Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

Summarify Pro-YouTube and Audio Transcript Summarizer*

1st Rajveer Singh
School of Computer Applications
Galgotias University)
Greater Noida, India
rajveersingh604668@gmail.com

2nd Anas Khan
School of Computer Applications
Galgotias University)
Greater Noida, India
ak0449369@gmail.com

3rd Md Rahmatullah
School of Computer Applications
Galgotias University)
Greater Noida, India
rajeevashish88@gmail.com

4th Dr. Suneet Gupta
School of Computer Applications
Galgotias University)
Greater Noida, India
email address or ORCID

***Abstract*—The digital information world has been given a paradigm shift in storing texts to storing the video content, and therefore a very large bottleneck in retrieval has been formed. As the number of hours of content uploaded to YouTube per minute exceeds 500, the ability of human consumption is swamped by production and thus requires a computer-based solution of extracting abstractive summaries of YouTube videos and audio transcripts which is proposed in this report, namely, Summerify Pro.[1] Summerify Pro, by combining the capabilities of the YouTube Data API v3, Whisper Automatic Speech Recognition (ASR) model with Open AI and state-of-the-art Large Language Models (LLMs) like GPT-4o and Gemini 1.5 Pro, overcomes the weaknesses of the traditional extractive summarization.[3] This discussion covers the modular architecture of the system, including data ingestion, acoustic modeling, semantic synthesis, and context window management. Additionally, it offers a strict comparative analysis of model performance on the basis of ROUGE, BERTScore, and Video-MME benchmark as well as critical analysis of ethical concerns on the topics of data privacy and algorithmic hallucination. The results indicate that despite the ongoing difficulties in optimization of costs and long-context processing, the combination of hierarchical merging approaches and multimodal LLM can be viewed as the game changer in knowledge management in the post-text world.**

***Index Terms*—YouTube, audio transcript, summarizer, GPT, large language models (LLMs), transcript extraction, summarization, productivity, accurate summary, timesaving, key information.**

I. Introduction

The dynamic explosion of the digital video content has radically changed the processes of information consumption. The YouTube channel, along with other sites like it, has become the main libraries on the planet, providing education content, technical documentation, news programs, and discussions of professionals in the world[1]. This however has brought in a linearity constraint. Video content does not allow time to be wasted as the textual information does by allowing non-linear scanning and quick search of keywords. A two-hour talk about quantum mechanics or a forty-minute technical brief has got a lot of information in it, but no mechanism exists to extract particular insights that has been buried in an unstructured video form, and the insight is not immediately retrieved. Studies have shown that cognitive load of the irrelevant content filtering of long video content and stream of information reduces the efficiency of learning material and information memorization[6]. The situation is shifting towards increasingly greater dependence on manual seeking (scrubbing) or inspection of user-generated timestamps, which tend to be imprecise and unavailable. Accordingly, the Automatic Video Summarization (AVS) requirement is no longer a convenient option but a mandatory productivity feature. Summerify Pro has been specifically designed to meet this requirement, as a way of transforming the time-based commitment of watching video into the instantaneous availability of systematic text.

Automatic Text Summarization (ATS) is a discipline that has been radically altered and passed through different eras of its evolution that shape the development of Summerify Pro. In the past, there were two major paradigms of approaches: extractive and abstractive summarization. Extractive summarization, which is common in the pre-2018 period, entailed the recognition and joining of the most salient sentences in a source text. Initial methods were based on statistical heuristic methods like term frequency-inverse document frequency (TF-IDF), sentence position and graph based ranking algorithms like TextRank,[2] though efficient computationally and free of factual errors, extractive summaries tend to be discontinuous in their narrative, unable to resolve anaphora (pronoun references) and incapable of synthesizing information spread across fragmented portions of a transcript. As an example, the premises made at the start of a video are often subject to conclusions made at the end; extractive methods do not always close the gap. Abstractive summarization with high fidelity has become possible with the introduction of the Transformer architecture and the Large Language Models (LLMs).

Identify applicable funding agency here. If none, delete this.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

Abstractive methods are similar to human thinking because they interpret the original text and produce new sentences with the essential meaning expressed, sometimes with words that are absent in the original input [2]. Before Transformers, Recurrent Neural Network (RNN) based Sequence-to-Sequence (Seq2Seq) models had problems with long range dependencies. However, the today versions of LLMs have the semantic processing ability to make out context, solve ambiguities and create fluent and coherent summaries. Summerify Pro uses this generative power to convert raw and disfluent spoken transcripts into structured and insightful knowledge artifacts[6].

