NEW YORK CITY COLLEGE OF TECHNOLOGY CITY TECH | CET4973

Summarization Chatbot for YouTube Videos

Bilal Akhtar[1]
Department of Computer Engineering Technology
New York City College of Technology
bilal.akhtar@mail.citytech.cuny.edu

Abstract

This project focuses on the design and development of an AI-driven, interactive chatbot intended to deliver accurate summarizations of YouTube video content. The chatbot, housed on a Streamlit platform for efficient deployment and increased user accessibility, fetches available video transcripts and generates concise summaries, alongside facilitating user-contextual inquiries. With the incorporation of OpenAI's GPT-3.5-turbo model, the tool ensures accurate summaries and fosters improved user engagement through interactive dialogues. The practical implication of this research manifests as a tool to efficiently manage and engage with the proliferating volume of YouTube content in the digital age.

1 Introduction

Navigating through the immense array of YouTube content is increasingly challenging due to its rapid growth. Valuable insights often get missed due to time constraints and the challenge of decoding complex content. Hence, there's a significant need for a system capable of concisely summarizing the essence of YouTube videos. Though, the development of such a system faces substantial challenges, including the complexity of localized NLP models, the exorbitant cost of cutting-edge AI models, and the need to ensure user-friendly interactivity.

To navigate these challenges, we designed an innovative AI-powered chatbot, which is available for public use at https://ytchatbot.streamlit.app/, deployed via a Streamlit app for easy accessibility. The chatbot fetches pre-existing transcriptions as the input and uses OpenAI's GPT-3.5-turbo based model to generate concise, yet comprehensive summaries as output. Alongside, it incorporates an interactive Q&A module for additional exploration of video content.

The paper delineates our approach to solve the stated problem, highlighting how the Streamlit-hosted chatbot operates, how it overcomes existing obstacles, and the potential benefits it can offer to users amidst the densely populated digital world of YouTube content.

2 Related work

Significant research has been conducted in the field of text summarization, with especially notable work utilizing a blend of Natural Language Processing (NLP) and Python-based web technologies.

a. "YouTube Transcript Summarizer Using Flask" by Bandabe et al. (2023)

Bandabe et al.'s work introduces an innovative tool for automatically summarizing YouTube

[1] Bilal Akhtar, Department of Computer Engineering Technology, New York City College of Technology. Email: bilal.akhtar@mail.citytech.cuny.edu. For more information, visit https://www.linkedin.com/in/iambilalakhtar/. EMPL ID: 23628765.

video transcripts, noteworthy for its versatility in processing both video transcripts and audio into textual summaries. Their broad application scope extends into numerous domains, highlighted by their clever use of accessible technologies including Flask, NLP, Python libraries, and Hugging Face Transformers.

Potential shortcomings, however, surface in the tool's static nature that solely provides summaries without enabling interactive user engagement. This lack of interactivity could limit in-depth user comprehension of content.

In contrast, our proposed AI-powered chatbot offers more than summarization. It provides an interactive platform allowing users a deeper immersion into video content via a Q&A module, thus promoting comprehension and information retention. While our work employs a similar technological framework, it innovates by adding a responsive chatbot, integrating an interactive, dynamic layer to the process.

3 Dataset and Features

The dataset for this project was dynamically created using YouTube's Data API, specifically targeting video transcript data. It involved acquiring transcripts from a vast range of YouTube videos across varied genres for diversity and comprehensiveness.

To acquire this data, we initiated a systematic plan. First, video IDs were collected using the user-defined search queries to cover a broad spectrum of topics. Then, transcripts for each video ID were extracted, constituting the raw dataset. It's crucial to note that this methodology only concerned videos that had accurate and accessible transcript data.

The raw dataset underwent extensive cleaning- timestamps, numerical data, and redundant phrases were removed meticulously. The aim was to preserve the essential textual content that would be interpretable by the NLP model, leading to optimal quality and preciseness of the produced summaries and chatbot interactions.

