EXECUTIVE SUMMARY FOR VIDEO ANALYSIS TASK

This project involves the analysis of video and textual data to classify advertisements based on specific questions. The main steps include data loading, preprocessing, feature extraction, model training, and evaluation. The goal is to predict answers to predefined questions for each advertisement and evaluate the performance of the models used.

Steps Done for The Analysis

1. **Data Loading**:
    - Loaded textual data from a CSV file.
    - Loaded video paths from a specified directory.
    - Loaded ground truth data from another CSV file.
2. **Data Preprocessing**:
    - Filled missing values in the ground truth data with a default value ('No').
    - Dropped columns with any `None` values to ensure data integrity.
    - Defined question columns by trimming spaces for consistency.
3. **Feature Extraction**:
    - **Textual Features**: Used BERT tokenizer and model to extract features from textual data.
    - **Video Features**: Used pre-trained ResNet-50 to extract features from video frames.
    - Combined textual and video features to create a comprehensive feature set.
4. **Model Training**:
    - Split the dataset into training and testing sets.
    - Defined parameter grids and performed hyperparameter tuning using GridSearchCV.
    - Trained classifiers: Gradient Boosting.
5. **Prediction**:
    - Predicted answers for each creative_data_id using the trained classifiers.
    - Ensured each creative_data_id was predicted only once and compiled the predictions.
6. **Evaluation**:
    - Calculated precision, recall, F1 score, and agreement percentage for each question.
    - Printed the evaluation metrics for insights.

Key Findings

The precision and recall scores varied across different questions, indicating the models' varying performance on different aspects of the advertisements.

General Observations

1. **Precision, Recall, and F1 Scores**:
    - The metrics for precision, recall, and F1 scores vary significantly across different questions, indicating variability in model performance.
    - High precision and recall scores suggest the model performs well on certain questions, while lower scores indicate challenges in accurate classification.

2. **Agreement Percentage**:
    - The agreement percentage is generally higher for questions related to explicit visual or textual elements (e.g., "Is there mention of something free?", "Does the ad show the brand logo?").
    - Lower agreement percentages suggest that some questions may be more ambiguous or difficult for the model to interpret.

Specific Questions and Insights

1. **Is there a call to go online (e.g., shop online, visit the Web)?**
    - **Precision**: 0.48, **Recall**: 0.49, **F1 Score**: 0.44, **Agreement Percentage**: 60%
    - **Insight**: The model has moderate performance, indicating it can detect calls to go online but with some uncertainty. Variability in how these calls are presented in different ads may contribute to lower scores.
2. **Is there online contact information provided (e.g., URL, website)?**
    - **Precision**: 0.47, **Recall**: 0.47, **F1 Score**: 0.47, **Agreement Percentage**: 46.67%
    - **Insight**: The model struggles with this question, likely due to the diverse ways online contact information can be embedded in ads.
3. **Is there a visual or verbal call to purchase (e.g., buy now, order now)?**
    - **Precision**: 0.56, **Recall**: 0.55, **F1 Score**: 0.51, **Agreement Percentage**: 53.33%
    - **Insight**: The performance is slightly better here, suggesting that explicit calls to purchase are more recognizable to the model.
4. **Does the ad portray a sense of urgency to act (e.g., buy before sales ends, order before ends)?**
    - **Precision**: 0.38, **Recall**: 0.33, **F1 Score**: 0.35, **Agreement Percentage**: 43.33%
    - **Insight**: This is one of the lowest-performing questions. Urgency cues can be subtle and context-dependent, making them hard to detect consistently.
5. **Is there an incentive to buy (e.g., a discount, a coupon, a sale or "limited time offer")?**
    - **Precision**: 0.61, **Recall**: 0.58, **F1 Score**: 0.57, **Agreement Percentage**: 60%
    - **Insight**: The model performs relatively well, indicating that incentives are clearer and more consistently presented in ads.
6. **Is there offline contact information provided (e.g., phone, mail, store location)?**
    - **Precision**: 0.62, **Recall**: 0.63, **F1 Score**: 0.62, **Agreement Percentage**: 66.67%
    - **Insight**: High scores suggest that offline contact information is easier for the model to detect, likely due to its explicit and consistent format.
7. **Is there mention of something free?**
    - **Precision**: 0.40, **Recall**: 0.50, **F1 Score**: 0.44, **Agreement Percentage**: 80%
    - **Insight**: The high agreement percentage but low precision indicates that while the model can often detect mentions of free offers, it also frequently misclassifies other statements as such.
8. **Does the ad show the brand (logo, brand name) or trademark multiple times?**
    - **Precision**: 0.65, **Recall**: 0.60, **F1 Score**: 0.61, **Agreement Percentage**: 76.67%
    - **Insight**: High performance on this question suggests that brand logos and trademarks are distinct and consistently recognizable features.

9. **Is the ad intended to be funny?**
    - **Precision**: 0.66, **Recall**: 0.62, **F1 Score**: 0.60, **Agreement Percentage**: 63.33%
    - **Insight**: Humor is more straightforward to identify, possibly due to common visual and auditory cues used in comedic ads.