Summerify Pro is theorized as an integrated system of the most advanced Natural Language Processing (NLP) and audio engineering. The system will facilitate a smooth ”Input URL = Output Insight” process to manage the challenges of data retrieval, speech recognition and semantic processing. The main goals of the system are fourfold:

- Robust Ingestion: In order to have a fault-tolerant system of retrieving transcripts using YouTube Data API or scraping generated captions, they need to be accessible even when a standard metadata is not available.[9]
- Audio Fallback: To give the videos without captions solid processing with the use of the OpenAI Whisper model to transcription serving the server-side, thus breaking the dependence on the quality of the in-house ASR in YouTube.[12]
- Semantic Synthesis: To use advanced LLM (particularly, GPT-4o and Gemini 1.5 Pro) to generate tiered summaries (between executive briefs and detailed educational notes) and reduce the problem of hallucinations with the help of citation protocols.[9]

II. Literature Review

A. Neural Abstractive Summarization

Summerify Pro is based upon a theoretical framework based on Neural Abstractive Summarization. This approach unlike the traditional statistical methods views summarization as a conditional language generation problem.The model generates a target summary of a source document by maximizing the probability. The attention mechanisms were introduced, which enabled models to pay attention to other relevant sections of the input sequence on a dynamic basis. Nonetheless, video transcripts have special complications in comparison to regular text files: the text is usually long (longer than standard token limits), disfluent and not punctuated[2]. The latest methods published in 2024 and 2025 have proposed a new method in long-document summarization which is known as Context-Aware Hierarchical Merging.[14] In this method, a long text is divided into chunks, summarized one by one, and the summaries are further combined. Recursive merging may enhance hallucinations, although effectively. Enriching this processing with contextual information in the source document, i.e., summaries are replaced with the input context in which the summaries are generated, or evidence is added to support the summaries, has been shown to help a lot in averting factual inaccuracies (Ou and Lapata 2025). These ideas are implemented in Summerify Pro.

B. Multimodal Large Language Models (MLLMs)

Although summarizing via text of transcripts is strong, it does not consider the visual aspect of video. Multimodal Large Language Model (MLLM) is a trend in video understanding[17]. Some models such as Gemini 1.5 Pro and GPT-4o are trained on huge amounts of interleaved text, image, and video data. MLLM theoretical models are based on matching visual features (extracted through Vision Transformers) to textual representations in a common latent space. This enables the model to reason across visual cues, like when a speaker points to a graph that are not found in the audio transcript, thus, [19]. Video-MME benchmark (2025) has introduced new criteria of measuring these models, in terms of their capability to process various types and lengths of video.Video-MME benchmark (2025) has also set new standards on how to gauge these models in terms of the capability to process different types and lengths of video.[18]

C. Automatic Speech Recognition (ASR) Architectures

The accuracy of a video summarizer depends on the accuracy of the text that is fed to it. Summerify Pro is based on the acoustic foundation of openAI Whisper model. Whisper is a Transformer-based encoder-decoder model that was trained on 680,000 hours of multilingual, multitask supervised data.[21] Unlike traditional ASR systems that require task-specific fine-tuning, Whisper has been trained on raw audio in 30-second segments, encoded into spectrograms in log-Mel form, and then fed through the encoder. The decoder does not only predict the text, it also predicts special tokens upon language identification, time stamp prediction and voice activity.[21]

III. System Architecture

Summerify Pro architecture is based on a microservices model that is scalable, maintains and fault tolerant. There are four major layers of the system, which include the Ingestion Layer, the Processing Layer, the Semantic Layer, and the Presentation Layer.

A. The Ingestion Layer: YouTube Data API and Transcript Handling

The Ingestion Layer deals with external interaction in terms of data sources. The major entry point is YouTube Data API v3. The system uses a You TubeHandler module, which encapsulates the google-api-python-client to get authentication and quota monitoring[1]. Metadata extraction starts the working process. Upon a user entering a URL, the system will extract the video ID and contact the API to a snippet of the video in the form of title, channel, description, and contentDetails (duration). This metadata is essential in formulating the next processing strategy; an example of such is the duration of the video that will decide whether the system will use a single-shot summarization strategy or a map-reduce strategy[10].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

The retrieval of transcripts is a major problem because of the limitations of the platform. The YouTube Data API does not have an easy mechanism to download the transcript of all videos. Summerify Pro would incorporate the youtube-transcript-api library to meet this need. The following priority is the order taken by this module to get transcripts:

