VIDEO ADS TRIAL TASK

1. Objective:

The task aimed to utilize video advertisements and their corresponding text descriptions and speech captions to answer 21 binary (yes/no) questions by developing a classifier or using an LLM, multi-modal LLM, etc. and calculate agreement percentages, compute F1 score, precision, and recall against the provided ground-truth data.

2. Methodology:

a. Data Preparation:

Input Data: The notebook processes the ground truth data and predicted results for 150 video ads, each mapped to 21 questions. The data is handled in pandas Data Frames for ease of manipulation and calculation. The notebook handles data sources like the Sample csv containing the details of the video ads, ground-truth that answers the questions we need to find using the classifier, the 21 questions questionnaire, and the video data with 150 video advertisements.

```
sample_df.head()
```

	creative_data_id	creative_data_title	creative_data_description	creative_data_duration	creative_data_lifetime_spend_estimated	creative_data_lifetime_airings_cc
0	2194673	30s Kim's Discount - 2194673	Kim is going for the State Farm Drive Safe & S...	30	29789808.73	13
1	2142915	30s New Flat - 2142915	Uncomfortable with her shabby apartment and ro...	30	5423001.70	10
2	1702851	30s Most Pills Don't Finish the Job - 1702851	Flonase guesses you wouldn't accept an incompl...	30	23072716.78	8
3	1671980	30s Box Vlog - 1671980	The Progressive Box starts his own vlog as he ...	30	44909836.61	7
4	1749291	30s Scars: One Way Out [T1] - 1749291	Chevrolet owners tell stories of how their Sil...	30	3490623.10	5

Data Conversion: Both the predicted and ground truth data are encoded from yes/no answers to binary 1/0 for computational ease.

Handled Missing data and renamed the column names wherever necessary.

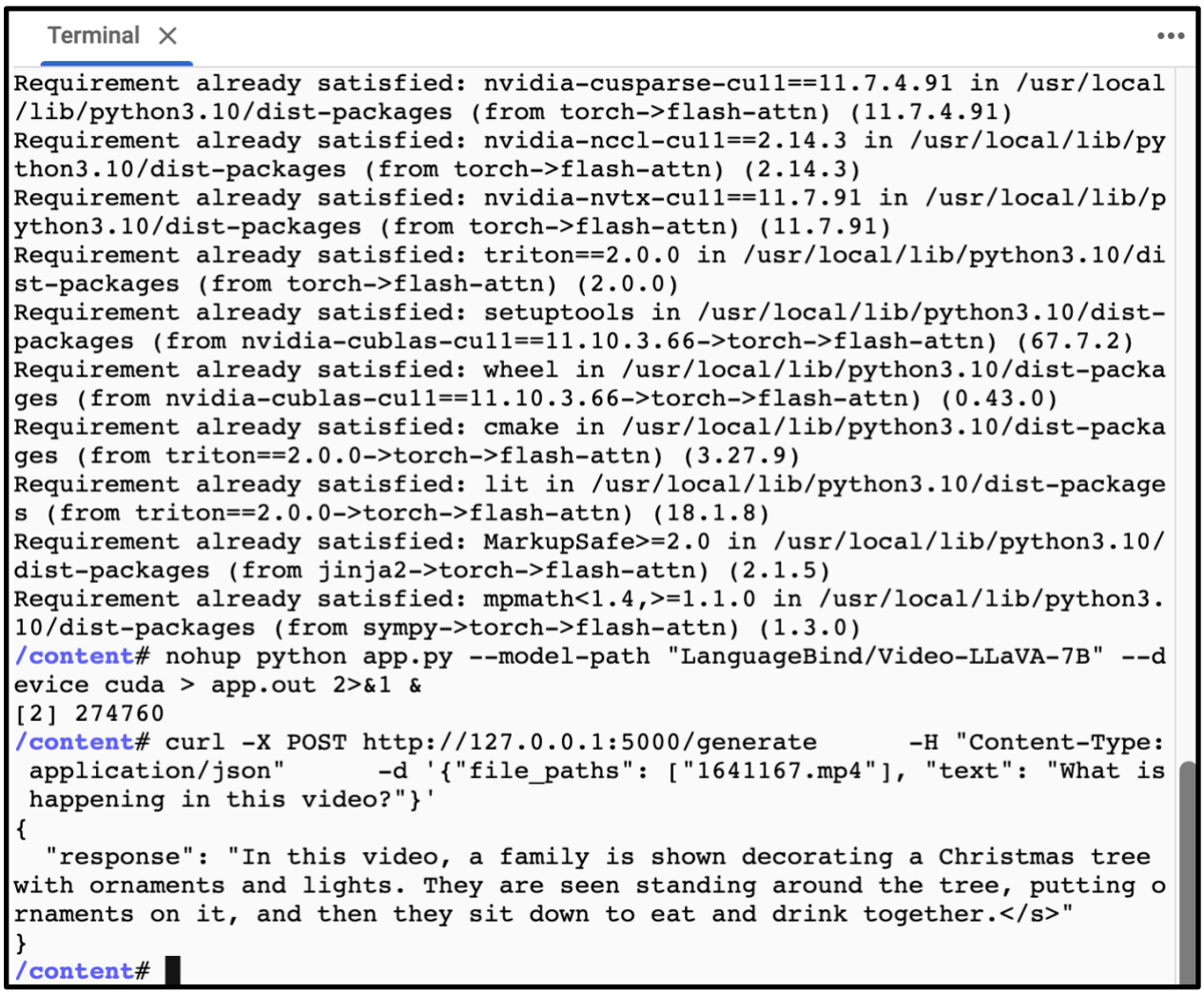
b. Classifier Configuration and Usage:

