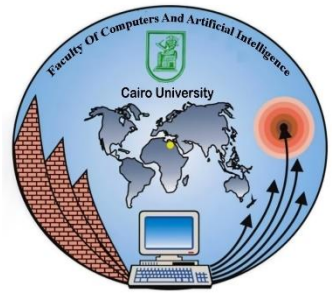

Faculty of Computers and Artificial Intelligence

Cairo University

Fine-Tuning of Gemini

Prepared By:

Mahmoud Wael Ebrahim Mohamed	20200505
Yousef Sherif Mohamed Ahmed	20200655

Artificial Intelligence Department

Cairo University

Academic Year: 2023/2024

July 4, 2024

Table of Contents

Table of Figures

List of Tables

Abstract

This project explores the integration of a chatbot powered by the advanced large language model (LLM) Gemini, enabling users to interact with video content in a more intuitive and engaging manner. By allowing users to pose questions related to video content, the chatbot interprets queries and generates responses based on subtitles and spoken dialogue, thereby enriching the viewing experience. Despite challenges in feature extraction and model performance, we implemented effective solutions to enhance usability and accuracy. A series of experiments were conducted to optimize model performance and identify the most suitable LLM for integration into the system. The findings demonstrate the significant potential of leveraging LLMs for interactive multimedia applications, paving the way for future enhancements in video content accessibility.

Chapter 1: Introduction

Large Language Models (LLMs) have transformed our daily lives by providing highly accurate and contextually relevant answers to a wide array of questions. Their ability to comprehend and generate human-like text has made them invaluable tools in various domains, including customer support, education, and personal assistance. Leveraging the power of LLMs, it is now possible to interact with video content in a more natural and intuitive manner, allowing users to ask questions about video content and receive textual answers to their queries. The rise of multimodal datasets, such as TVR and TVQA, underscores the need for systems that can effectively integrate visual and textual data, facilitating deeper understanding and interaction.

In this context, our project focuses on utilizing LLMs to enable users to interact with video content through conversational queries. By integrating advanced natural language processing capabilities with video analysis, we aim to enhance the user experience, making information retrieval from video content more seamless and efficient. This initiative not only addresses the growing demand for interactive video tools but also opens avenues for further research and development in the fields of artificial intelligence and multimedia content management.

Chapter 2: Related Work

2.1. The Datasets:

TVR Dataset: [1]

The authors present a new multimodal retrieval dataset called the TV show Retrieval (TVR) dataset. Unlike previous datasets, TVR requires systems to comprehend both videos and their accompanying subtitle (dialogue) texts, reflecting real-world scenarios more accurately. The dataset comprises 108,965 queries obtained from 21,793 videos sourced from 6 TV shows spanning diverse genres. Each query is linked to a specific timeframe within the video. TV shows were chosen as the data source due to their characteristic rich social interactions and dialogue exchanges between actors, offering varied content. The dataset includes labels indicating whether each query is primarily related to the video, subtitle, or both, facilitating detailed dataset analysis and method evaluation. On average, the videos are 76.2 seconds long and are paired with corresponding subtitles. To collect the TVR dataset, the authors utilized Amazon Mechanical Turk (AMT), where workers were tasked with generating queries based on video and/or subtitle information. They were also required to specify the start and end timestamps defining a moment matching the query. Each query-moment pair had to be uniquely identifiable within the video, ensuring precise localization. Workers were prompted to classify queries into three types: video-only, relevant to visual content; sub-only, pertinent to subtitles; and video + sub, involving both modalities.