- Manually Created Captions: These have been prioritized as they are very accurate and well punctuated.
- Auto-Generated Captions: Used as a fallback. Although it is rather comprehensive, these captions are usually not punctuated and have recognition mistakes.[9]
- Audio Fallback Protocol: In case the API gives out an error of TranscriptsDisabled or the video does not contain any captions whatsoever, the system activates the audio extraction pipeline.[9]

B. The Processing Layer: Audio Extraction and ASR

In case the transcript is not available using the YouTube API, the system triggers the AudioProcessor module. This element uses yt-dlp, a media downloader which is a command line utility, to strip the audio track of the video stream. The raw audio is then fed to ffmpeg to convert to 16kHz mono WAV or MP3 format which is the most suitable input format to the Whisper model.[3] Summerify Pro is based on the OpenAI Whisper architecture in order to transcription. The system chooses either the whisper-turbo model (when the user requires a higher level of speed) or the whisper-large-v3 model (when the user requires the utmost accuracy).[12] The implementation takes care of chunking audio files so as to comply with the 25MB file size restriction of the OpenAI API or does the batch processing when using Azure AI Speech services[23]. More importantly, the word-timestamps parameter is used in the system when transcription is involved. This aspect matches all words of the text generated to their corresponding start and end times of the audio to generate the essential data format of deep linking citations in the end summary[21].

C. The Semantic Layer: LLM Integration and Context Management

The Semantic Layer is the very heart of intelligence of Summerify Pro which converts the raw text into structured knowledge. The system is model-agnostic and is implemented with LangChain to coordinate the communication with Large Language Models.[24] It is optimized to use GPT-4o and Gemini 1.5 Pro.

1) Context Window Management Strategies: The main technical challenge of video summarization is the size of its context window, the maximum size of the text that an LLM can read at once. In two hours there can be more than 30,000 tokens of a lecture. Summerify pro uses dynamic strategy choice:

- The ”Stuff” Strategy: In the case of short videos (up to 15 minutes), the full transcript will be contained within one prompt. This gives the model international consistency since it is able to look at the full picture at a given time.
- The Map-Reduce Strategy: With longer videos, the transcript is divided into blocks of roughly 4,000 tokens, with the 200-token gap to maintain the semantic continuity across the boundaries to the best of its ability.[3]. In the Map phase, the transcript blocks are summarized separately. These segment overviews are then combined together and sent to the LLM in the final synthesis, the Reduce phase.[3]
- Hierarchical Merging with Contextual Augmentation: In order to reduce the risk of detail loss in the Map-Reduce algorithm, Summerify Pro uses a hierarchical strategy in which intermediate summaries are reinforced with citations to the underlying text until the final merge. This is in line with the recent discovery of the reduction of hallucinations in long-document summarization.[14]

2) Multimodal Processing: In case the user needs visual analysis (say, the summarization of a tutorial about coding or a presentation with many slides), the system turns to a Multimodal Pipeline with Gemini 1.5 Pro. Gemini can digest the video file (or a sequence of sampled frames) together with the audio track due to its huge context window (as many as 2 million tokens), where such hints as The speaker pointed to the upper-right quadrant of the scatter plot would be inaccessible to a text analysis, alone.

D. The Presentation Layer: User Interface and Experience

The user is to be provided with the user interface which allows consuming the information quickly. The prototype is created with the help of Streamlit, which is a Python-based framework which enables quick iterations of data-centric applications[9]. The UI features:

- Video Preview: An inbuilt YouTube video that autoplay at particular timestamps whenever a user clicks on a citation in the summary.[9]
- Configuration Panel: It allows users to choose the output format (e.g. Executive Summary, Detailed Notes, Mind Map), as well as a target language.
- Interactive Chat: RAG-based Chat with Video feature, during which the user is allowed to pose certain questions. This is done by inserting chunks of transcripts in a VECTOR database (i.e. Pinecone) and finding related segments when a user makes a query.[11]

IV. Model Selection and Comparative Analysis

The actual performance of Summerify Pro is greatly reliant on the underlying foundational models. The comparison between the two main competitors, OpenAI GPT-4o and Google Gemini 1.5 Pro shows that the two companies have different trade-offs in regards to cost, performance, and capabilities.