LLaVA (Large Language-and-Vision Assistant) is an end-to-end trained large multimodal model that connects a vision encoder and LLM for general-purpose visual and language understanding. This model exhibits remarkable interactive capabilities between images and videos

LLaVa LLM Setup:

- Cloned the necessary LLaVa LLM repository using Git.
- Navigated into the repository directory and installed the package along with its dependencies.
- Specialized Python packages for video and ML model handling were installed.
- Initiated the server script in the background using nohup to allow continuous operation.
- Monitored the server operation through log outputs and managed the process lifecycle using system commands.

Repository Setup:



- `pip install -e.`
- `pip install flash-attn --no-build-isolation`
- `pip install decord opencv-python git+https://github.com/facebookresearch/pytorchvideo.git@28fe037d212663c6a24f373b94cc5d478c8c1a1d`
- `nohup python app.py --model-path "LanguageBind/Video-LLaVA-7B" --device cuda > app.out 2>&1 &`

- Monitor the application output: `tail -f app.out`
- Manage the application process: Check running processes: `ps aux | grep python`

c. Prompt Engineering:

```
Processing video: 2750416.mp4
Prompt: You are an assistant tasked to watch videos. You will be given description and speech for each video.
                            You will be asked a question related to the video. You only have to answer as Yes or No
                            Here is the description: A customer at Lowe's tells an employee that he is in the middle
                            Here is the speech:  Hi. Hey. So what are you working on? I'm in kind of a yard off Rob B
                            Here is your question: Is the ad visually pleasing? Guidelines: Say yes if the ad is aest


Processing video: 2750416.mp4
Prompt: You are an assistant tasked to watch videos. You will be given description and speech for each video.
                            You will be asked a question related to the video. You only have to answer as Yes or No
                            Here is the description: A customer at Lowe's tells an employee that he is in the middle
                            Here is the speech:  Hi. Hey. So what are you working on? I'm in kind of a yard off Rob B
                            Here is your question: Does the ad have cute elements? Guidelines: Say yes only if the ad


Processing video: 2750416.mp4
['2750416', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'yes', 'y
Processing video: 1742915.mp4
Prompt: You are an assistant tasked to watch videos. You will be given description and speech for each video.
                            You will be asked a question related to the video. You only have to answer as Yes or No
                            Here is the description: Travelling can have a harsh impact on your health and well-being
                            Here is the speech:  The more you travel, the more your well being can get left behind. F
```

Prompt Design: Developed and refined textual prompts to effectively interact with the LLM, aiming to improve clarity and specificity in the responses.

Combined the columns from sample.csv to get description and speech information of the video, added the guidelines for answering each of the 21 questions.

Iterative Testing: Conducted iterative testing to fine-tune prompts based on initial performance outcomes, focusing on reducing ambiguities and enhancing the relevance of the questions to video content.

