1

Datasets

In this segment, we will explore the fundamental datasets that serve as the foundation for our project. We will examine the unique characteristics and purposes of the TVR and TVQA datasets.

Gemini Fine Tuning

Team Members

Names	IDs
Mahmoud Wael	20200505
Yousef Sherif	20200655

1

Datasets

In this segment, we will explore the fundamental datasets that serve as the foundation for our project. We will examine the unique characteristics and purposes of the TVR and TVQA datasets.

2

Extractive VS. Abstractive

In this section, we will analyze the fundamental differences between extractive and abstractive models

3

Choosing the Model

In this section, we will outline our approach to selecting and evaluating various large language models (LLMs) for processing video transcripts, specifically focusing on LLaMA2, BERT, and Gemini.

4

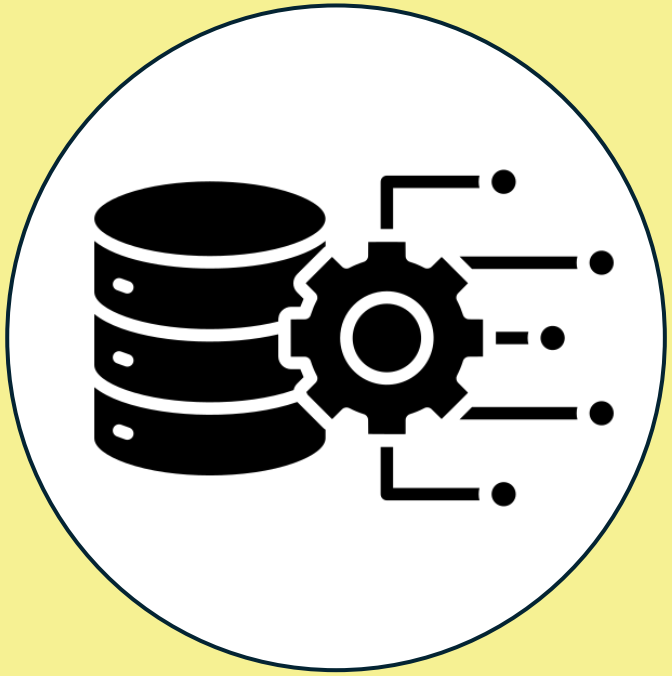
LLM Fine-Tuning

In this section, we will discuss the fine-tuning process of our model. Additionally, we will describe our approach to preparing the dataset by integrating subtitles with question-answer pairs.

5

Running Demo

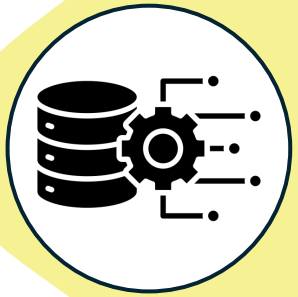
In the Demo section, we showcase a live demonstration of our project's capabilities. This segment provides a hands-on experience, illustrating how the system functions in real-world scenarios.



Datasets

TVR Dataset

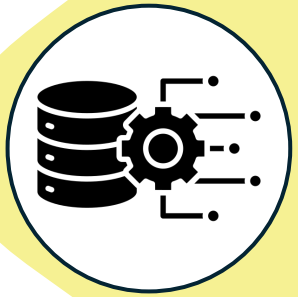
- Unlike previous datasets, TVR requires systems to comprehend both videos and their accompanying subtitle (dialogue) texts. The dataset comprises 108,965 queries obtained from 21,793 videos sourced from 6 TV shows spanning diverse genres. Each query is linked to a specific timeframe within the video



- The dataset includes labels indicating whether each query is primarily related to the video, subtitle, or both

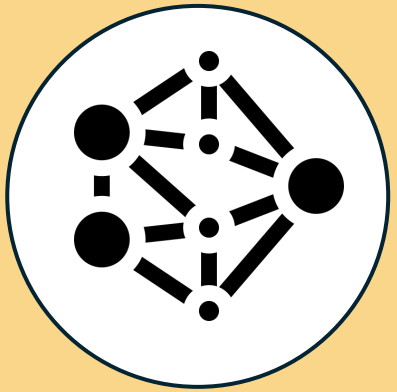
TVQA Dataset

- TVQA dataset is built on six long-standing TV shows, comprising 7.3 seasons each, totaling 925 episodes and 461 hours of content (same as the TVR dataset).
- The dataset includes 152,545 human-written question-answer pairs, featuring compositional questions and aligned subtitles for each clip, 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
- 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



