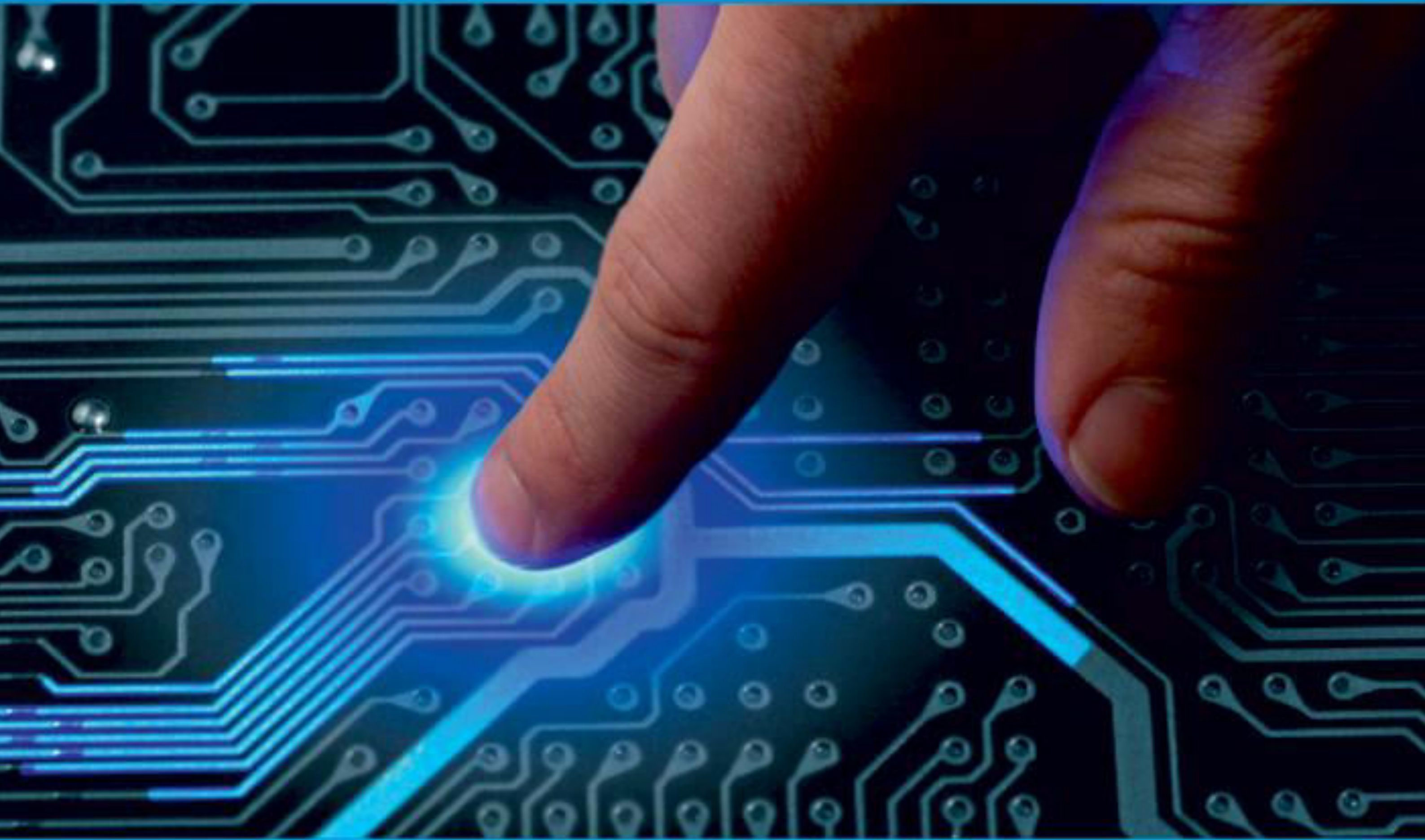
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204113 |

Interactive Audio Question Answering System with Whisper and Rag-Enhanced LLMS

[1] Sayantan Maity, [2] Avinash Reddy, [3] Megavath Shankar, [4] N.M.S. Desai

[1, 2, 3.] Students, Department of Computer Science and Engineering, Anurag University, Telangana, India