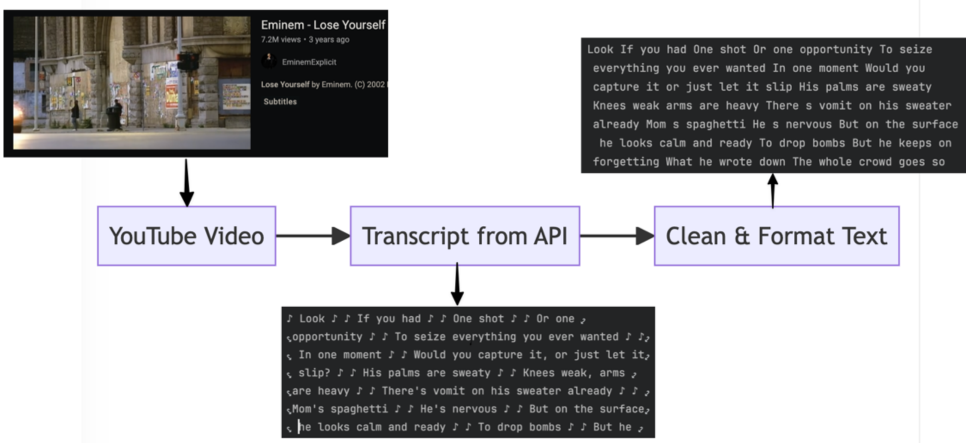
Feature extraction from the cleaned transcripts was performed using techniques like word2vec to transform the text data into a format that the model could utilize effectively.

As for the ethical and privacy aspects, we closely abided by YouTube's API usage rules, making sure all data used was public and non-personalized. We were careful not to infringe on privacy rights; data retrieval respected public access standards, and no data that could identify individual users was gathered or used.

4 Methods

In our system, we employ three critical modules — video content transcription, text summarization, and AI-empowered chatbot interactivity via Streamlit, leveraging the capabilities of OpenAI's GPT-3.5-turbo.

a. Video content transcription: The approach initially requires textual data from YouTube videos which underpins the subsequent steps. The choice of using video transcripts over audio or visual data roots in the convenience and cost-effectiveness of Google's YouTube Data API, which permits us to access already generated transcripts for numerous YouTube videos, thereby refraining from expensive and time-consuming transcription operations.



```
Look If you had One shot Or one opportunity To seize
 everything you ever wanted In one moment Would you
 capture it or just let it slip His palms are sweaty
 Knees weak arms are heavy There s vomit on his sweater
 already Mom s spaghetti He s nervous But on the surface
  he looks calm and ready To drop bombs But he keeps on
 forgetting What he wrote down The whole crowd goes so
```

```
♪ Look ♪ ♪ If you had ♪ ♪ One shot ♪ ♪ Or one
opportunity ♪ ♪ To seize everything you ever wanted ♪ ♪
 In one moment ♪ ♪ Would you capture it, or just let it
 slip? ♪ ♪ His palms are sweaty ♪ ♪ Knees weak, arms
are heavy ♪ ♪ There's vomit on his sweater already ♪ ♪
Mom's spaghetti ♪ ♪ He's nervous ♪ ♪ But on the surface
 he looks calm and ready ♪ ♪ To drop bombs ♪ ♪ But he
```

Fig. 1 Outline for Fetching a YouTube Video Transcript.

b. **Text Summarization:** The acquired transcripts get transformed into brief summaries using several NLP methodologies incorporated by the GPT model. To cope with extensive content, our strategy is to segment the text into chunks, manageable for the GPT model, while maintaining the context's coherence. Each chunk is independently summarized by the GPT model, producing a collective output that stands as the final summary. Thus, the summarization handles text of variable lengths without any context compromise.

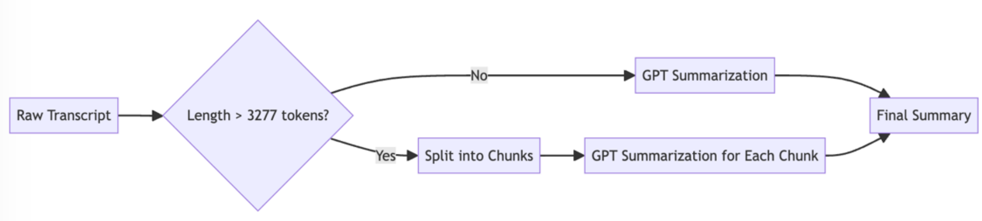


Fig. 2 Outline from Transcript to GPT to Summary.

c. **AI-powered chatbot interactivity:** The solution, in addition to offering text summarization, retains an interactive feature in the form of a context-aware chatbot, hosted on the Streamlit platform for easy deployment and improved user accessibility. This chatbot employs the same underlying ChatGPT model, showcasing the model's seamless transition capability between summarization and dynamic, context-sensitive query processing. Users can directly interact with the chatbot via a user-friendly interface, making inquiries based on the video summary, promoting an engaging interaction.

