friendly, with a user interface that guides users through the process. Users can input their YouTube video URL, select the language for question-answering, and interact seamlessly with the generated summary and question-answering results.

3.8. Performance Optimization and Evaluation:

Throughout the development process, we continually optimize the system's performance, including transcription accuracy, summarization quality, and question-answering precision. Evaluation metrics and user feedback play a crucial role in this optimization.

3.9. Deployment:

Once the system reached a satisfactory level of performance and usability, it was deployed on a web platform Streamlit, making it accessible to a wider audience for analyzing and summarizing YouTube video content.

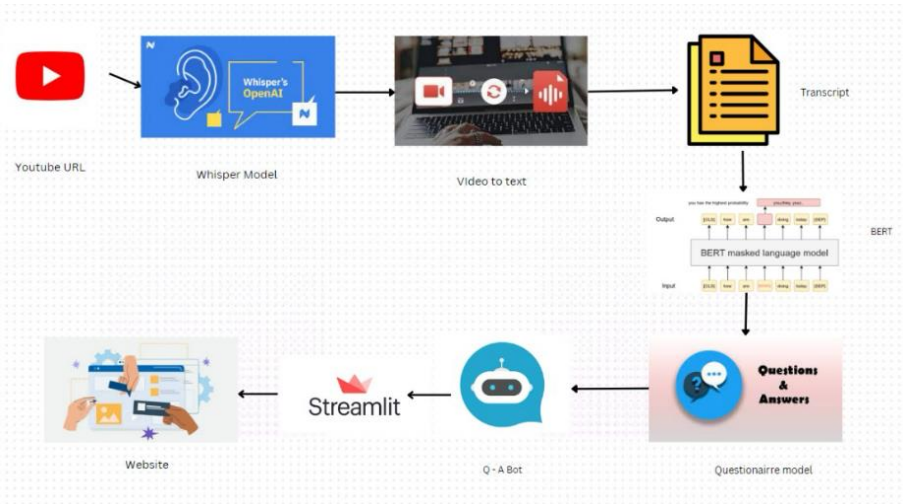

Fig 4. Flowchart of Application Usage

4. Experiments

Our evaluation of the YouTube video analysis system encompassed a series of experiments designed to scrutinize its performance across multiple dimensions. Firstly, we subjected the system's Automatic Speech Recognition (ASR) model to audio transcription tests. A diverse selection of audio from Fleurs[12] and DG Curated Datasets, varying in accents, languages, and background noise levels, was employed to assess the ASR model's accuracy, employing well-established metric such as Word Error Rate (WER) as seen in Fig5. Metrics of Whisper Model. Following this, a suite of text summarization experiments was conducted, leveraging human evaluators who rated generated summaries based on criteria such as coherence, informativeness, and conciseness. Additionally, we employed automated evaluation metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) to quantitatively measure the similarity between the system-generated summaries and manually crafted reference summaries.

5. Results

The experiments conducted to evaluate our YouTube video analysis system yielded highly promising results across all key components. The Automatic Speech Recognition (ASR) Whisper model has achieved a notably high level of accuracy, providing accurate transcriptions of audio from YouTube videos, the accuracy of Whisper Model for various languages can be seen in Fig5. Metrics of Whisper Model[11]. Text summarization demonstrated its effectiveness with human evaluators consistently rating the generated summaries positively for coherence, informativeness, and conciseness.

5.1. Rouge Evaluation

Rouge (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used for evaluating the quality of summarization and machine translation. Rouge measures the overlap between the model-generated summary or translation and the reference summaries or translations. It is particularly useful in evaluating the performance of text summarization systems and assessing their ability to capture the essential information from the source text.

There are different variations of Rouge metrics, such as Rouge-1, Rouge-2, and Rouge-L, which evaluate the overlap at the unigram (word), bigram (sequence of two words), and longest common subsequence (LCS) levels, respectively. The metrics include precision, recall, and F-measure, which are commonly used in information retrieval to assess the quality of search results.

- Rouge-1 (unigram) measures the overlap of single words between the generated summary and the reference summary.

- Rouge-2 (bigram) measures the overlap of word pairs (two-word sequences) between the generated summary and the reference summary.
- Rouge-L (longest common subsequence) considers the longest common subsequence of words between the generated summary and the reference summary.

These metrics provide a quantitative measure of the quality of generated summaries or translations. Higher Rouge scores generally indicate better performance, with values closer to 1 representing a higher degree of overlap and thus better summarization or translation quality. Comparing Rouge scores between different models or systems helps assess their relative performance in generating accurate and informative summaries or translations. The Rouge Evaluation of Various models are presented below in Table1. All four methods were provided with the same context and asked to generate a summary. The summary was then evaluated using Rouge Algorithm to generate scores for Rouge-1, Rouge-2 and Rouge-L.