[4.] Associate Professor, Department of Computer Science and Engineering & Anurag University, Telangana, India

ABSTRACT: This paper addresses the limitations of passive listening in educational audio content, such as podcasts and lectures. We propose a novel interactive audio question answering (QA) system that empowers listeners to become active participants in their learning journeys. The core innovation lies in integrating Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs). Whisper, a state-of-the-art automatic speech recognition (ASR) tool, transcribes the audio into text, enabling the system to understand user queries within the context of the content. The LLM leverages RAG to access its internal knowledge base and retrieve relevant information from external sources using embeddings and a vector store. This retrieved information enhances the answer generation process, fostering a deeper understanding of the subject matter beyond the information explicitly presented in the audio. This interactive QA system with RAG-enhanced LLMs has the potential to revolutionize how listeners engage with educational audio content by enabling a more dynamic and personalized learning experience. Listeners can delve deeper into topics, clarify ambiguities, and ultimately gain a more comprehensive understanding of the presented material.

KEYWORDS: Audio Transcription, Whisper, Generative AI, Large Language Model (LLM), Retrieval Augmented Generation (RAG), Embeddings, Vector Store.

I. INTRODUCTION

The surge in popularity of audio-based platforms like podcasts and audiobooks has revolutionized information access. However, the current model of audio consumption often suffers from limitations. Listeners passively absorb information presented in a linear fashion, hindering deeper engagement and exploration. This lack of interactivity becomes particularly problematic for educational and informative audio content. Learners seeking a more active experience, or those requiring a nuanced understanding of complex topics, often encounter roadblocks. They lack the ability to ask clarifying questions or revisit confusing sections during audio playback. This linear format can lead to knowledge gaps and hinder a complete understanding of the presented material.

Recent advancements in Generative AI, particularly Large Language Models (LLMs), offer promising solutions for enhancing audio information consumption. LLMs possess vast knowledge and the ability to generate human-quality text, making them ideal candidates for developing interactive audio experiences. However, inherent limitations exist in LLM knowledge bases. This paper proposes a novel approach that leverages Retrieval-Augmented Generation (RAG) to overcome these limitations and create a robust interactive audio question answering (QA) system.

By combining Whisper-based audio transcription with RAG-enhanced LLMs, this proposed QA system fosters a more dynamic and personalized learning experience. It empowers listeners to delve deeper into topics, clarify ambiguities, and ultimately gain a more comprehensive understanding of the presented material. The following sections will delve deeper into the limitations of current audio information consumption, the functionalities of the proposed QA system, and the potential impact on educational content engagement.

II. RELATED WORK

The field of question answering (QA) has witnessed significant progress in recent years. Traditional methods focused on written text data, relying on manual information extraction and keyword matching to answer factual queries. However, advancements in natural language processing (NLP) have paved the way for more sophisticated systems.

These systems can handle complex questions and understand the context of the text using techniques like deep learning and transformers.

One crucial advancement in NLP is Automatic Speech Recognition (ASR). Tools like Whisper achieve impressive accuracy in converting spoken language to text, opening doors for audio-based QA systems. This eliminates the need for pre-processing audio data into text, a significant hurdle for traditional methods.

Furthermore, the rise of Retrieval-Augmented Generation (RAG) for Large Language Models (LLMs) has brought a new dimension to QA. RAG allows LLMs to not only leverage their internal knowledge but also focus on specific, relevant parts of retrieved information. This approach holds immense potential for audio-based QA, where the retrieved information comes from the ASR transcript

III. EXISTING METHOD

Traditional question answering (QA) systems primarily rely on pre-processed text data. This process often involves manual tasks like document cleaning, annotation, and feature engineering. These methods are not only labour-intensive and time-consuming but also may not scale well for large datasets. Additionally, the pre-processing step can lead to information loss, potentially affecting the accuracy and completeness of the answers.

While some existing audio-based QA systems exist, they may have limitations:

- **Transcript Dependency:** They might rely on pre-existing transcripts, which can be limited in availability and may not perfectly align with the audio content.
- **ASR Accuracy Limitations:** They might employ older, less accurate ASR models, leading to errors in the transcript and consequently, inaccurate answers.