A graph representing the query type distribution of the dataset:

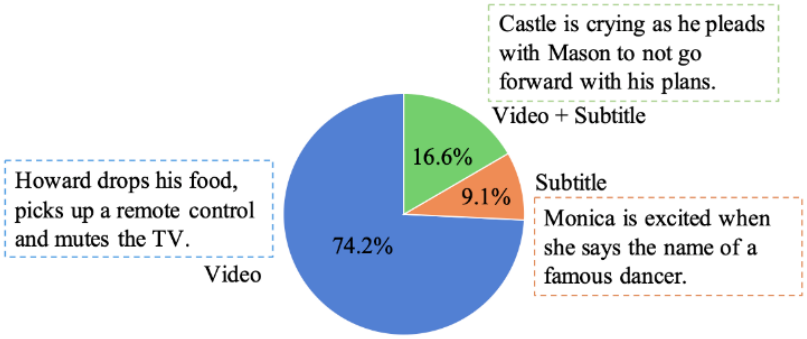

Figure 1: The query type distribution of the TVR dataset

TVQA Dataset: [2]

The dataset is built on six long-standing TV shows spanning three genres: sitcoms, medical dramas, and crime drama (same as the TVR dataset). This collection comprises 7.3 seasons for each show, with a total of 925 episodes. Altogether, it adds up to 461 hours of content, enabling extensive character interactions and evolving relationships over time. To facilitate analysis and understanding, each episode was further segmented into shorter clips resulting in 21,793 video clips ranging from 60 to 90 seconds long. The authors carefully shifted temporal boundaries to avoid splitting subtitle sentences between clips.

Additionally, the authors collected 152,545 human-written question-answer pairs (QA pairs), where each video clip is associated with seven questions and five answers (one of them is correct) provided for each question. Notably, negative answers were crafted by human annotators to ensure relevance and challenge. Each clip is accompanied with subtitles and aligned with transcripts, which include character names, emphasizing the significance of understanding the relationship between the provided dialogue and the QA pairs for accurate responses.

Finally, the dataset features compositional questions, necessitating algorithms to localize relevant moments within the videos. For each question, start and end points are provided, aiding in this localization process. To gather the question-answer pairs, the authors utilized Amazon Mechanical Turk, where workers were presented with both videos and aligned

named subtitles. This setup encouraged the creation of multimodal questions that require both visual and language understanding to answer effectively. Workers were asked to create questions using a compositional-question format: What/How/Where/Why/When/Before/After. The second part of each question serves to localize the relevant video moment within a clip, while the first part poses a question about that moment. This compositional format also serves to encourage questions that require both visual and language understanding to answer, since people often naturally use visual signals to ground questions in time.

A graph representing the distribution of question types based on answer types:

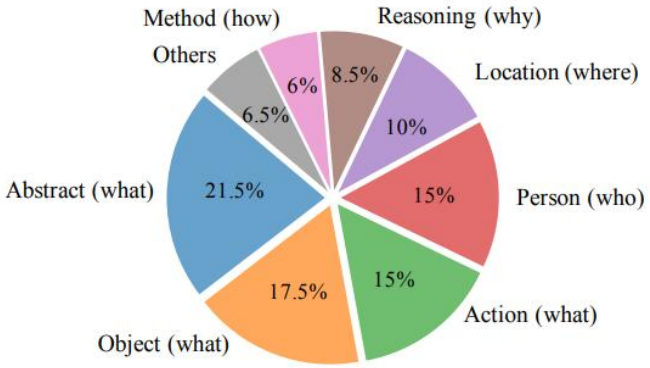

Figure 2: The TVQA dataset question types distribution based on answer types

Chapter 3: Methodology

3.1. LLMs:

In recent years, the advent of large language models (LLMs) has revolutionized the field of natural language processing (NLP). These models, with their ability to understand and generate human-like text, have found applications in numerous domains, including conversational agents, text summarization, and question answering. In our project, we aim to leverage the power of LLMs to enhance video content accessibility and interactivity.

Our primary objective is to evaluate the performance of three prominent LLMs ─LLaMA2, BERT, and Gemini─ on video transcripts. By doing so, we intend to create a system where users can ask questions related to the video's content, and the model can provide accurate textual answers. Such a system has broad applications in educational platforms, corporate training modules, and multimedia content management, where quick and precise access to information is paramount.

3.1.1. Overview of the Models:

LLaMA2 (Large Language Model Meta AI):

LLaMA2, developed by Meta (formerly Facebook), is one of the latest advancements in the field of large language models. Designed to handle a wide variety of natural language processing tasks, LLaMA2 leverages a sophisticated transformer architecture that allows it to understand and generate human-like text with remarkable accuracy. Its large-scale pre-training on diverse datasets enables it to capture complex linguistic patterns and nuances, making it a powerful tool for tasks such as text summarization, question answering, and text generation. LLaMA2's architecture is particularly notable for its scalability and efficiency, which are crucial for applications requiring real-time responses and high throughput.