A. Cost and Efficiency Analysis

- Scalability depends on a very important factor like cost. However, by 2025 the pricing situation will bear a considerable difference among the providers.
- GPT-4o: costs around 2.50 million dollars per million input tokens and 10.00 million doller output tokens. It is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

characterized by great reasoning power and with a higher operational cost when operating over great contexts.[25]
- Gemini 1.5 Pro: Google has priced this model at or below cost-plus: about 1.25 / million input tokens and 5.00 / million output tokens, although not as efficient as GPT-4o.[25]]because the architecture is better at long-context retrieval.[27]

B. Performance and Context Capabilities

The context window is the most important distinguishing variable.

- The GPT-4o has a 128k token context window. Although it is enough in case of most papers, and even most medium-length videos, it demands the Map-Reduce architecture in case of very long content which creates complexity and leads to information loss at the boundaries of the chunks.[25]
- Gemini 1.5 Pro has a context window of up to 2 million tokens.[20] this generational advancement gives Summerify Pro the capacity to load an entire 10-hour audio book, or a lengthy conference proceeding, into the model in a single prompt. This removes the necessity of chunking and enables the model to do needle-in-a-haystack retrieval across the entire dataset with more than 99

C. Evaluation Benchmarks: Video-MME

In order to have an objective estimate of these models, we use the Video-MME (Multi-Mode Evaluation) benchmark, launched at CVPR 2025. It is the first large-scale test of MLLMs on video analysis, encompassing a variety of areas and time ranges, 11 seconds to 1 hour long.[18]. Based on the benchmark findings:

- Gemini 1.5 Pro was found to have an average accuracy of 75.0 percent and is dramatically better than open-source models and better than GPT-4o, with 71.9 percent.[18]
- The review shows that the outstanding performance of Gemini is especially strong when it comes to working with long-term temporal dynamics and the combination of multiple modalities (subtitles + audio + frames).[18]

According to this data, Summerify Pro will have a hybrid routing logic: simple, short text queries are run on GPT-4o (or its less expensive version GPT-4o-mini) to use its fast reasoning, whereas complex, long-form, or visually-intensive tasks are run on Gemini 1.5 Pro.

TABLE I
COMPARATIVE ANALYSIS OF FOUNDATION MODELS FOR VIDEO SUMMARIZATION

Feature	GPT-4o	Gemini 1.5 Pro	OpenAI Whisper (Large-v3)
Primary Function	Text/Multimodal Generation	Native Multimodal (Video/Audio)	Automatic Speech Recognition
Context Window	128,000 Tokens	2,000,000 Tokens	N/A (Streaming/File based)
Input Cost (per 1M)	~$2.50	~$1.25	~$0.006 per minute (API)
Video Handling	Sampled Frames + Text	Native Video Tokenization	Audio Only
Benchmark (Video-MME)	71.9% Accuracy	75.0% Accuracy	N/A
Optimal Use Case	Complex reasoning, Structured text	Long-form video (>1hr), Visual analysis	High-fidelity transcription

V. EVALUATION AND METRICS

Assessment of abstractive summarization is also famous to be hard due to the subjectivity of quality. Summerify Pro uses a multi dimensional assessment system which integrates conventional statistical measures with new LLM-generated judging systems.

A. Statistical Metrics: ROUGE and BERTScore

In the past, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure has been the one used. It is the overlap between N-grams (patterns of N words) of the generated summary and a human generated gold standard reference.[30]

- ROUGE-1 is denoted by the coincidence of the unigrams (words).
- ROUGE-L uses the longest common wording, which entails sentence structure.

Although ROUGE can be applied in regression testing, it has a shortcoming since it penalizes legitimate synonyms. The summary that the film was great will score low ROUGE on a reference that the film is outstanding even though it means the same thing.[31]

Summerify Pro uses BERTScore to deal with this. BERTScore is used instead of the exact string matching: the cosine similarity of the contextual meaning of the words in the candidate and reference summaries is computed as a result of the embedding of the words in the context of this computation. This enables the system to identify that excellent and outstanding are semantically equivalent giving it a significant better correlation with human judgment.[31]

B. LLM-as-a-Judge: The G-Eval Frameworks

Since reference-based metrics are too costly (summaries need to be tediously written by humans), the industry has shifted to reference-free evaluation with the help of LLM-as-a-Judge, i.e., G-Eval framework wherein a powerful LLM (such as GPT-4) would receive the source transcript and the summary generated and then a specific rubric. The model produces a score (1-5) according to the set criteria:

- Coherence
- Consistency
- Fluency
- Relevance

C. Quantitative Results

A ROUGE-L score of 0.45 was obtained when the generated workflow of the ”Map-Reduce” was internally tested by using GPT-4 on a dataset of lecture videos, which is much more pertinent than the 0.29 score of extractive baselines.[3] It proves that the generative model is a better narrative flow and structural coherence when compared to extractive counterparts. But the drawback is latency; 30-60 seconds to run a long video over Map-Reduce is not quick at all, and extractive processes are almost instant.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

VI. Methodologies

Summify Pro goes beyond basic TL;DR generation to provide the reader with a sophisticated level of insight. It uses definite methods of managing the organization and concentration of information.