Drawbacks of Existing Methods

Here's a more specific breakdown of the drawbacks associated with current approaches:

- **Limited Scalability:** Manual pre-processing of text data becomes cumbersome for large datasets, hindering scalability.
- **Information Loss:** Pre-processing steps might remove or alter information from the original source, leading to inaccurate or incomplete answers.
- **Limited Context:** Many current text-based and even some audio-based QA systems are restricted to the information explicitly provided within the text or transcript. This hinders their ability to answer questions that require reasoning or drawing connections to external knowledge that might be relevant to the conversation but not explicitly mentioned.

IV. PROPOSED METHOD

This paper proposes a novel interactive audio question answering system designed to empower users to engage actively with educational audio content. Here's a breakdown of the system's workflow with a focus on the information flow:

1. Preprocessing and Embedding Generation:

- The system begins by taking an audio file as input (podcast episode, lecture recording, etc.).
- Whisper, a state-of-the-art Automatic Speech Recognition (ASR) model, transcribes the audio into text, creating a complete textual representation of the spoken content.
- This transcript is then segmented into smaller, manageable chunks. These chunks can be individual sentences, paragraphs, or other meaningful units depending on the chosen strategy.
- Each text chunk is converted into a numerical representation called an embedding with the help of an embedding model. This embedding captures the semantic meaning of the text chunk in a high-dimensional space.

- These embeddings are then uploaded to a vector store, which acts as a high-performance database specifically designed for storing and retrieving vectors. This allows for efficient searching and retrieval based on semantic similarity.

2. User Interaction and Retrieval:

- When a user poses a question, the system processes the user's query to understand its intent and key terms. Similar to the text chunks, the user's question is also converted into an embedding.
- The vector store is then queried with the user's question embedding. This search retrieves the most semantically similar text chunk embeddings from the store.
- By considering the context provided by the retrieved transcript snippets, the system identifies the most relevant portions of the audio content based on their semantic closeness to the user's query.

3. Answer Generation with LLM:

- The retrieved transcript passages, along with the user's original question, are provided as input to a Large Language Model (LLM).
- The LLM leverages its knowledge and understanding of language to generate a comprehensive and informative answer to the user's query. Importantly, the context provided by the retrieved transcript snippets allows the LLM to tailor its response to the specific content and intent of the user's question.

Our proposed system offers several advantages:

- **Direct Audio Input:** This system eliminates the need for pre-processed text data, allowing users to interact directly with audio recordings.
- **Enhanced Efficiency:** By automating the transcription process with Whisper, the system reduces manual effort and increases overall efficiency.
- **Improved Contextual Understanding:** Natural Language Processing techniques facilitate a deeper understanding of the user's question and the context of the audio transcript. This allows the system to generate more relevant and accurate answers.
- **Access to External Information (via RAG):** Retrieval-Augmented Generation empowers the LLM to go beyond the audio content itself and access relevant information from external sources. This broadens the knowledge base used for answer generation, potentially leading to richer and more informative responses.
- **Interactive Learning Experience:** This system fosters a more interactive learning experience by allowing users to ask questions and receive answers directly within the context of the audio content.