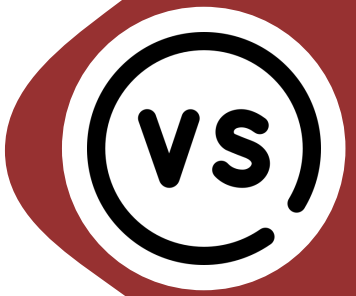
Extractive vs. Abstractive

Choosing the Model



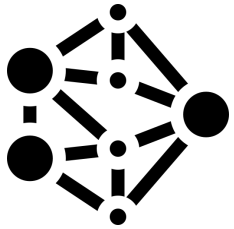
LLM Fine-Tuning

Extractive vs. Abstractive



Extractive vs. Abstractive Models

- **Extractive Models:** 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 (Ex. BERT)



- **Abstractive Models:** 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 (Ex. Gemini , LLaMA2)

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.

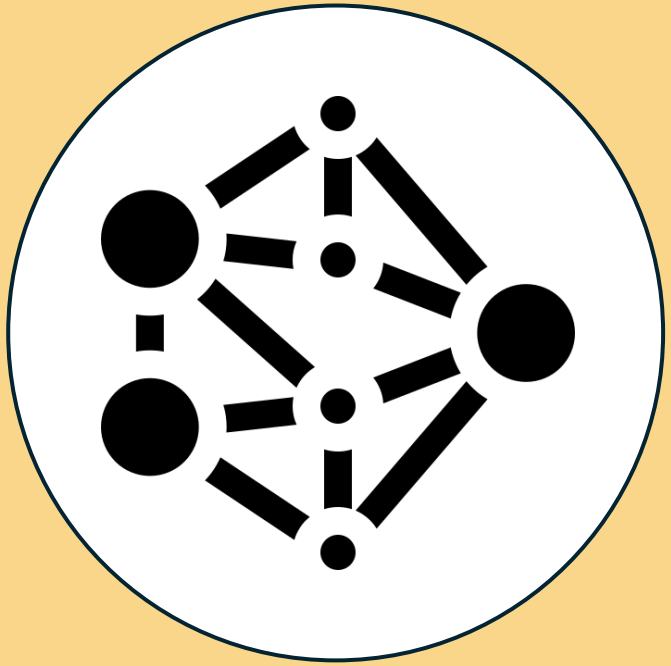
- To further guide the models, we employed prompt engineering for Gemini and LLaMA2, but not for BERT, as it does not support prompt engineering To ensure fair comparison, each model was fed with the same CSV file containing the transcript of one of the videos from the TVR dataset and was asked the same set of questions

Comparison Results

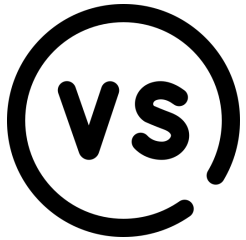
- Based on our experimental results, we concluded that Gemini and LLaMA2 exhibited superior contextual understanding compared to BERT and provided responses closer to the human answers as they answer abstractively.

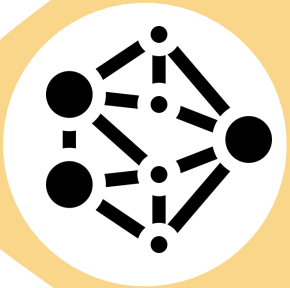
- BERT showed limitations due to its lack of support for prompt engineering and its extractive way in answering questions, resulting in less accurate and sometimes irrelevant answers

- 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 that the text doesn't mention any information about the asked question



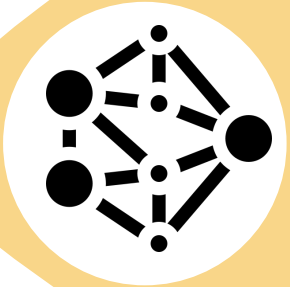
LLM Fine-Tuning





- 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, as the XML model was trained on the TVR dataset, we aimed to fine-tune Gemini using data from the same TV shows featured 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.



Preparing the Dataset

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.