Anomalies and Potential Causes

1. **Inconsistent Performance on Emotional and Contextual Questions**:
    - Questions about emotions, story arcs, and creativity have lower precision and recall scores.
    - **Cause**: These features are more subjective and context-dependent, making them harder for a model to learn and predict accurately.
2. **High Agreement but Low Precision/Recall**:
    - For questions like "Is there mention of something free?" the high agreement percentage paired with low precision suggests that while the model detects the feature often, it also produces many false positives.
    - **Cause**: This may be due to ambiguous phrasing or varied presentation styles that confuse the model.

Conclusion

By addressing the observed patterns and anomalies through improved data quality, enhanced feature extraction, balanced training data, and robust model training techniques, the overall performance of the classifier can be significantly improved. Implementing a hybrid approach that combines human judgment and machine learning can also help mitigate some of the challenges and leverage the strengths of both methods.

BONUS QUESTION:

1. Why Certain Videos Might Not Work Well with the Classifier or Some Questions Yield Inconsistent Answers

1. **Feature Extraction Challenges**:
    - **Textual Data**: If the text within the video is ambiguous, uses slang, or is contextually dependent, it can be difficult for the BERT model to extract meaningful features. This can result in poor classifier performance for questions that rely on textual data.
    - **Video Data**: Variations in video quality, such as low resolution, poor lighting, or excessive noise, can hinder the ResNet-50 model's ability to extract relevant features. Videos with quick cuts or rapid scene changes might also miss key frames, leading to incomplete data for the classifier.
2. **Ambiguity and Subjectivity**:
    - **Question Ambiguity**: Some questions might be inherently ambiguous or open to interpretation. For example, "Does the ad portray a sense of urgency to act?" can be subjective and interpreted differently by various annotators, leading to inconsistent answers.

    - **Subtle Features**: Subtle cues like small brand logos or implicit calls to action can be challenging to detect consistently. These features might be missed or misinterpreted by the classifier, leading to lower precision and recall.
3. **Training Data Quality**:
    - **Inconsistent Annotations**: If the ground truth data used for training the classifier contains inconsistent annotations, possibly due to human error, the classifier's performance will be affected. Variability in human coder responses can introduce noise into the dataset, leading to less reliable model predictions.
4. **Class Imbalance**:
    - **Underrepresented Classes**: If certain categories (e.g., urgency cues, incentives) are underrepresented in the training data, the classifier might not learn to recognize these features effectively. This results in poor performance for those categories.
5. **Overfitting**:
    - **Overfitting to Training Data**: The classifier might perform well on the training data but poorly on unseen test data if it has overfitted to the specific examples in the training set. This can lead to high variability in performance across different videos and questions.

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

1. **Human Coders' Responses**:
    - **Consistency**: Human coders may show variability in their annotations due to fatigue, differing interpretations, or lack of clear guidelines. Intra-coder consistency (consistency of the same coder over time) and inter-coder consistency (consistency between different coders) are crucial for reliable annotations.
    - **Accuracy**: Coders might miss subtle cues or context-dependent elements, leading to errors in the ground truth data. Detailed and consistent guidelines can help mitigate these issues.
    - **Bias and Subjectivity**: Personal biases and subjective interpretations can affect the consistency and accuracy of human annotations. This is particularly true for questions that require subjective judgment, such as those related to emotions or creativity.
2. **Classifier Performance**:
    - **Precision, Recall, and F1 Scores**: The classifier shows varying performance across different questions. High precision indicates fewer false positives, while high recall indicates fewer false negatives. The F1 score balances both metrics. **Example**: For the question "Is there an incentive to buy?", the classifier shows high precision (0.61) and recall (0.58), indicating good performance. In contrast, for "Does the ad portray a sense of urgency?", the precision (0.38) and recall (0.33) are much lower, highlighting challenges in detecting urgency cues.
    - **Agreement Percentage**: Higher agreement percentages suggest that the classifier's predictions align well with the ground truth. Lower percentages indicate areas where the model struggles. **Example**: The agreement percentage for "Is there mention of something free?" is 80%, showing high alignment with the ground truth. However, for "Does the ad portray a sense of urgency?", the agreement percentage is only 43.33%, indicating significant inconsistencies.

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

1. **Patterns**:
    - **Higher Performance on Explicit Features**: The classifier performs better on questions related to explicit visual or textual elements, such as brand logos or mentions of discounts. These features are more consistently presented and easier to detect.
        - **Example**: Questions like "Does the ad show the brand logo?" have high precision (0.65) and recall (0.60), indicating strong performance on explicit features.
    - **Lower Performance on Subjective or Context-Dependent Features**: Questions requiring subjective judgment or contextual understanding, such as emotions or urgency, show lower performance. **Example**: The question "Is the ad intended to affect the viewer emotionally?" has lower precision (0.54) and recall (0.56), indicating difficulties in detecting emotional content.
2. **Anomalies**:
    - **Low Precision with High Agreement Percentage**: Some questions show high agreement percentages but low precision, indicating that while the classifier detects the feature often, it also produces many false positives. **Example**: For "Is there mention of something free?", the high agreement percentage (80%) coupled with lower precision (0.40) suggests frequent false positives.
    - **High Variability in Subjective Questions**: Questions related to emotions, creativity, and story arcs exhibit high variability, reflecting the challenges in modeling subjective and complex features.
        - **Example**: The question "Does the ad have a story arc?" has lower precision (0.44) and recall (0.42), highlighting inconsistencies in detecting narrative structures.