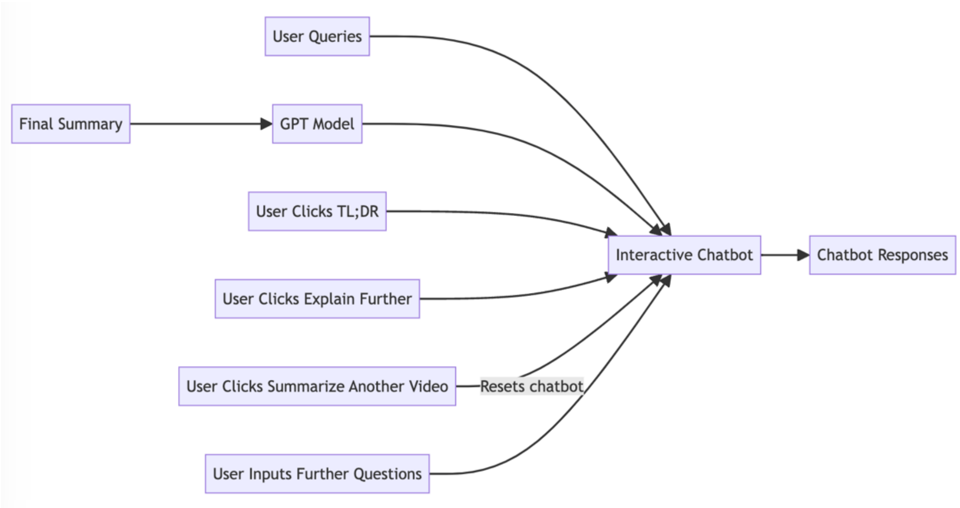
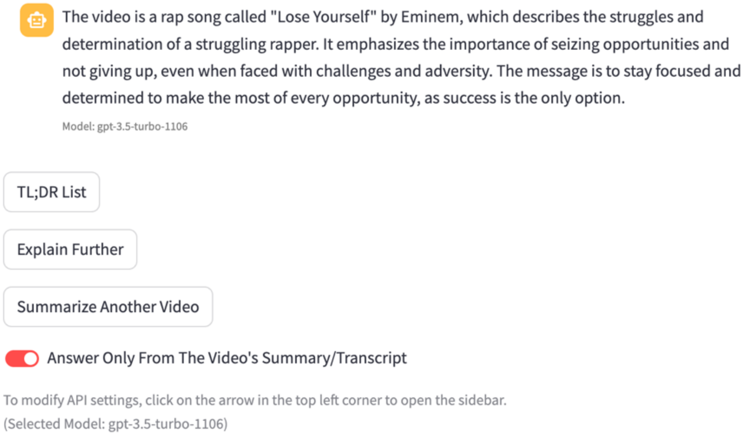


Fig. 3 Outline of the user receiving the final summary, then either using the predefined button or posing their own question.



The video is a rap song called "Lose Yourself" by Eminem, which describes the struggles and determination of a struggling rapper. It emphasizes the importance of seizing opportunities and not giving up, even when faced with challenges and adversity. The message is to stay focused and determined to make the most of every opportunity, as success is the only option.

Model: gpt-3.5-turbo-1106

● Answer Only From The Video's Summary/Transcript

To modify API settings, click on the arrow in the top left corner to open the sidebar.
(Selected Model: gpt-3.5-turbo-1106)

Fig. 4 Screenshot of the Streamlit chatbot displaying the entire procedure.