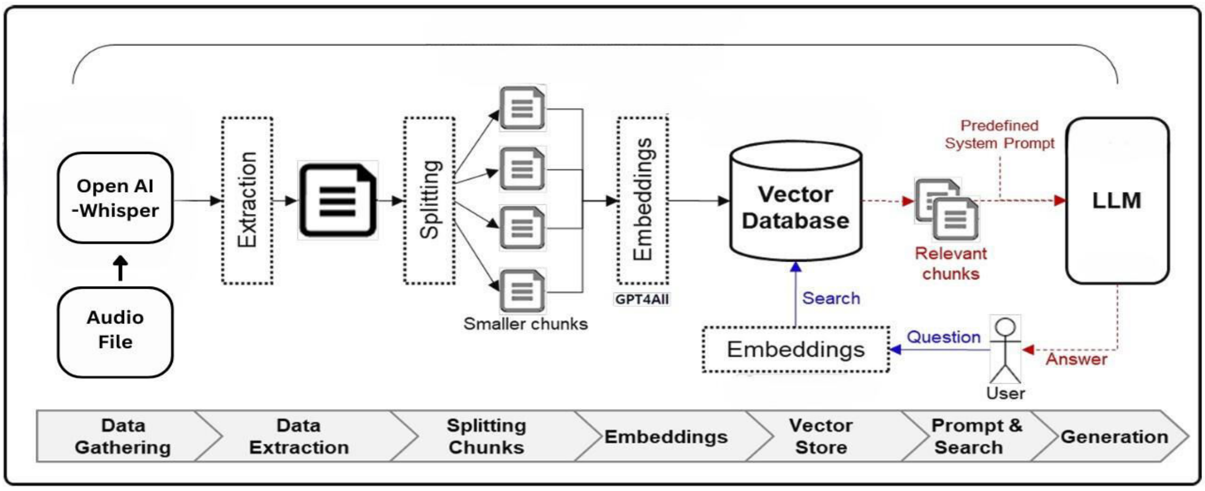


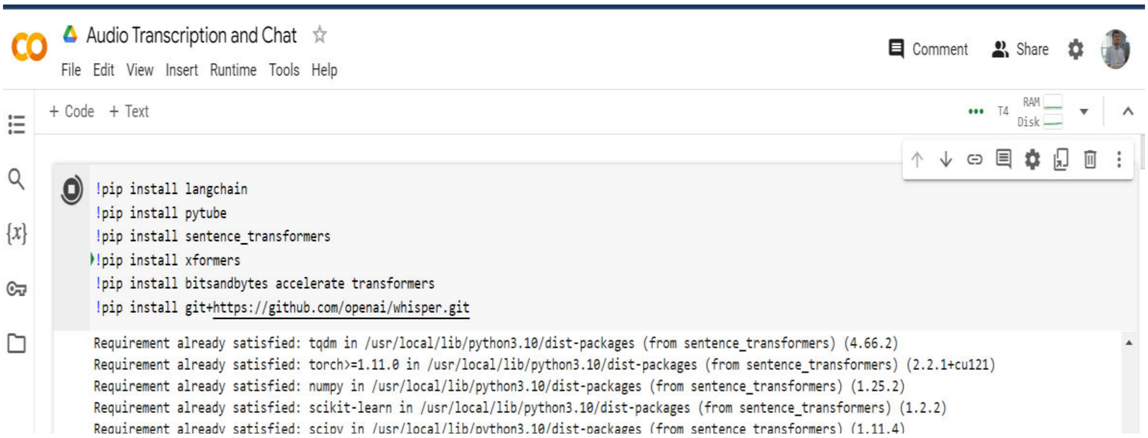
Fig 1: Flow Chart

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204113 |

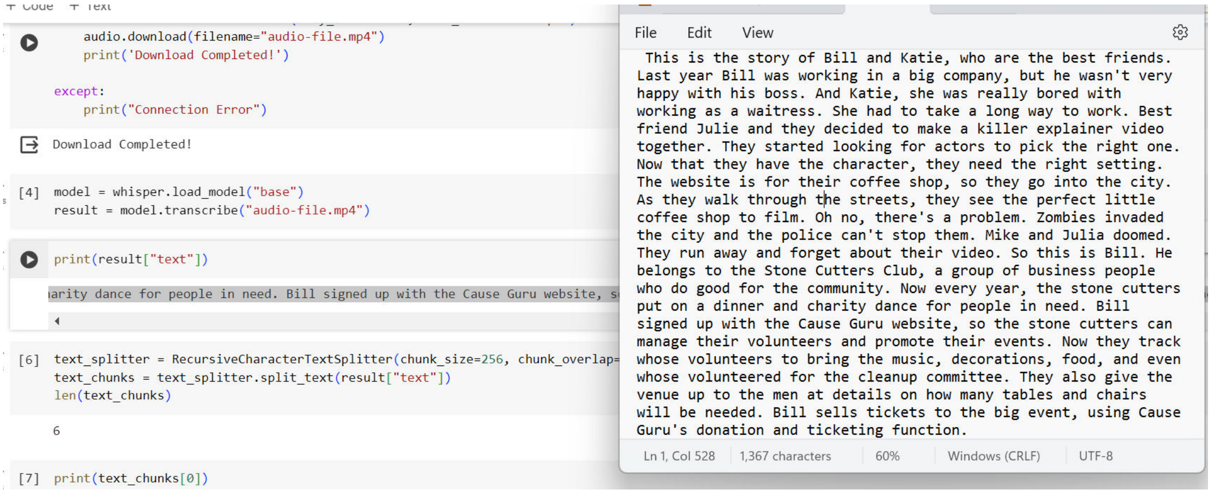
V. SIMULATION RESULTS


Audio Transcription and Chat
File Edit View Insert Runtime Tools Help
Comment
Share
+ Code + Text
T4
RAM
Disk

```
!pip install langchain
!pip install pytube
!pip install sentence_transformers
!pip install xformers
!pip install bitsandbytes accelerate transformers
!pip install git+https://github.com/openai/whisper.git
```

```
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from sentence_transformers) (4.66.2)
Requirement already satisfied: torch>=1.11.0 in /usr/local/lib/python3.10/dist-packages (from sentence_transformers) (2.2.1+cu121)
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (from sentence_transformers) (1.25.2)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.10/dist-packages (from sentence_transformers) (1.2.2)
Requirement already satisfied: scipy in /usr/local/lib/python3.10/dist-packages (from sentence_transformers) (1.11.4)
```

Fig 2: Environment – Google Colab with GPUs



```
        audio.download(filename="audio-file.mp4")
        print('Download Completed!')

    except:
        print("Connection Error")
```

```
Download Completed!
```

```
[4] model = whisper.load_model("base")
    result = model.transcribe("audio-file.mp4")
```

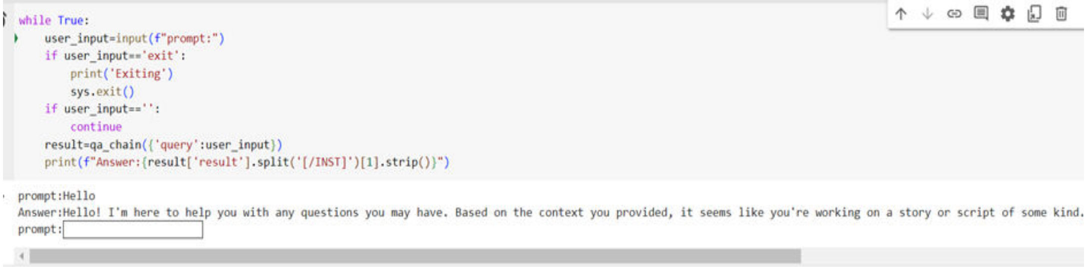
```
print(result["text"])
```

```
arity dance for people in need. Bill signed up with the Cause Guru website, s
```

```