A. Hierarchical Merging with Refinement

In extremely long documents a concatenation of chunk summaries can result in the final product being disjointed especially when it is done in a simple manner. A Hierarchical Merging strategy is adopted in Summerify Pro.[14]

- Level 1: The transcript is divided into 10-minute instructions.
- Level 2: Summaries of such segments are created.
- Level 3: These again are summarized and summed up.

More importantly, the system employs Refinement step. The model does not merely sum up the Level 2 summaries, but rather encourages the user to include details of Segment X in the global summary. This recursion process will guarantee that particular data points (such as dates, names, or figures) will rise to the upper level to avoid the blandness commonly experienced in recursive summarization.[14]

B. Prompt Engineering and Hallucination Mitigation

The hallucination or the creation of believable and false information is a major danger. Chain-of-Thought (CoT) prompting is used to alleviate this in summerify Pro. [15]The prompt gives the following instructions to the model: First, you should list the main facts that you have discovered in the text. Second, create a summary with the help of those facts only. Third, include time of all facts quoted. This Citation Mode requires the model to base its generation on particular parts of the transcript. In case the model fails to discover a timestamp of a claim, the model will be told to eliminate that claim. The method has been found to give high values on the factual consistency score in G-Eval assessments.[15].

C. Competitor Analysis

- Eightify: A Chrome add-in that is speed-oriented. It offers an 8 point summary of videos. Its strong aspect lies in being closely incorporated into the YouTube player, yet the weakness is in the lengthiness of its summaries; it is a bit lacking in detail in complicated technical tutorials.[34]
- Glasp: It differentiates itself by means of social highlighting. It is more of a knowledge management tool than a mere summarizer so that users can highlight pieces of transcripts and share it. It attracts the community using the Second Brain (Obsidian/Notion users).[35]
- Recall.ai: Located as a high-end knowledge base tool. It abstracts videos and classifies them as a knowledge graph. It provides quality briefs at a greater cost level.[—human—¿It provides quality briefs, however, at a premium price point.[—human—¿It provides quality briefs but at a premium price point.[35]
- NoteGPT: Catering to the educational niche, it offers such features as screenshot capture and organized notes to students.[35]

D. Summerify Pro's Differentiation

The features of Summerify Pro include Deep Contextualization and Audio-First design.

- Hybrid Multimodality: Summerify Pro, having lost Eightify because of its text-only product, uses Gemini 1.5 Pro to view the video, which provides a visual context that the other competitors fail to capture.[29]
- Robust Audio Handling: A lot of competitors are unable to survive when disabling YouTube captions. The implementation of the Whisper pipeline in Summerify Pro guarantees that it can handle any video or audio file, thus it can also be used to summarize podcasts.[3]
- Hierarchical Depth: The depth of the summary can be selected by the user. A student may ask to get study notes (extracting definition and concepts), whereas the executive may ask to receive a decision brief (concentrating on results and action steps).

TABLE II
Feature Comparison of Leading Video Summarizers

Feature	Summerify Pro	Eightify	Glasp	Recall.ai
Core Model	GPT-4o / Gemini 1.5	GPT-3.5 / GPT-4	GPT-3.5	GPT-4
Video Vision	Yes (Gemini)	No (Text Only)	No	Limited
Audio Fallback	Yes (Whisper)	No	No	Yes
Pricing Model	Freemium / Usage	Subscription	Free	Subscription
Key Differentiator	Hierarchical Merging	Speed / Integration	Social Sharing	Knowledge Graph

VII. Security and Privacy

Automation of content processing attracts certain ethical and legal issues which Summerify Pro has to answer to conduct responsible deployment.