The efficacy of the method is judged by the relevance and clarity of the produced summaries and the accuracy of the chatbot responses. The cohesive performance of these elements transforms how YouTube content is interacted with and understood by users, thus making it more accessible and intuitive.

5 Experiments/Evaluation

- **Context**
  In evaluating the summarization capabilities of a chatbot using the GPT-4 model, a comprehensive approach was adopted. Three YouTube videos, representing different genres, were selected to test the model's proficiency across various content types:

  a. **Entertainment**: Eminem's "Lose Yourself" - a cornerstone in rap culture, well-known for its deep and motivational lyrics.
  b. **Educational**: "How to build a capsule wardrobe" - a practical guide offering insights into fashion minimalism and wardrobe management.
  c. **Science & Nature**: "How Mosquitoes Use Six Needles to Suck Your Blood | Deep Look" - an in-depth look into the intricate biology of mosquitoes.

These selections aim to challenge the model's ability to handle diverse subjects - from the poetic and abstract lyrics of a rap song to the practical tips of fashion, and the scientific details of nature.

- **Reason for Using GPT-4**
  Given the extensive pre-training and optimized default settings of the GPT-4 model, we didn't find it necessary to alter any of the parameters. The model was utilized with its default settings including the temperature being set to 1, which already provide a good balance between generating diverse responses and ensuring relevance to the prompt. We utilized this model mainly due to two reasons:

  a. **Advancement:** GPT-4 is one of the most advanced language models to date, capable of understanding and generating human-like text with a high degree of accuracy and coherence. It is excellently suited for tasks such as summarization that require rich semantic understanding.
  b. **Subjectivity Handling:** Evaluating text summarization cannot be thoroughly managed with metrics like BERTScore or ROUGE due to the inherent subjectivity involved in the task. What might seem like an important point to one person may not be to another, hence an advanced language model like GPT-4 can better comprehend and reflect this subjectivity in its summaries.

- **Evaluation Metrics**
  The metrics for evaluation chosen were:

  a. **Relevance**: The summaries should capture the essence of each video, omitting non-essential details.
  b. **Coherence**: The summaries should be well-structured, presenting information in a logical and seamless manner.
  c. **Consistency**: The summaries should accurately reflect the content of the videos without misrepresenting any facts.
  d. **Fluency**: The language should be grammatically correct, with appropriate use of vocabulary and syntax.

- **Methodology**
  Each video was processed through the chatbot, and the summaries were evaluated against these metrics. We also compared the chatbot's summaries with human-written summaries for a comprehensive assessment.

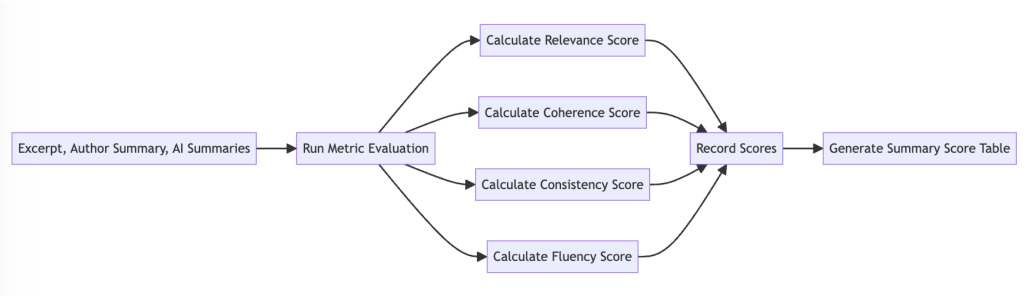
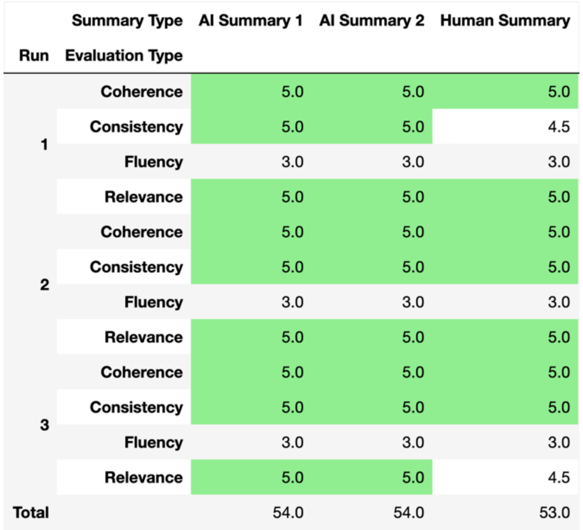