[6] text_splitter = RecursiveCharacterTextSplitter(chunk_size=256, chunk_overlap=
    text_chunks = text_splitter.split_text(result["text"])
    len(text_chunks)

    6

[7] print(text_chunks[0])
```

```
 This is the story of Bill and Katie, who are the best friends.
Last year Bill was working in a big company, but he wasn't very
happy with his boss. And Katie, she was really bored with
working as a waitress. She had to take a long way to work. Best
friend Julie and they decided to make a killer explainer video
together. They started looking for actors to pick the right one.
Now that they have the character, they need the right setting.
The website is for their coffee shop, so they go into the city.
As they walk through the streets, they see the perfect little
coffee shop to film. Oh no, there's a problem. Zombies invaded
the city and the police can't stop them. Mike and Julia doomed.
They run away and forget about their video. So this is Bill. He
belongs to the Stone Cutters Club, a group of business people
who do good for the community. Now every year, the stone cutters
put on a dinner and charity dance for people in need. Bill
signed up with the Cause Guru website, so the stone cutters can
manage their volunteers and promote their events. Now they track
whose volunteers to bring the music, decorations, food, and even
whose volunteered for the cleanup committee. They also give the
venue up to the men at details on how many tables and chairs
will be needed. Bill sells tickets to the big event, using Cause
Guru's donation and ticketing function.
```