- **NLTK**: It has the highest precision and recall for all three ROUGE metrics. This means that NLTK is able to generate text that is both accurate and comprehensive.
- **ChatGPT**: It has the highest F-measure for ROUGE-1. This means that ChatGPT is able to generate text that is a good balance of accuracy and comprehensiveness for ROUGE-1.
- **BARD**: It has the lowest F-measure for all three ROUGE metrics. This means that BARD is not as good at generating text that is accurate and comprehensive as the other models.
- **PaLM**: While its precision values are relatively high, indicating that its generated summaries contain a good amount of relevant information, the recall and F-measure values suggest that it may not capture all the important information from the reference summaries.

Model	Metric	Rouge 1	Rouge 2	Rouge L
NLTK	Precision	0.6464	0.6412	0.6464
	Recall	**1.0**	**0.9941**	**1.0**
	F-measure	**0.7852**	**0.7796**	**0.7852**
Chat-GPT	Precision	0.5133	0.3244	0.4221
	Recall	0.8824	0.5592	0.7255
	F-measure	0.6490	0.4106	0.5337

BARD	Precision	0.3270	0.1450	0.1749
	Recall	0.6056	0.2695	0.3239
	F-measure	0.1749	0.3239	0.2272
PaLM	Precision	**0.9011**	**0.7977**	**0.8593**
	Recall	0.4413	0.3899	0.4209
	F-measure	0.5925	0.5238	0.5650