Excerpt, Author Summary, AI Summaries
Run Metric Evaluation
Calculate Relevance Score
Calculate Coherence Score
Calculate Consistency Score
Calculate Fluency Score
Record Scores
Generate Summary Score Table

Fig. 5 Evaluation Model Outline

  a. For "Lose Yourself," a comparison was made with a human summary from a blog titled "Eminem's Lose Yourself Lyrics: A Masterpiece of Rap Culture" by Alex Harris, specifically under the header "The Meaning of Lose Yourself by Eminem."
  b. For the other two videos, human-authored summaries from the author's website were used, providing insights into the chatbot's summarization of educational and scientific content.

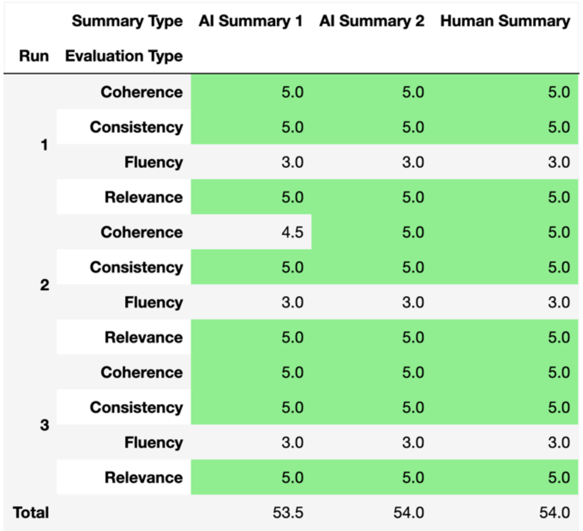
- **Summary Evaluation Findings**
  As seen from the previous figures, the evaluation yielded the following insights:

  a. **Relevance:** All summaries consistently captured the essential elements of their respective videos. Both AI and human summaries achieved high relevance scores, demonstrating the GPT-4 model's ability to discern and condense key information from diverse content.
  b. **Coherence:** Coherence scores were high across the board, with AI-generated summaries matching or exceeding the human-generated summary in structuring information in a logical flow.
  c. **Consistency:** Consistency was generally high, indicating a strong factual alignment between the summaries and the source videos. The AI model effectively preserved the accuracy of the content without introducing discrepancies.
  d. **Fluency:** While all summaries were understandable and followed logical sentence structures, this was the area with the lowest scores, indicating that the fluidity and grace of natural language could be enhanced in AI-generated summaries.



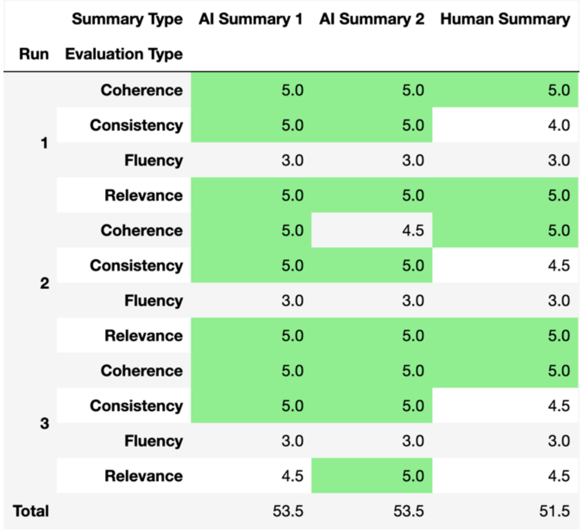
	Summary Type	AI Summary 1	AI Summary 2	Human Summary
Run	Evaluation Type			
1	Coherence	5.0	5.0	5.0
	Consistency	5.0	5.0	4.5
	Fluency	3.0	3.0	3.0
	Relevance	5.0	5.0	5.0
2	Coherence	5.0	5.0	5.0
	Consistency	5.0	5.0	5.0
	Fluency	3.0	3.0	3.0
	Relevance	5.0	5.0	5.0
3	Coherence	5.0	5.0	5.0
	Consistency	5.0	5.0	5.0
	Fluency	3.0	3.0	3.0
	Relevance	5.0	5.0	4.5
Total		54.0	54.0	53.0

Fig. 6 Evaluation output for "Lose Yourself."