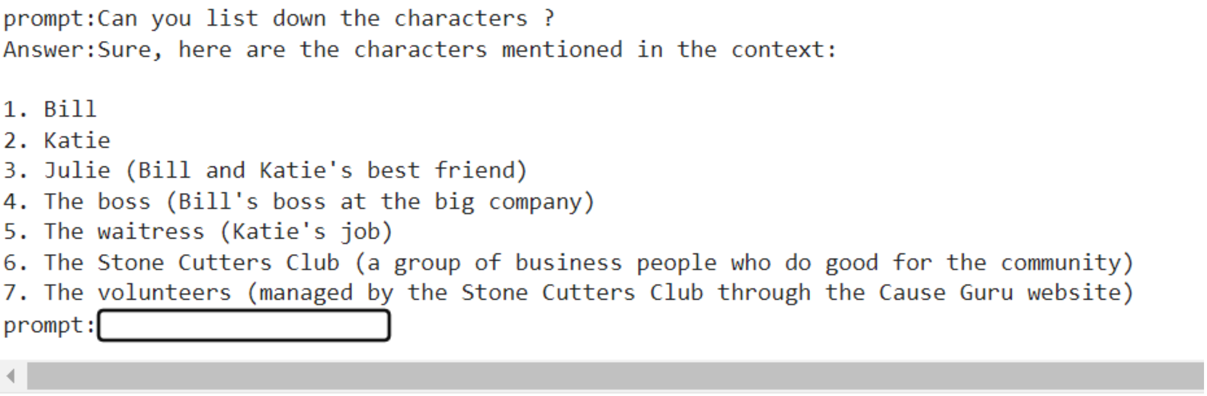
Fig 3: The transcribed audio text using Whisper.



```
while True:
    user_input=input(f"prompt:")
    if user_input=='exit':
        print('Exiting')
        sys.exit()
    if user_input=='':
        continue
    result=qa_chain({'query':user_input})
    print(f"Answer:{result['result'].split('[/INST]')[1].strip()}")
```

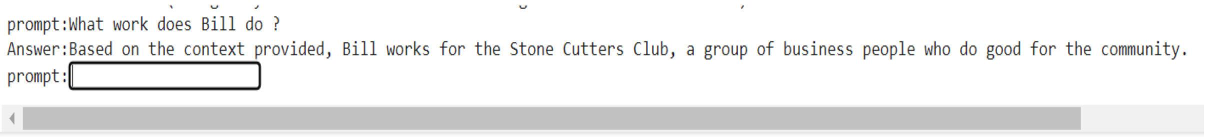
```
prompt:Hello
Answer:Hello! I'm here to help you with any questions you may have. Based on the context you provided, it seems like you're working on a story or script of some kind.
prompt:
```

Fig 4: Interactive Question Answering Loop



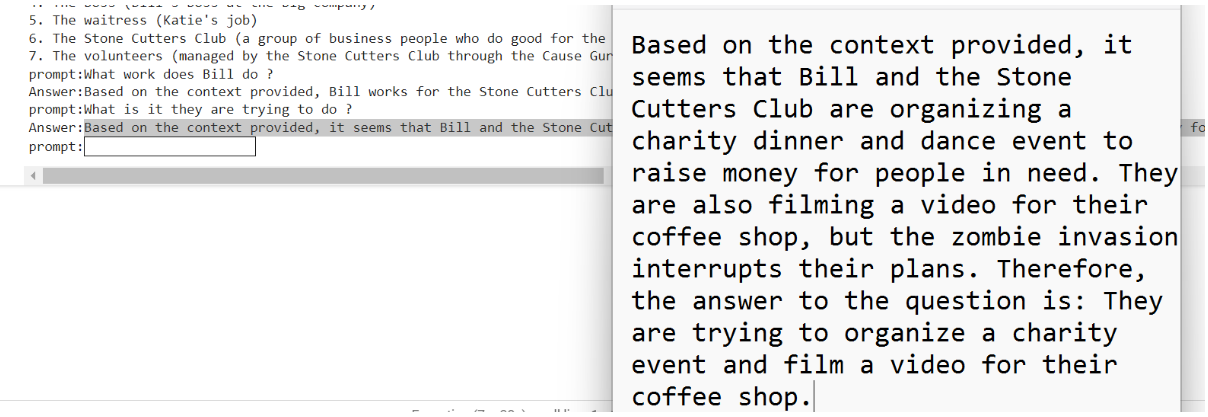
```
prompt:Can you list down the characters ?
Answer:Sure, here are the characters mentioned in the context:

1. Bill
2. Katie
3. Julie (Bill and Katie's best friend)
4. The boss (Bill's boss at the big company)
5. The waitress (Katie's job)
6. The Stone Cutters Club (a group of business people who do good for the community)
7. The volunteers (managed by the Stone Cutters Club through the Cause Guru website)
prompt:
```