BERT (Bidirectional Encoder Representations from Transformers):

BERT, introduced by Google in 2018, revolutionized the field of NLP with its bidirectional training approach. Unlike traditional models that process text sequentially, BERT reads the entire sequence of words simultaneously, allowing it to consider the context from both directions. This bidirectional mechanism significantly enhances BERT's understanding of context and semantics, making it highly effective for various NLP tasks, including named entity recognition, question answering, and language inference. BERT's pre-training on vast amounts of text data followed by task-specific fine-tuning has set new benchmarks for performance across numerous NLP applications.

Gemini:

Gemini, developed by Google, is designed specifically for conversational AI applications. It is fine-tuned to excel in generating coherent and contextually appropriate responses in dialogue systems. Gemini's architecture is optimized for understanding the nuances of human conversation, making it ideal for tasks involving customer service, virtual assistants, and interactive chatbots. With its ability to adapt to different conversational contexts and maintain the flow of dialogue, Gemini provides a user-friendly interface for engaging with AI-driven systems. Its performance in generating accurate and contextually relevant responses makes it a valuable asset in enhancing user interactions with automated systems.

3.1.2. Extractive vs. Abstractive Methods:

First of all, LLMs can answer the user's questions in two ways: abstractive and extractive.

The abstractive method involves generating new sentences that may not be present in the original text but convey the necessary information. This method allows the model to synthesize and reformulate information in a coherent and concise manner. Models like LLaMA2 and Gemini employ the abstractive method, which enables them to provide more natural and human-like responses, often summarizing or interpreting the content rather than quoting it verbatim.

The extractive method, on the other hand, locates and extracts relevant portions of the text to form an answer directly from the given content. This method ensures that the response

is directly grounded in the original text, maintaining a high degree of accuracy. BERT is an example of a model that utilizes the extractive method, highlighting its strength in pinpointing specific information within the text and presenting it as the answer.

3.1.3. Testing the Models' Performance:

To assess the performance of the three mentioned models, we passed each one a transcript from a video to test their response accuracy. We focused solely on evaluating the correctness of the responses provided by each model.

3.1.4. Prompt Engineering:

To further refine the performance of LLaMA2 and Gemini, we employed prompt engineering, a technique used to guide LLMs by crafting specific input prompts that influence their responses. Prompt engineering can significantly impact the effectiveness of an LLM by providing context and framing questions in a way that elicits the most relevant and accurate answers. However, this technique was not applicable to BERT, as it does not support prompt engineering in the same way.

Through prompt engineering, we guided LLaMA2 and Gemini to better understand and respond to user queries, enhancing their usability and accuracy

By comparing these methodologies, we aim to identify which model yields the most accurate and useful responses for our application.

3.2. Mobile Application:

The resulting model, designed to enhance multimedia retrieval, and the LLM, will be deployed on a mobile app platform developed using Flutter framework and Dart programming language, providing users with seamless access to our product.

3.2.1. Mobile App Features:

- **Chatbot Conversation Page:**

The chatbot conversation page provides a dynamic interaction platform with the following features:

Welcome Message: Users are greeted with a welcome message at the start of the conversation.

Text Field: Users can enter their messages, which are sent to the LLM. The model then generates a response, which appears in the chat.

- ## Settings Page:

  The settings page includes several important features:

  User Data Management: Users can check and update their personal data.

  Log Out Option: Users can log out of their accounts from this page.

- ## Integration:

  Integration with Flask was a critical component of our application architecture. We utilized Flask, a lightweight web framework for Python, to serve as the intermediary between the mobile application and our backend models

- ## Additional Features:

  Data Storage: All user data is securely stored on Google's Firebase platform.

  Caching: Onboarding pages and sign-in/sign-up processes are cached to appear only once unless the user logs out.