A. Data Privacy and PII Redaction

In the case where users summarize their own videos (e.g. company-only meetings or user interviews), they may risk sharing Personally Identifiable Information (PII) with third-party model providers. Summerify Pro should be built to alleviate this with the addition of Amazon Bedrock Guardrails or other PII detection libraries (such as Microsoft Presidio) into the preprocessing pipeline.[12]. A transcript must first go through a scrubber that recognizes and masks any transcript content like names and telephone numbers, email addresses and credit card numbers and replaces them with tokens (e.g. ¡PHONE NUMBER¿).[12] This will make it compliant with GDPR and HIPAA policies concerning data processing.

B. Copyright and Fair Use

The legal question of whether YouTube transcripts should be used to process AI is complicated. Generally, the Terms of Use of YouTube do not allow scraping of information to construct competing sets of data. Nevertheless, Summerify Pro is a User Agent, i.e., it is a single video that is run on express command of a user to serve his or her own purposes. This

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

type of usage is typically argued to be subject to Fair Use, namely, in the category of a transformative work that adds value (summarization) but does not substitute the market of the original work (the video experience).[42] However, the system has a hard and fast policy of no-storage of raw transcripts to curtail copyright liability; transcripts are temporary and disappear when the processing session is complete.

C. Algorithmic Bias and Hallucination Risks

LLM can replicate the bias of their training data or do hallucinations that do not accurately reflect the intent of the speaker. This especially is risky in healthcare or monetary scenarios.[39].

- Risk: A summary may recommend a stock trade which the speaker had in fact recommended against.
- Mitigation: The interface has conspicuous disclaimers indicating that the content is generated by an AI. Besides, the ”Citation Mode” enables users to confirm any assertion.[44] When a user clicks on the timestamp accompanying a summary point, he or she is redirected to the corresponding audio source enabling instant human confirmation.

D. Cloud Security Challenges

The risks associated with the operation of a cloud-based AI service are data breach, insider threats, and API key leaks[40]. Summerify Pro uses a Shared Responsibility Model of security. Securing the infrastructure is provided by industry-standard encryption (TLS 1.3) of data in transit and AES-256 of data at rest (user summaries). YouTube and OpenAI API keys are stored as secret management (such as AWS Secrets Manager) instead of being hardcoded, avoiding unauthorised access[45].

VIII. Future Directions

A. The Road to Multimodal RAG

The short-term future of Summerify Pro is Multimodal Retrieval Augmented Generation (RAG).[13]. Where the existing systems depend so much on text transcripts, the following generation will index the visual vectors of the video frames. This will allow users to search by visual concepts e.g. Find the part of the video where the chart has a dip in Q3 sales even though the speaker did not refer to the chart. Such models as CLIP and SigLIP will be instrumental aspects in alignment strategies of this feature.

B. Real-Time and Agentic Workflows

Another way it will develop in the future is with Agentic Workflows.[48]. Instead of summarizing, the system will be in a position to act. As an illustration, an integrated agent could automatically fill a shopping list application after summarizing a video about the cooking process. Or, it may create a Pull Request containing a code in a codet tutorial. This takes the system out of the passive mode of a Summarizer and puts it into the active mode of an Assistant.

IX. Conclusion

Summerify Pro is a convergence of developed ASR technology and the emerging Generative AI potential. It provides invaluable value in education, career advancement, and content control by overcoming the major bottleneck of video usage time. The architectural choice of the hybrid system, i.e. lightweight transcript extraction when applicable and other robust approaches, i.e. ASR and map-reduce, when there is a complex content, offers the best tradeoff between performance and reliability. The barrier to the so-called total video recall will be eliminated as the size of the context windows will diminish and the price will drop, which will essentially transform the relationship with digital media, as it will be no longer about consumption but about interaction. The study confirms that issue of hallucination and cost are yet to be resolved, but the future of the technology is such that AI summarizing will become a common utility to all video materials.[1]