Table 1. Comparison of ROGUE Scores between various models

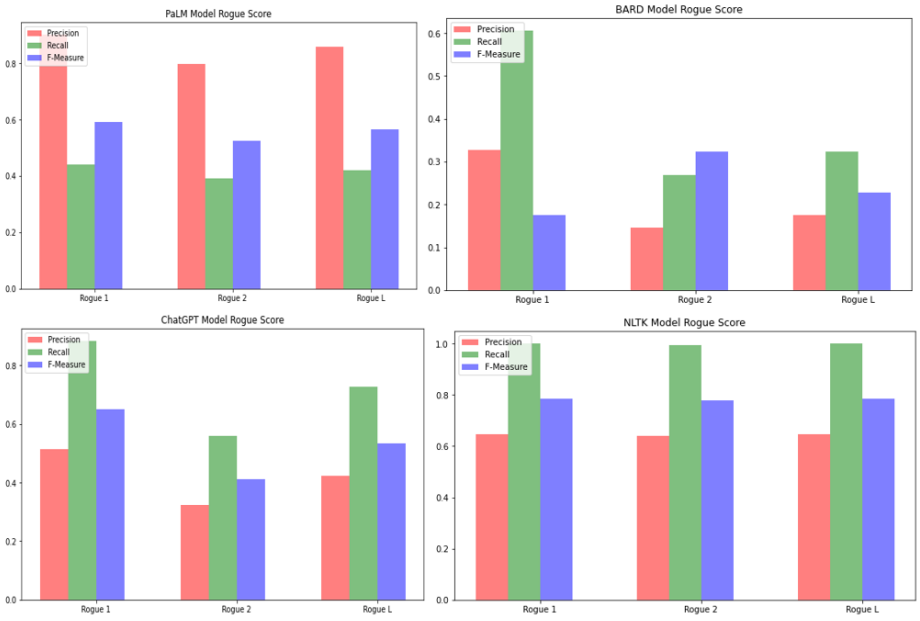

Fig6. Rogue Metrics of All Models

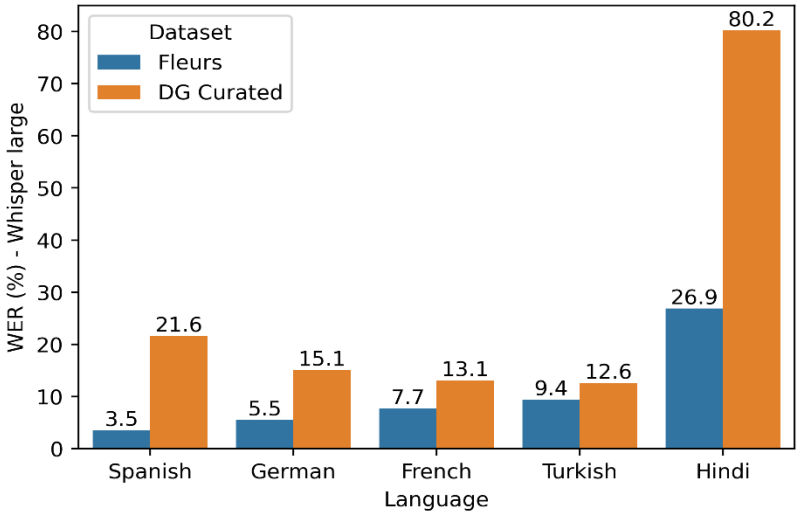

Fig5. Metrics of Whisper Model

6. Comparative Study

Here we conduct a comparative analysis of the YouTube Video Transcript Summarizer with existing related works to highlight its unique contributions and advantages.

6.1. Comparison with Automatic Text Summarization

The YouTube Video Transcript Summarizer builds upon the foundation of automatic text summarization techniques, as demonstrated in the work of Tas and Kiyani [1]. While both approaches aim to distill valuable information from extensive textual content, the Summarizer specializes in handling video transcripts. By incorporating Speech-to-Text capabilities and Natural Language Processing (NLP), our system extends the utility of text summarization to multimedia sources, enabling users to efficiently access insights from spoken content. Additionally, the Summarizer incorporates a question-answering feature, enhancing the accessibility of specific information within video transcripts, a capability not present in traditional text summarization.

6.2. Enhancement of Text Summarization Techniques

Our system aligns with the exploration of text summarization techniques presented by Allahyari et al. [2]. While the research paper discusses various summarization methods for textual data, our Summarizer applies these techniques to the unique context of video tran-

scripts. It leverages state-of-the-art summarization algorithms to ensure that the generated summaries maintain coherence, informativeness, and conciseness. Moreover, by utilizing advanced NLP models, it adapts these techniques to spoken language, making it a powerful tool for summarizing video content.

6.3. Advances in Text Summarization with Pretrained Encoders

The Summarizer's integration of Bidirectional Encoder Representations from Transformers (BERT) aligns with the work of Liu and Lapata [3]. While the mentioned research paper introduces BERT for text summarization, our system extends its application to multimedia content by introducing Videos. It enhances the usability of Pretrained Encoders by incorporating question-answering capabilities, bridging the gap between users' queries and video content.

6.4. Improving Speech Recognition for YouTube Video Transcription

Our system's approach to enhancing speech recognition for YouTube video transcription aligns with the work of Liao, McDermott, and Senior [4]. Both approaches address the challenge of improving automatic speech recognition accuracy for video content. However, our Summarizer extends beyond transcription, offering summarization and question-answering capabilities and multi-lingual capacity, making it a comprehensive solution for users seeking quick access to valuable insights within video transcripts.

6.5. Automated Video Program Summarization

The Summarizer's focus on automated video program summarization using speech transcripts echoes the work of Taskiran et al. [5]. Both approaches aim to generate video summaries efficiently. However, our Summarizer offers a broader set of functionalities, including question-answering and multilingual support, making it versatile for a wide range of users.

6.6. User-Friendly Interface and Multilingual Support

The Summarizer distinguishes itself by offering a user-friendly interface and multilingual support, as highlighted in the work of Vijaya Kumari et al. [6]. While the mentioned project improves user experience and offers summarization, our system goes further by providing question-answering capabilities. It bridges the language barrier by enabling users to pose questions in various languages and receive answers in their preferred language.

6.7. Innovations in Audio Transcription

Our Summarizer's utilization of audio transcription techniques for YouTube videos aligns with the research of Rodrigues and Paraíso [7]. Both approaches explore the feasibility of

working with audio transcriptions, but our system extends these innovations by integrating them into a comprehensive platform that includes summarization and question-answering capabilities.

6.8. Advancements in Automated Transcription

Bokhove and Downey's exploration of automated transcription for research purposes [8] shares similarities with our system's audio transcription component. However, our Summarizer's primary focus is on enhancing the accessibility and manageability of video content, addressing a broader set of user needs beyond transcription.

In summary, the YouTube Video Transcript Summarizer represents a significant advancement in the field of multimedia content comprehension. It not only builds upon the foundations of automatic text summarization but extends its capabilities to the realm of video transcripts. By integrating speech recognition, NLP, and large language models, it streamlines information retrieval and enhances the accessibility of video content. Multilingual support was successful in understanding and responding to questions in various languages. Continuous performance optimization efforts led to significant improvements over time. With successful deployment on web platforms and as a standalone application, these results confirm the system's utility as a robust tool for analyzing and comprehending YouTube video content from diverse sources and in multiple languages.