3.2.2. UI Design:

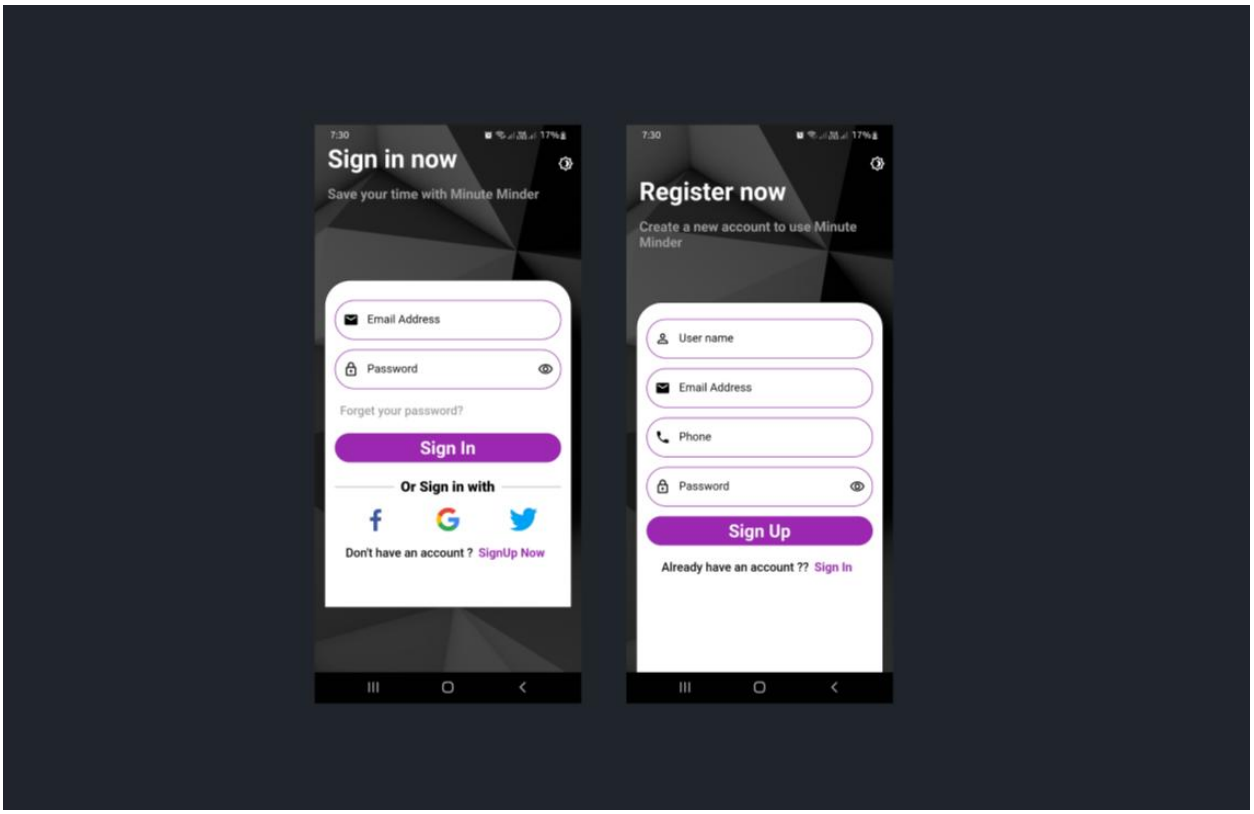

Figure 3: Login & Register

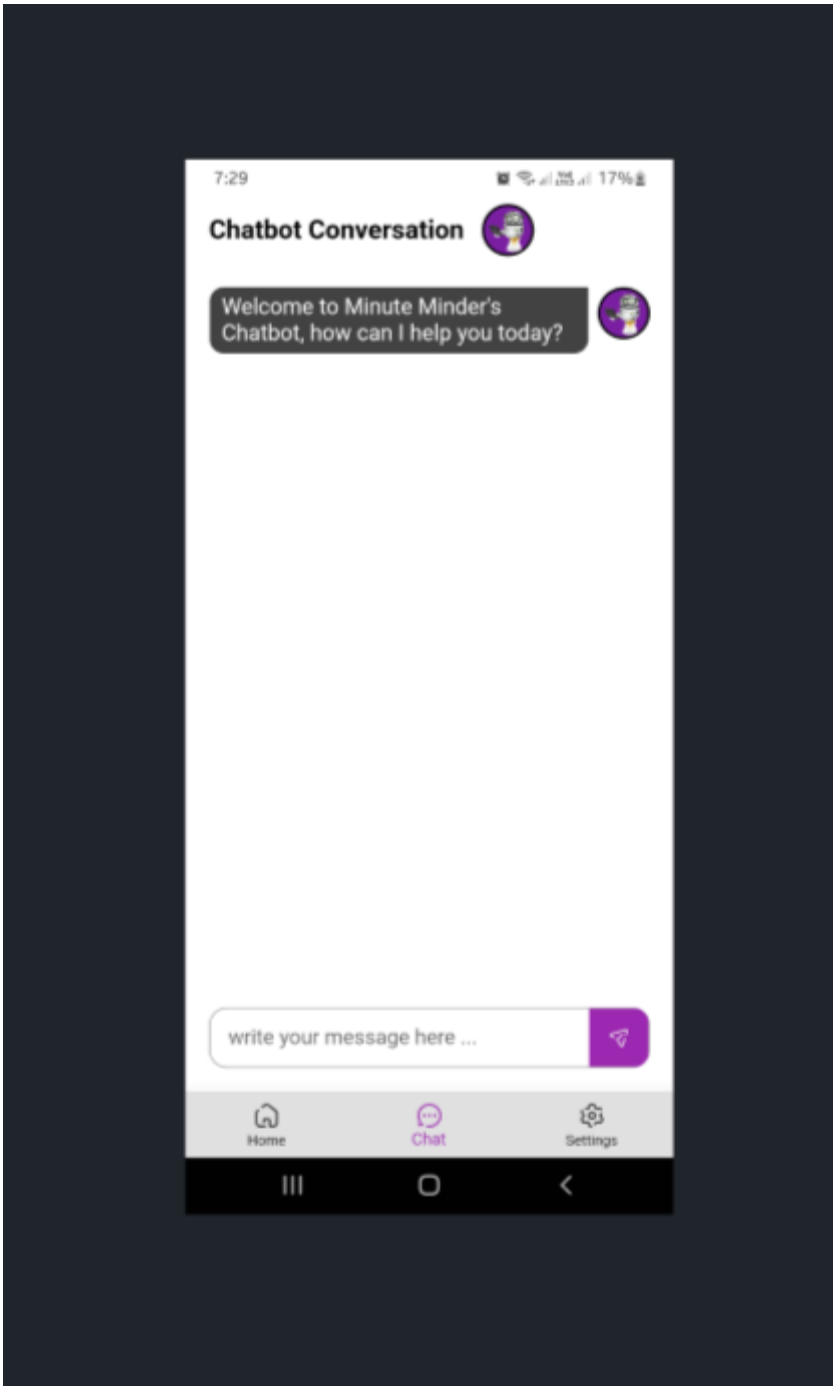

Figure 4: Chatbot Conversation Page

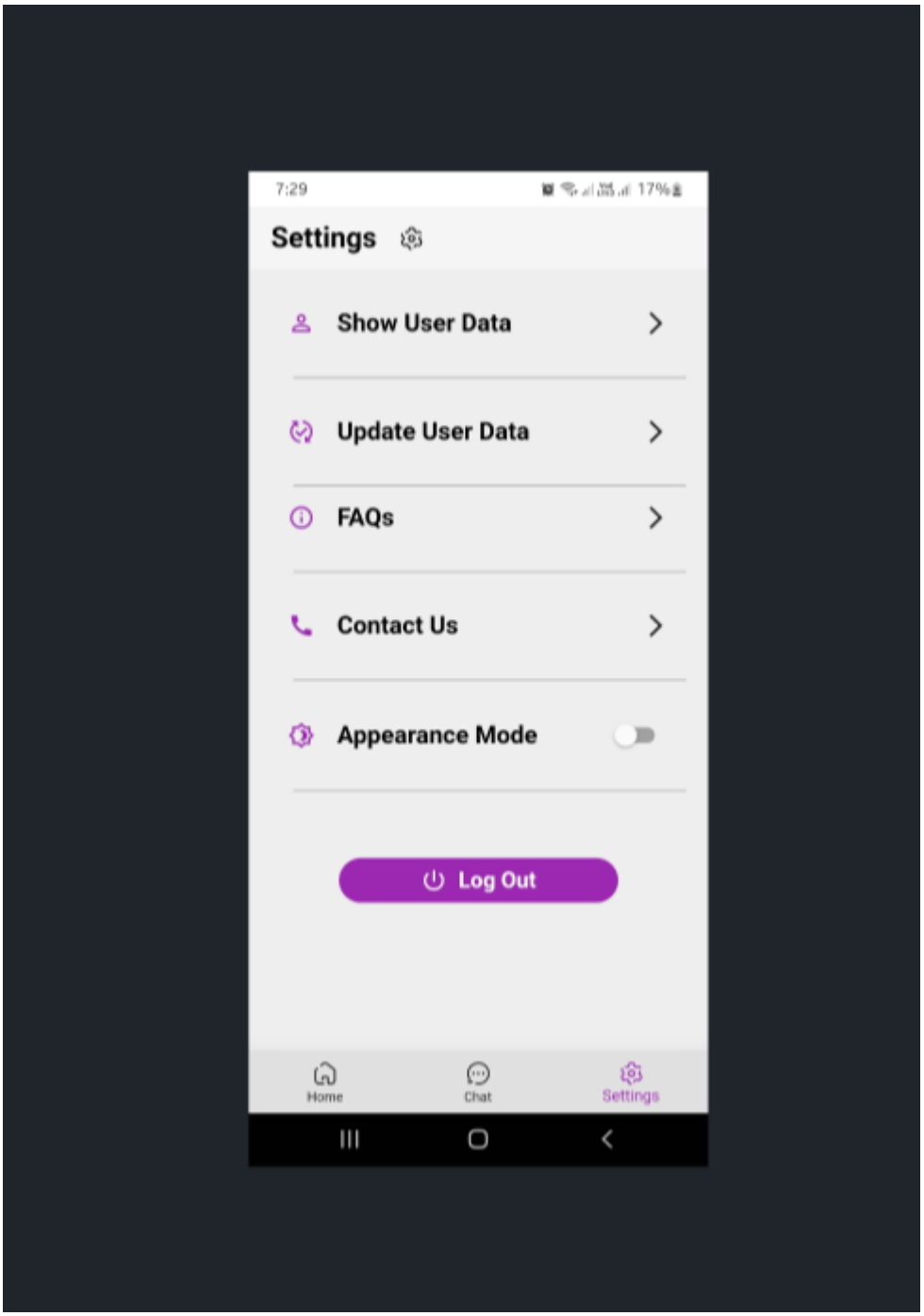

Figure 5: Settings Page

Chapter 4: Results & Analysis

4.1. Fine Tuning of Gemini:

Fine-tuning large language models (LLMs) involves adapting a pre-trained model to a specific task or dataset. This process allows the model to learn task-specific patterns and nuances, enhancing its performance and accuracy in generating relevant responses. Fine-tuning can significantly improve an LLM's ability to understand and answer user queries by aligning its predictions more closely with the specific context and content of the data it has been exposed to.

4.1.1. Choosing the Model:

Our primary objective was to develop LLMs—specifically LLaMA2, BERT, and Gemini—on video transcripts. To evaluate these models' performance, we conducted a series of experiments, comparing the answers generated by each model. The evaluation focused on the accuracy and relevance of the answers provided in response to user queries about the video's content.

To further guide the models, we employed prompt engineering for Gemini and LLaMA2, but not for BERT, as it does not support prompt engineering. The prompt used was:

"You are an expert assistant and a QA conversational chatbot. I will provide you with a CSV file containing three columns: start_time, end_time, and text. Your task is to consolidate all the entries under the 'text' column to comprehend the content as a whole. Once done, I will begin asking you questions about this consolidated text. Your responses should be based solely on the information provided in the CSV file.
If you cannot find the answer within the CSV file, please respond with 'I can't find an answer for your question in the provided video's transcript."