Contextual Embedding: Enhanced prompts by embedding contextual cues from the video and textual data, helping the LLM better understand and interpret the content.

Feedback Incorporation: Adjusted prompts based on observed LLM outputs to optimize question framing and maximize response accuracy.

d. Performance Metrics Calculation:

Metric Functions: Utilizes scikit-learn library functions to compute accuracy (as agreement percentage), F1 score, precision, and recall for each question based on the binary data.

Iterative Analysis: The script iterates through each of the 21 questions for each video under the sample folder, applying the metric calculations and printing the results.

e. Error Handling:

Incorporates warnings and error handling particularly for cases where certain metrics like recall may not be well-defined due to the absence of positive samples in the ground truth. Addresses potential issues where some metrics might be undefined due to data imbalances or absence of certain classes in the ground truth.

3. Results:

The Final Evaluation Metrics for the 21 questions for each of the 150 videos are as follows:

```
Metrics for question_1:
  Agreement Percentage: 36.00%
  F1 Score: 0.44
  Precision: 0.29
  Recall: 0.97

Metrics for question_2:
  Agreement Percentage: 45.33%
  F1 Score: 0.61
  Precision: 0.45
  Recall: 0.98

Metrics for question_3:
  Agreement Percentage: 40.00%
  F1 Score: 0.54
  Precision: 0.37
  Recall: 0.98

Metrics for question_4:
  Agreement Percentage: 26.00%
  F1 Score: 0.40
  Precision: 0.25
  Recall: 1.00

Metrics for question_5:
  Agreement Percentage: 42.67%
  F1 Score: 0.59
  Precision: 0.42
  Recall: 1.00

Metrics for question_6:
  Agreement Percentage: 20.00%
  F1 Score: 0.29
  Precision: 0.17
```

```
  Recall: 1.00

Metrics for question_7:
  Agreement Percentage: 20.00%
  F1 Score: 0.15
  Precision: 0.08
  Recall: 1.00

Metrics for question_8:
  Agreement Percentage: 83.33%
  F1 Score: 0.91
  Precision: 0.85
  Recall: 0.98

Metrics for question_9:
  Agreement Percentage: 10.67%
  F1 Score: 0.00
  Precision: 0.00
  Recall: 0.00

Metrics for question_10:
  Agreement Percentage: 83.33%
  F1 Score: 0.91
  Precision: 0.84
  Recall: 0.98

Metrics for question_11:
  Agreement Percentage: 83.33%
  F1 Score: 0.91
  Precision: 0.85
  Recall: 0.98

Metrics for question_12:
  Agreement Percentage: 78.00%
  F1 Score: 0.88
  Precision: 0.79
  Recall: 0.98

Metrics for question_13:
  Agreement Percentage: 81.33%
  F1 Score: 0.90
  Precision: 0.82
  Recall: 0.98

Metrics for question_14:
  Agreement Percentage: 23.33%
  F1 Score: 0.38
  Precision: 0.23
  Recall: 1.00

Metrics for question_15:
  Agreement Percentage: 16.67%
  F1 Score: 0.22
  Precision: 0.13
```

```
  Recall: 0.95

Metrics for question_16:
  Agreement Percentage: 45.33%
  F1 Score: 0.62
  Precision: 0.45
  Recall: 1.00

Metrics for question_17:
  Agreement Percentage: 62.67%
  F1 Score: 0.77
  Precision: 0.63
  Recall: 0.99

Metrics for question_18:
  Agreement Percentage: 26.67%
  F1 Score: 0.34
  Precision: 0.20
  Recall: 0.97

Metrics for question_19:
  Agreement Percentage: 55.33%
  F1 Score: 0.71
  Precision: 0.55
  Recall: 1.00

Metrics for question_20:
  Agreement Percentage: 68.67%
  F1 Score: 0.81
  Precision: 0.69
  Recall: 1.00

Metrics for question_21:
  Agreement Percentage: 27.33%
  F1 Score: 0.37
  Precision: 0.23
  Recall: 0.94
```