References

[1] AI-driven Video Summarization, ResearchGate. https : / / www . researchgate . net / publication / 381456155 AI - driven Video _ Summarization
[2] A Comprehensive Survey on Automatic Text Summarization with Exploration of LLM-Based Methods, arXiv. https://arxiv.org/html/2403.02901v2
[3] Building a Custom Video Summarization Workflow with Whisper and GPT Models, ResearchGate. https://www.researchgate.net/publication/400096227_Building_a Custom_Video_Summarization_ Workflow_ with_Whisper_and_GPT_Models
[4] AI-Powered Audio Summarization and Ethical Content Analysis Using OpenAI Whisper, SSRN. https : / /papers.ssrn.com/sol3/papers.cfm?abstract_id=5219188
[5] A Comparative Study of Quality Evaluation Methods for Text Summarization, arXiv. https://arxiv.org/html/2407.00747v1
[6] Lee et al., “Video Summarization with Large Language Models,” CVPR 2025. https : / / openaccess . thecvf . com / content / CVPR2025/papers/Lee _Video _Summarization _with _Large _Language _ Models _CVPR _2025 _paper.pdf
[7] The Impact of Artificial Intelligence on Students’ Academic Development, MDPI. https://www.mdpi.com/2227-7102/15/3/343
[8] YouTube Transcript Summarization Using Abstractive and Extractive Approaches, ResearchGate. https://www.researchgate.net/publication/384645085_YouTube_ Transcript_ Summarization_ Using_ Abstractive_ and_Extractive_Approaches
[9] Abstractive Summarization of YouTube Videos Using LaMini-Flan-T5 LLM, ResearchGate. https : / / www. researchgate . net / publication / 384644200_Abstractive_Summarization_ of_ YouTube_ Videos_ Using_ LaMini-Flan-T5 _LLM
[10] David Yoo, “Building a Simple YouTube AI Video Summarizer,” Medium. https://medium.com/@david_ yoo/building-a-simple-youtube-ai-video-summarizer-07d372bfcb6a
[11] YouTubeGPT: A Deep Dive into Building and Running It, Microsoft TechCommunity. https://techcommunity.microsoft.com/blog/educatordeveloperblog / youtubegpt - a - deep - dive - into - building - and - running-it/4270867
[12] Build a Serverless Audio Summarization Solution with Amazon Bedrock and Whisper, AWS Machine Learning Blog. [Online]. Available: https: //aws.amazon.com/blogs/machine- learning/build- a- serverless- audio-summarization-solution-with-amazon-bedrock-and-whisper/
[13] Hierarchical Document Refinement for Long-context Retrieval-augmented Generation, arXiv. [Online]. Available: https://arxiv.org/html/2505.10413v1
[14] Context-Aware Hierarchical Merging for Long Document Summarization, arXiv. [Online]. Available: https://arxiv.org/abs/2502.00977
[15] Context-Aware Hierarchical Merging for Long Document Summarization, ACL Findings 2025. [Online]. Available: https://aclanthology.org/2025.findings-acl.289/