	Summary Type	AI Summary 1	AI Summary 2	Human Summary
Run	Evaluation Type			
1	Coherence	5.0	5.0	5.0
	Consistency	5.0	5.0	5.0
	Fluency	3.0	3.0	3.0
	Relevance	5.0	5.0	5.0
2	Coherence	4.5	5.0	5.0
	Consistency	5.0	5.0	5.0
	Fluency	3.0	3.0	3.0
	Relevance	5.0	5.0	5.0
3	Coherence	5.0	5.0	5.0
	Consistency	5.0	5.0	5.0
	Fluency	3.0	3.0	3.0
	Relevance	5.0	5.0	5.0
Total		53.5	54.0	54.0

Fig. 7 Evaluation output for "How Mosquitoes Use Six Needles to Suck Your Blood | Deep Look."



	Summary Type	AI Summary 1	AI Summary 2	Human Summary
Run	Evaluation Type			
1	Coherence	5.0	5.0	5.0
	Consistency	5.0	5.0	4.0
	Fluency	3.0	3.0	3.0
	Relevance	5.0	5.0	5.0
2	Coherence	5.0	4.5	5.0
	Consistency	5.0	5.0	4.5
	Fluency	3.0	3.0	3.0
	Relevance	5.0	5.0	5.0
3	Coherence	5.0	5.0	5.0
	Consistency	5.0	5.0	4.5
	Fluency	3.0	3.0	3.0
	Relevance	4.5	5.0	4.5
Total		53.5	53.5	51.5

Fig. 8 Evaluation output for "How to build a capsule wardrobe."

6 Results/Discussion

The figures provide a visual representation of scores across each metric for each video. The following observations were made:

a. **AI vs. Human Summary Performance:** The AI-generated summaries often matched or even exceeded the human-written summaries in terms of Coherence and Consistency. This demonstrates the AI's strong understanding of the videos' structures and its ability to reflect content accurately.

b. **Comparison in Coherence and Consistency:** In a few instances, human summaries did not perform as well as the AI in Coherence and Consistency. This could be attributed to the AI's algorithmic ability to detect patterns and logical sequences in the text.

c. **Challenges in Fluency:** Fluency was identified as the most challenging metric for the AI, suggesting that the AI, while adept at constructing grammatically correct sentences, has room to grow in emulating the more nuanced aspects of natural language that human writers naturally exhibit.

Building on these observations, it is clear that AI technology is approaching the threshold of human-like summarization capabilities. While the AI excelled in structural and factual aspects of the content, the human touch in language fluency remains a distinctive edge, especially in more creative or subjective interpretations.

The experiments underscore the progress AI has made in semantic comprehension and synthesis. Yet, they also highlight the complexities of language that AI has yet to master fully. As AI technology continues to evolve, we anticipate that subsequent models will close this gap, refining their linguistic output to match the stylistic nuances and expressive quality of human language.

Moreover, it is worth noting that the AI's performance in the realm of Fluency did not significantly detract from the overall clarity or informativeness of the summaries. Users seeking quick, accurate overviews of video content will find the AI summaries highly effective. However, for applications requiring a more narrative or engaging quality of language, further improvements in AI-generated text will enhance the user experience.

7 Conclusion/Future Work

This project successfully delivered a chatbot, available at https://ytchatbot.streamlit.app/, adept at summarizing YouTube video content, integrating the YouTube Transcript API for transcript extraction and OpenAI's GPT-3.5-turbo for generating summaries. The Streamlit-based user interface enables an accessible and efficient user experience for swift summarization retrieval.

The chatbot has showcased its efficacy as a tool for users to quickly digest and interact with video content. Summaries produced by the chatbot were coherent, succinct, and effectively distilled the crux of the videos, offering a streamlined approach to information consumption.