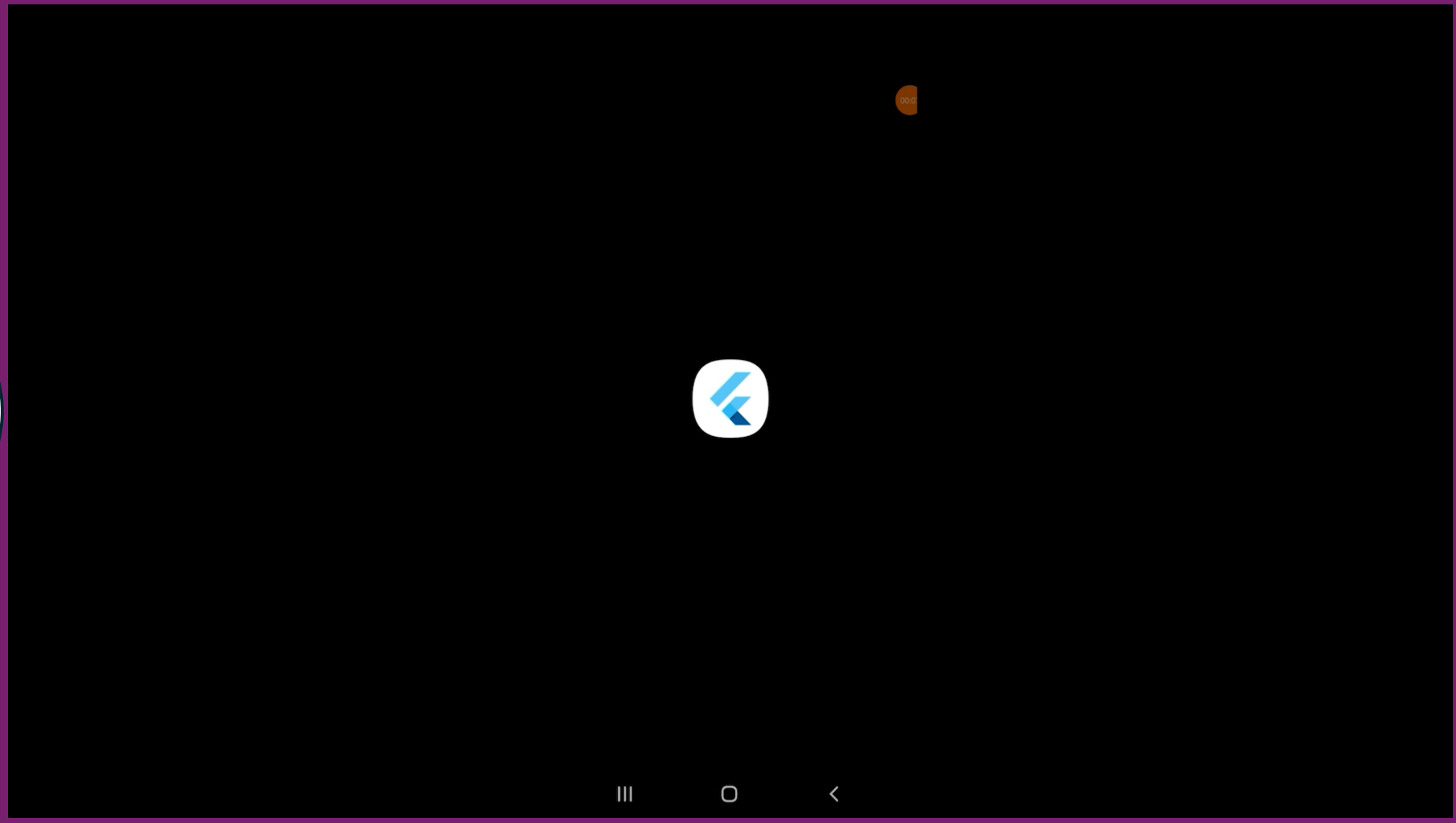
3. To prepare the filtered data to be in the Input-Output format, we combined the context (subtitles) with the questions from the QA pairs to be the ‘Input’. Each question was prefixed with "Q:" to clearly mark it, and combined the answers from the QA pairs, each prefixed with "A:" to be the ‘Output’.

Fine-Tuning of Gemini

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." The performance of the model will be shown in the demo

Running Demo


DEMO



Thank You