Fig 5: Posing Questions - 1.



```
prompt:What work does Bill do ?
Answer:Based on the context provided, Bill works for the Stone Cutters Club, a group of business people who do good for the community.
prompt:
```

Fig 6: Posing Questions - 2.



```
4. The boss (Bill's boss at the big company)
5. The waitress (Katie's job)
6. The Stone Cutters Club (a group of business people who do good for the
7. The volunteers (managed by the Stone Cutters Club through the Cause Gur
prompt:What work does Bill do ?
Answer:Based on the context provided, Bill works for the Stone Cutters Clu
prompt:What is it they are trying to do ?
Answer:Based on the context provided, it seems that Bill and the Stone Cut
prompt:
```

```
Based on the context provided, it
seems that Bill and the Stone
Cutters Club are organizing a
charity dinner and dance event to
raise money for people in need. They
are also filming a video for their
coffee shop, but the zombie invasion
interrupts their plans. Therefore,
the answer to the question is: They
are trying to organize a charity
event and film a video for their
coffee shop.
```

Fig 7: Posing Questions - 3.

VI. CONCLUSION AND FUTURE WORK

This paper has presented a novel interactive audio question answering system designed to empower listeners to actively engage with educational audio content. The core innovation lies in the integration of Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs). By leveraging Whisper, a state-of-the-art automatic speech recognition (ASR) tool, the system can directly process audio data and generate informative answers to user queries within the context of the audio content.

The proposed system offers several advantages over traditional methods. It eliminates the need for pre-processed text data, allowing for a more efficient and user-friendly experience. Additionally, the combination of RAG and LLMs allows the system to access and incorporate information from both the audio transcript and external sources, potentially leading to richer and more comprehensive answers. This ability to access out-of-context information is a significant benefit compared to traditional systems. Furthermore, the interactive nature of the system fosters a more dynamic learning experience for users.

Future Work

This research lays the groundwork for further exploration and development in the field of interactive audio question answering. Here are some potential areas for future work:

- **Expanding the Knowledge Base:** Investigating techniques to integrate domain-specific knowledge bases into the LLM to enhance its understanding of particular educational topics.
- **Multilingual Support:** Exploring the feasibility of extending the system to handle audio content in multiple languages. This would require incorporating multilingual ASR models and LLMs trained on multilingual datasets.
- **Real-time Question Answering:** Developing the system to handle real-time audio streams, enabling users to ask questions and receive answers during live lectures or podcasts.
- **User Interface and User Experience Design:** Designing a user-friendly interface that facilitates a seamless question-answering experience within the chosen audio platform.
- **Evaluation and Refinement:** Conducting user studies to evaluate the system's effectiveness and accuracy in real-world educational settings. This feedback can be used to further refine the system and improve its capabilities.

By addressing these areas, we can further enhance the capabilities of this interactive audio question answering system, ultimately promoting a more engaging and enriching learning experience for users of educational audio content.

REFERENCES

[1]. Koh Matsuda;Ian Frank; **LangChain Unleashed:Advancing Education Beyond Chat-GPT's Limits**,Future University Hakodate,2024

[2]. Sriramaraju Sagi; **GenAI: RAG Use Cases With Vector Db to solve the limitations of LLMs**,(IJCET) Volume 15, Issue 2, March-April 2024, pp. 56-62, Article ID: IJCET_15_02_008

[3]. Alec Radford;Jong Wook Kim;Tao Xu1 Greg Brockman;Christine McLeavey;Ilya Sutskever - **Robust Speech Recognition via Large-Scale Weak Supervision,2022**.

[4]. Cheonsu Jeong; Principal Consultant & the Technical Leader for AI Automation Platform at SAMSUNG SDS,Seoul 05510, Korea - **A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture**.

[5]. Muhammad Usman Hadi, Qasem Al-Tashi, Rizwan Qureshi, Abbas Shah - **Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects**

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details