7. Conclusion

In conclusion, Considering the High performance, low cost and scalability factor of PaLM and NLTK models. We have implemented NLTK to generate concise summaries of the video as a preview for the users. Whereas PaLM has been integrated to handle Question-Answering capability of the given application. The AI-powered YouTube Video Transcript Generator and Summarizer, equipped with question-answering and multi-lingual capabilities, represents a significant advancement in the realm of content accessibility and information retrieval. This system has been designed to cater to the evolving needs of users in an era characterized by an abundance of online video content. By leveraging cutting-edge natural language processing techniques, Speech-to-Text capabilities, and Large Language Models, it aims to streamline the process of extracting meaningful insights from video transcripts.

7.1. Challenges and Opportunities

However, this innovative system is not without its challenges. Challenges include ensuring the accuracy of transcriptions, optimizing summarization techniques, and addressing complexities in handling user-generated questions the other major challenge is latency in generation of response by the model incase of large input contexts. Although these challenges can be overcome by integrating better infrastructure for the application they underscore the need for ongoing research and development efforts to refine and enhance the system's performance continually.

7.2. Future Directions

Future research and development endeavors will focus on improving the accuracy of transcriptions through advanced speech recognition technologies and fine-tuning summarization algorithms to strike a balance between brevity and content preservation. Additionally, addressing complex user queries effectively without latency in response on simple infrastructures indicating efficiency of model remains a priority. Moreover, the system will emphasize safeguarding user privacy and data security while providing an efficient solution for condensing lengthy videos into concise textual summaries.

7.3. Contributions

Development of the YouTube Video Transcript Summarizer: This paper introduces a novel solution, the "YouTube Video Transcript Summarizer," designed to streamline the process of extracting valuable insights from YouTube video transcripts. This innovative system leverages advanced technologies, including Natural Language Processing (NLP) techniques, Speech-to-Text capabilities, and Large Language Models (LLM), making it a comprehensive tool for multimedia content comprehension. It also compares the various LLMs available in generating text summaries to further widen the scope for research and development in the field of LLMs.

8. Reference:

1. Taş, O., & Kiyani, F. (2017). A survey automatic text summarization. *Journal of Business, Economics and Finance*, *5*(1), 205–213. https://doi.org/10.17261/pressacademia.2017.591
2. Allahyari, M. (2017, July 7). *Text Summarization Techniques: A Brief survey*. arXiv.org. https://arxiv.org/abs/1707.02268
3. Liu, Y. (2019, August 22). *Text Summarization with Pretrained Encoders*. arXiv.org. https://arxiv.org/abs/1908.08345
4. *Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription*. (2013, December 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/6707758
5. Taskiran, C. M., Amir, A., Ponceleón, D., & Delp, E. J. (2001). . *Proceedings of SPIE*. https://doi.org/10.1117/12.451107
6. Sravani, P. V. K. . M. C. K., . C. N. ,. P. a. ,. K. (2022, July 31). *Youtube Transcript Summarizer Using Flask And Nlp*. https://journalppw.com/index.php/jpsp/article/view/9886
7. Rodrigues, J. P., & Paraíso, E. C. (2020, October 20). *From audio to information: Learning topics from audio transcripts*.https://doi.org/10.5753/kdmile.2020.11967
8. Bokhove, C., & Downey, C. (2018, May 1). *Automated generation of 'good enough' transcripts as a first step to transcription of audio-recorded data*. Methodological Innovations; SAGE Publishing. https://doi.org/10.1177/2059799118790743

9. Albeer, R. A., Al-Shahad, H. F., Aleqabie, H. J., & Al-Shakarchy, N. D. (2022, June 1). *Automatic summarization of YouTube video transcription text using term frequency-inverse document frequency*. Indonesian Journal of Electrical Engineering and Computer Science; Institute of Advanced Engineering and Science (IAES). https://doi.org/10.11591/ijeecs.v26.i3.pp1512-1519
10. Bandabe, S., Zambre, J., Gosavi, P., Gupta, R., & Gaikwad, P. J. A. (2023, April 30). Youtube Transcript Summarizer Using Flask. *International Journal for Research in Applied Science and Engineering Technology*, *11*(4), 98–104.
11. *Benchmarking OpenAI Whisper for non-English ASR - Deepgram Blog ⚡ | Deepgram*. (n.d.). Deepgram. https://deepgram.com/learn/benchmarking-openai-whisper-for-non-english-asr
12. *A Complete Guide to Audio Datasets*. (n.d.). A Complete Guide to Audio Datasets. https://huggingface.co/blog/audio-datasets.