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

[16] Context-Aware Hierarchical Merging for Long Document Summarization (PDF), ACL Anthology. [Online]. Available: https://aclanthology.org/2025.findings-acl.289.pdf
[17] Bridging Multimodal and Video Summarization: A Unified Survey, ACL Anthology. [Online]. Available: https://aclanthology.org/2025.newsum-main.11.pdf
[18] Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis, CVPR 2025. [Online]. Available: https://cvpr.thecvf.com/virtual/2025/poster/33002
[19] Video-MME Benchmark Paper, CVF Open Access. [Online]. Available: https : / / openaccess . thecvf . com / content / CVPR2025 / html / Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025 paper.html
[20] Gemini 1.5: Unlocking Multimodal Understanding, Google DeepMind Technical Report. [Online]. Available: https://storage.googleapis.com/deepmind-media/gemini/gemini_v1 5 report.pdf
[21] Using OpenAI Whisper and Mixtral8x7B to Transcribe and Correct Grammar from Videos, Medium. [Online]. Available: https : / / brandolosaria.medium.com/using- openai- whisper- and- mixtral8x7b-to-transcribe-and-correct-grammar-from-videos-da1d243fc157
[22] Speech-to-Text API Documentation, OpenAI Developers. [Online]. Available: https://developers.openai.com/api/docs/guides/speech-to-text/
[23] Whisper Model Overview, Microsoft Learn. [Online]. Available: https://learn.microsoft.com/en-us/azure/ai-services/speech-service/whisper-overview
[24] YouTube Transcript Summarizer, Semantic Scholar. [Online]. Available: https : / / www . semanticscholar . org / paper / YouTube - Transcript - Summarizer - Kumar - Vashistha / 7f7f4887b146b6c84cbc4bf8f6e18117f2c4ad42
[25] GPT-4o vs Gemini 1.5 Pro Performance Comparison. [Online]. Available: https://docsbot.ai/models/compare/gpt-4o/gemini-1-5-pro-002
[26] Gemini 1.5 Pro vs ChatGPT-4o Full Guide, EaseMate AI. [Online]. Available: https://www.easemate.ai/ai-pdf-solutions/gemini-1.5-pro-vs-chatgpt-4o.html
[27] Comparing API Costs of Grok AI, Gemini API, and OpenAI, Rogue Marketing. [Online]. Available: https://the-rogue-marketing.github.io/generative-ai-gemini-vs-other-api-providers-cost-comparison/
[28] Fu et al., "Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis," CVPR 2025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2025/papers/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025 paper.pdf
[29] How MXT-1.5 Performs Against Leading AI Models, Moments Lab Blog. [Online]. Available: https : / /www.momentslab.com/blog/mxt-outperforms-leading-ai-models
[30] RAG Evaluation Metrics: A Journey Through Metrics, Elasticsearch Labs. [Online]. Available: https : / /www.elastic.co/search- labs/blog/evaluating-rag-metrics
[31] LLM Evaluation for Text Summarization, Neptune.ai. [Online]. Available: https://neptune.ai/blog/llm-evaluation-text-summarization
[32] FACTS Grounding: A Benchmark for Evaluating the Factuality of Large Language Models, Google DeepMind Blog. [Online]. Available: https://deepmind.google/blog/facts-grounding-a-new-benchmark-for-evaluating-the-factuality-of-large-language-models/
[33] EdinburghNLP, Awesome Hallucination Detection Repository, GitHub. [Online]. Available: https : / / github . com / EdinburghNLP / awesome - hallucination-detection
[34] 10 Best YouTube Video Summarizer AI Tools in 2025, Side Space. [Online]. Available: https://www.sidespace.app/blog/10-best-youtube-video-summarizer
[35] 10 Best YouTube Video Summarizers in 2024: Features and Pricing, Recall.ai Blog. [Online]. Available: https://www.getrecall.ai/post/10-best-youtube-video-summarizers-in-2024
[36] Top YouTube Video Summarizer Tools Discussion, Reddit ProductivityApps. [Online]. Available: https : / / www . reddit . com / r / ProductivityApps / comments / 1bk4q17 / i tested 40 youtube video _ summarizers_that all/
[37] Top 10 YouTube AI Video Summary Tools in 2025, BibiGPT Blog. [Online]. Available: https://bibigpt.co/blog/posts/top- 10- youtube- ai-video-summary-tools-2025-en
[38] Best AI Summary Apps of 2025, AI Tutor Blog. [Online]. Available: https://ai-tutor.ai/blog/best-ai-summary-apps/
[39] Ethical Considerations in Automated Transcription Services, Globibo Blog. [Online]. Available: https://globibo.blog/ethical-considerations-in-automated-transcription-services/
[40] The Challenges of Data Privacy and Cybersecurity in Cloud Computing and Artificial Intelligence Applications, PMC. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12743334/
[41] Data Security within AI Environments, Cloud Security Alliance. [Online]. Available: https://cloudsecurityalliance.org/artifacts/data-security-within-ai-environments
[42] The Use of Automatic AI-based Notes and Transcription Services in Qualitative Research: Ethical and Methodological Concerns, ResearchGate. [Online]. Available: https : / / www . researchgate . net / publication / 385065048_The_Use_of_Automatic_AI - based_Notes_and_Transcription_Services_in Qualitative_ Research_ Ethical_ and_ Methodological_Concerns
[43] Practical and Ethical Issues of Automated Transcription, Blog Article. [Online]. Available: https : / / caitlinhafferty . blogspot . com / 2021 / 07 / practical-and-ethical-considerations-for-automated-transcription.html
[44] IEEE Conference Template Document. [Online]. Available: https://sst-reu.fiu.edu/wp-content/uploads/2019/07/IEEE-Template.doc
[45] Cloud Data Security in 2026: Dangers, Safeguards, and More, Coursera. [Online]. Available: https : / / www. coursera . org / articles / cloud - data - security
[46] Educational Impacts of Generative Artificial Intelligence on Learning, PMC. [Online]. Available: https : / / pmc . ncbi . nlm . nih . gov / articles / PMC12280127/
[47] AI and Teacher Productivity: Time-Saving and Workload Reduction in Education, ResearchGate. [Online]. Available: https : / / www.researchgate.net/publication/387980511_AI_and_Teacher_Productivity_A_Quantitative_Analysis_of_Time - Saving_and_ Workload_ Reduction_in_Education
[48] Long-Context Modeling Overview, Emergent Mind. [Online]. Available: https://www.emergentmind.com/topics/long-context-modeling
[49] Learning to Summarize by Learning to Quiz: Adversarial Agentic Collaboration for Long Document Summarization, arXiv. [Online]. Available: https://arxiv.org/html/2509.20900v2