The resulting dataframe with the answers as provided by the classifier is as follows:

```
results_df
```

	creative_data_id	question_1	question_2	question_3	question_4	question_5	question_6	question_7	q
0	2750416	no	yes	yes	yes	yes	yes	yes	
1	1742915	yes	yes	yes	yes	yes	yes	yes	
2	2194673	yes	yes	yes	yes	yes	yes	yes	
3	2385082	yes	yes	yes	yes	yes	yes	yes	
4	2808275	yes	yes	yes	yes	yes	yes	yes	
...	...	...	...	...	...	...	...	...	
145	1625396	yes	yes	yes	yes	yes	yes	no	
146	2238004	yes	yes	yes	yes	yes	yes	no	
147	2764983	yes	yes	yes	yes	yes	yes	no	
148	3351059	yes	yes	yes	yes	yes	yes	yes	
149	2650588	no	yes	yes	yes	yes	yes	yes	

150 rows × 22 columns

4. Insights and Observations:

We can draw a few insights and conclusions from the above results:

1. Variable Performance: The script's output indicates significant variability in the performance across questions, highlighting some areas where the classifier excels and others where it struggles.

2. Recall and Precision Discrepancies: High recall across many questions suggests that the classifier can identify most positives but struggles with precision—indicating a tendency to over-predict positives.

3. Challenges with Ambiguity: Lower performance metrics in some questions may point towards inherent ambiguities in the video content or subjective interpretation required by the questions.

5. Recommendations for Improvement:

Refinement of Features and Data Handling: We could enhance the feature selection and engineering process to improve classifier sensitivity and specificity and also consider techniques to balance the dataset where some classes may be underrepresented.

Classifier Optimization: Optimizing the classifier through parameter fine tuning and possibly exploring different modeling techniques that might handle the nuances of the dataset better.

Focused Analysis on Troublesome Questions: Conduct a deeper dive into the questions where performance lags, by maybe incorporating more textual data and clearer prompts that could clarify ambiguous cases.

6. BONUS QUESTIONS:

1. Analyzing Classifier Performance with Certain Videos

Inconsistencies in Classifier Responses:

- Complexity of Videos: Videos that feature rapid scene changes, multiple overlays of text and visuals, or poor audio quality might confuse the classifier, leading to inconsistent answers.

- Ambiguity in Content: Ads that are abstract or highly artistic might lack clear verbal or visual cues that the classifier relies on to answer questions, especially those requiring subjective interpretation (e.g., emotional impact).

- Cultural and Contextual Nuances: Some ads might incorporate cultural references or humor that are location-specific and might not be well understood by a classifier not trained on such diverse datasets.

Technical Limitations:

- Feature Extraction Issues: Inadequate or ineffective feature extraction from videos can result in the classifier missing key information necessary for answering specific questions accurately.

- Model Generalization: If the training data isn't sufficiently diverse, the classifier might not generalize well to new, unseen examples, particularly those that differ significantly from the training set in style or content.

2. In-Depth Analysis of Human Coders' Responses and Performance of Classifier

- Agreement Levels: We can compare the agreement levels between the human coders and the classifier. High variability in human responses could indicate questions that are subjective or ambiguous, which also challenge the classifier.

- Consistency Across Videos: Identify if certain types of videos consistently receive mixed reviews from humans, which could also correlate with poor classifier performance.

- Questions that ask about emotional impact or the creativity of the ad often have lower agreement rates among human coders and between coders and the classifier. These are inherently subjective and can vary widely in interpretation.

- Precision vs. Recall: Analyze cases where the classifier had high recall but low precision (or vice versa) to identify if it's overly cautious or overly aggressive in its predictions, compared to human coders.

3. Observed Patterns or Anomalies in the Data and Potential Causes

Patterns:

- High Recall, Low Precision: Common across several questions, suggesting the classifier might be erring on the side of predicting 'yes' too frequently, possibly due to imbalanced training data where 'yes' responses were more common.
- Low Performance on Some Questions: If specific questions consistently show poor metrics, it could indicate either poor training data quality for those questions or that the questions themselves are not well framed.

Anomalies:

- Outlier Videos: Videos that significantly deviate from the normal ones in terms of length, style, or content and that consistently confuse both human coders and the classifier.
- Unusual Classifier Behavior: Any unexpected spikes or drops in classifier performance across different batches of the data could indicate underlying issues with data processing or model application.