To ensure fair comparison, each model was fed with the same CSV file containing the transcript of one of the videos from the TVR dataset and was asked the same set of questions. The table below illustrates the results of our comparison study.

Question	Human Answer	LLaMA2 answer	BERT's answer	Gemini's answer
What does the neighbors called to complain about?	Neighbors called to complain about the music	The neighbors called to complain about the music	The music	Neighbors called to complain about the music
What was unusual about Alison Tisdale's living situation?	Despite being a social work student, Alison lived in a nice place, which was due to her father's money	Alison Tisdale's living situation was unusual because she was a grad student at NYU	She wasn't seeing anyone	Alison Tisdale lived in a nice place for a social worker, which indicates she had money
What does Castle say about romance?	Castle sarcastically comments that he says romance is dead every Saturday night.	Castle says that romance is dead	Dead	Castle says that romance is dead
How was Alison Tisdale found?	Alison Tisdale was found with two shots to the chest from a small caliber weapon	Alison Tisdale was found with two shots to the chest	Dead	She was found dead
What was found at the crime scene involving Alison Tisdale?	Castle noted that the setup of roses on her body and sunflowers on her eyes was familiar from previous cases	Roses on her body, sunflowers on her eyes	The rose petals	The crime scene involving Alison Tisdale was found to have no signs of a struggle, flowers, and roses on her body.
What does Castle's daughter, Alexis, prioritize even during a party?	Alexis prioritizes her homework, showing her responsibility.	She is a student at NYU	She bored money. The little she had; she gave to charity	Alexis prioritizes her homework even during a party.
Who is Cristiano Ronaldo?	Not in the transcript	Cristiano Ronaldo dos Santos Aveiro is a Portuguese professional footballer…	He already caught the killer	This conversation does not contain any information about Cristiano Ronaldo

Based on these experimental results, we can conclude that Gemini and LLaMA2 exhibited superior contextual understanding compared to BERT and provided responses closer to the human answers. This is due to the way each model answers questions, Gemini and LLaMA2 answer extractively, while BERT answers abstractively. Additionally, BERT showed limitations due to its lack of support for prompt engineering, resulting in less accurate and sometimes irrelevant answers

Choosing between LLaMA2 and Gemini, we observed that Gemini excelled in understanding context and providing human-like answers based on the context it was fed. Even when we asked both models a question not included in the transcript, Gemini responded appropriately with, 'This conversation does not contain any information about Cristiano Ronaldo,' demonstrating the effectiveness of our prompt engineering. As a result, we chose Gemini to fine-tune on the dataset we prepared.

4.1.2. Dataset Preparation:

For the fine-tuning process, the model needed the data to be in Input-Output format, this is where the TVQA dataset comes into play. TVQA dataset was selected due to its alignment with the TV shows present in the TVR dataset.

However, directly feeding the QA pairs from the TVQA dataset for fine-tuning was not feasible, as the model required contextual understanding to generate coherent answers. Instead, the QA pairs were utilized as guiding information. To achieve this, we integrated subtitles from the respective TV shows with the QA pairs.

The data preparation involved several methodical steps:

1. The subtitles for each clip were combined into a single context string. This string provided a comprehensive view of the dialogue within each clip.
2. For each clip, the corresponding questions and their answers were extracted and paired with the clip's context. This resulted in a detailed data structure containing the context and associated QA pairs for each clip.

3. To prepare the filtered data to be in the Input-Output format, we combined the context (subtitles) with the questions from the QA pairs to be the ‘Input’. Each question was prefixed with "Q:" to clearly mark it, and combined the answers from the QA pairs, each prefixed with "A:" to be the ‘Output’.

4.1.3. Fine Tuning the Model:

Due to constraints limiting the training data to a maximum of number of samples × number of epochs, we opted to fine-tune the model using data from two specific shows: "Castle" and "Friends." This focused approach allowed us to effectively manage training resources while tailoring the model to these selected datasets.