Future endeavors to enhance this chatbot could include:

a. **Summarization Quality Enhancement**: Applying more sophisticated NLP techniques and model refinements using specialized datasets could heighten the summarization precision, particularly for content dense with technical terminology or niche subjects.
b. **Real-Time Summarization:** Investigating real-time summarization could allow for dynamic summary generation, providing users with insights as they view the content, which could significantly benefit learning and content retention.
c. **Personalization:** Tailoring summaries to individual user preferences through adaptive algorithms could offer a more customized experience, ensuring relevance and user engagement.
d. **Interactive Features Expansion**: Augmenting the chatbot with capabilities to answer topic-specific inquiries and offering contextual information on demand would increase the chatbot's utility as an interactive learning tool.
e. **User Interface Enhancement:** Continual refinement of the user interface, potentially integrating voice commands and support for multiple languages, would improve accessibility and broaden the chatbot's user base.
f. **Performance Optimization:** Streamlining the chatbot's performance to accommodate a higher volume of simultaneous requests would ensure its scalability and reliability.
g. **Dataset Diversification**: Expanding the training dataset to encompass a wider array of content types would better equip the model to handle diverse summarization tasks effectively.

The chatbot stands as an embodiment of AI's potential to revolutionize our engagement with digital media, presenting a scalable solution to the challenge of navigating the extensive information available on platforms like YouTube. Anticipated future improvements aim to push the envelope of AI capabilities in content summarization and enhance the interactive experience for users.

8 Contributions

The entire project was a solo venture that spanned the spectrum of conceptualization, development, user experience design, and documentation. It demanded a multifaceted skill set—encompassing software development, natural language processing, API integration, and interface design—all of which were single-handedly orchestrated. This end-to-end project management underscores a proficiency in handling complex project requirements and the delivery of a functional, cutting-edge application. The project's fruition not only demonstrates technical competency but also a commitment to innovation in the AI summarization space.

9 References

[1] Streamlit.app, 2023. https://ytchatbot.streamlit.app/ (accessed Dec. 22, 2023).

[2] Bandabe, S., Zambre, J., Gosavi, P., Gupta, R., & Gaikwad, J. A. (2023). Youtube Transcript Summarizer Using Flask. In IJRASET Volume 11 Issue IV. Datta Meghe College of Engineering, Navi Mumbai, Maharashtra, India: IJRASET. DOI: 10.22214/ijraset.2023.50

[3] D. Look, “How Mosquitoes Use Six Needles to Suck Your Blood | Deep Look,” *YouTube*. Jun. 07, 2016. Available: https://www.youtube.com/watch?v=rD8SmacBUcU

[4] “How Mosquitoes Use Six Needles to Suck Your Blood,” *KQED*, Sep. 27, 2016. https://www.kqed.org/science/728086/how-mosquitoes-use-six-needles-to-suck-your-blood

[5] “how to build a capsule wardrobe,” *www.youtube.com*. https://www.youtube.com/watch?v=y-TPFKTnG_4 (accessed Dec. 22, 2023).

[6] “How to Build a Capsule Wardrobe – OnPointFresh.” https://onpointfresh.com/capsule-wardrobe/ (accessed Dec. 22, 2023).

[7] “Eminem’s Lose Yourself Lyrics: A Masterpiece of Rap Culture - Neon Music - Digital Music Discovery & Showcase Platform,” *neonmusic.co.uk*. https://neonmusic.co.uk/eminems-lose-yourself-lyrics-a-masterpiece-of-rap-culture/

[8] EminemExplicit, “Eminem - Lose Yourself (Official Video) (Explicit),” *YouTube*. Sep. 28, 2020. Available: https://www.youtube.com/watch?v=7YuAzR2XVAM

[9] “Build a basic LLM chat app - Streamlit Docs,” *docs.streamlit.io*. https://docs.streamlit.io/knowledge-base/tutorials/build-conversational-apps

[10] “How to evaluate a summarization task | OpenAI Cookbook,” *cookbook.openai.com*. https://cookbook.openai.com/examples/evaluation/how_to_eval_abstractive_summarization (accessed Dec. 22, 2023