The fine-tuning process involved training the model for 15 epochs with a learning rate set at 0.001 and a batch size of 4. These parameters were carefully chosen to optimize learning and model performance within the given computational constraints.

By fine-tuning on a targeted dataset comprising episodes from "Castle" and "Friends," we aimed to enhance the model's understanding and responsiveness specifically within the context of these shows. This approach not only streamlined training but also aimed to improve the model's ability to generate accurate responses based on the content and style unique to each series.

4.1.3. Fine Tuning Results:

The outcomes of the fine-tuning process yielded impressive results, with the model consistently providing accurate answers, exhibiting minimal error margins across its responses. Below are examples of questions crafted by humans based on the show, along with the corresponding responses generated by the model:

Question: What are the names of Beckett's partners?
Response: Ryan and Esposito

Question: Does Castle have a daughter?
Response: Yes, castle has a daughter named Alexis.

Question: What is the show 'Castle' about?

Response: Castle is a crime drama television series that follows the lives of Richard Castle, a mystery novelist, and Kate Beckett, a homicide detective.

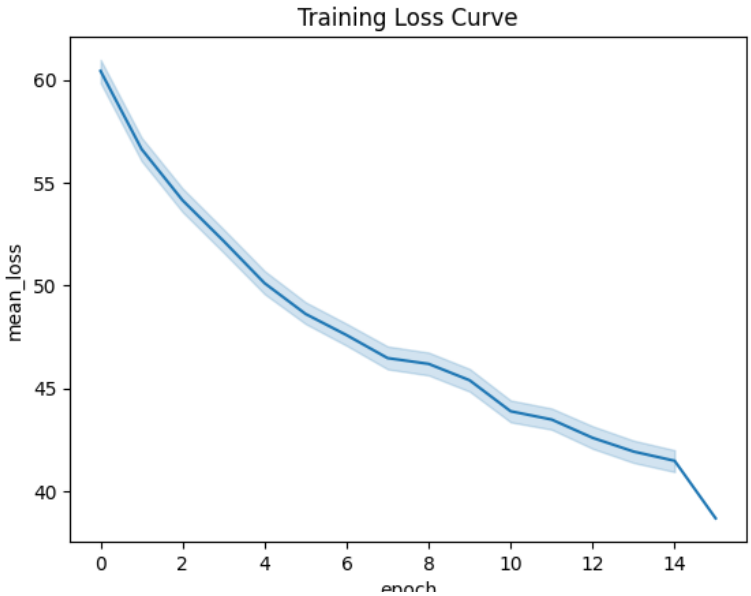

Figure 6: Gemini Fine-tuning training loss over 15 epochs

Conclusion

The centerpiece of our project was the integration of a chatbot powered by the Gemini LLM, which enables users to ask questions about video content and receive detailed, contextually relevant answers. By leveraging the video's subtitles and spoken dialogue, the chatbot provides a comprehensive interactive experience that enhances user engagement. Throughout the development and experimentation phases, we encountered and addressed various challenges, including optimizing feature extraction processes, fine-tuning model parameters, and ensuring the system's overall usability and accuracy.

Our results highlighted the effectiveness of combining visual and textual data for moment retrieval, showcasing the potential of LLMs in enhancing user interaction with video content. Although the current scope of our project is limited to a specific dataset, our vision is to generalize the system to work with diverse video content in the future. This project establishes a foundation for further advancements in video content analysis and interactive multimedia applications, addressing the growing need for efficient and intuitive tools to manage and interact with video data. Future work will involve expanding the dataset scope and exploring additional multimodal integration techniques to further enrich the interactive experience.

References

[1] Jie Lei, Licheng Yu, Tamara L. Berg and Mohit Bansal "TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval" published in Aug. 2020. [2001.09099] TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval (arxiv.org)

[2] Jie Lei, Licheng Yu, Mohit Bansal and Tamara L. Berg "TVQA: Localized, Compositional Video Question Answering" published in May. 2019. [1809.01696] TVQA: Localized, Compositional Video Question Answering (arxiv.org)

[3] TVR dataset examples: UNC TVR/TVC Dataset

[4] TVQA dataset examples: TVQA Dataset (unc.edu)