CLARITY Shared Task: Exploratory Data Analysis Report

1 Introduction & Motivation

Political discourse is inherently rich in ambiguity, particularly in high-stakes environments such as televised interviews, debates, and press conferences. Politicians often rely on evasive or equivocal communication strategies that allow them to address a topic without providing a direct or verifiable answer. Such strategies can leave audiences with multiple interpretations of a statement and create the false impression that a question has been adequately addressed. This phenomenon—commonly referred to as *equivocation* or *evasion*—has been widely examined in political communication research, yet it remains comparatively underexplored in computational linguistics.

Empirical evidence demonstrates the strategic and pervasive nature of this behavior. For instance, Bull (2003) reports that politicians provide clear responses to only 39–46% of questions in televised interviews, whereas non-politicians offer substantially higher rates of direct answers (70–89%). These differences highlight an important gap: despite extensive theoretical literature in political science, there is limited computational work capable of automatically detecting ambiguous or evasive responses at scale.

The CLARITY shared task aims to address this gap by proposing an automated approach for identifying and categorizing response ambiguity in question–answer (QA) pairs extracted from presidential interviews. The task is grounded in well-established theories of political equivocation and leverages contemporary advances in natural language processing and language modeling. A key contribution of this task is its hierarchical framework, derived from Thomas et al. (2024), which introduces:

1. A high-level classification distinguishing between clear and ambiguous responses; and
2. A fine-grained taxonomy of nine evasion techniques commonly used in political discourse.

This two-level taxonomy enables a richer and more systematic analysis of political communication. Preliminary findings suggest that modeling the two levels jointly can enhance classification performance, providing both coarse- and fine-grained insight into how ambiguity manifests in political speech.

By offering a standardized dataset, annotation scheme, and evaluation setup, the CLARITY task aims to support large-scale, data-driven research on political interviews and public communication. It invites participation from researchers across natural language processing, discourse analysis, political science, fact-checking, and dialogue systems. Ultimately, this work contributes toward developing tools that promote transparency, accountability, and deeper understanding of the linguistic strategies deployed in political communication.

2 Dataset Overview

The CLARITY dataset consists of question–answer pairs extracted from presidential interviews. Each sample includes:

- Interview question text
- Interview answer text
- GPT-3.5 summary and prediction fields
- High-level clarity labels
- Fine-grained evasion technique labels

The dataset supports multimodal extensions (text, audio, and potentially alignment metadata), although this report focuses primarily on the text-based components.

3 Team Members & Work Division

The CLARITY shared task EDA project was conducted collaboratively by four team members, each responsible for different aspects of data exploration and analysis:

- **Muhammad Huzefa (Member 1, Qalam ID: 508819):** Responsible for overall LaTeX report preparation, understanding data analysis concepts, and technical writing. Tasks include deciding which plots to include in the report and interpreting results, reviewing methods like token length distribution, n-grams, vocabulary size, label distributions, and addressing dataset challenges such as imbalance and noise. Implementation is not required; the focus is on summarizing and documenting insights.
- **Abdullah Rashid (Member 2, Qalam ID: 522694):** Conducted initial EDA using Python and Matplotlib, fetching data from HuggingFace. Focused on token length analysis for questions and answers and generating corresponding histograms.
- **Rafay Ahmed (Member 3, Qalam ID: 511295):** Downloaded the dataset from HuggingFace and performed second-stage EDA. Generated plots and analyses for frequent n-grams, top phrases across answers, word counts, and vocabulary size to provide insights into textual patterns and dataset richness.
- **Amna Imran (Member 4, Qalam ID: 516513):** Also worked with HuggingFace data, focusing on label distribution and dataset challenges. Explored noise, imbalance, sentiment analysis, and multimodal alignment. Produced plots to visualize label distribution, data inconsistencies, and other quality issues.

4 Exploratory Data Analysis (EDA)

4.1 Token Length Analysis

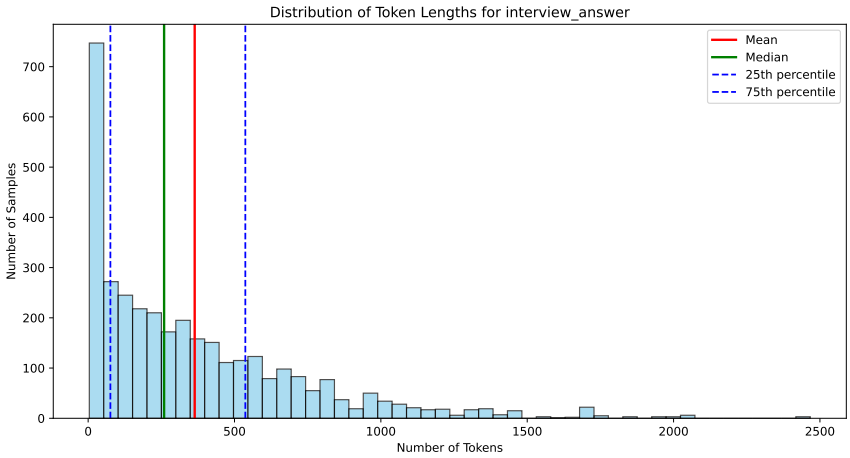

Figure 1: Token Length Distribution: Interview Answers. Most answers are moderately long, but there is a long tail of very verbose responses, indicating high variability in answer length.

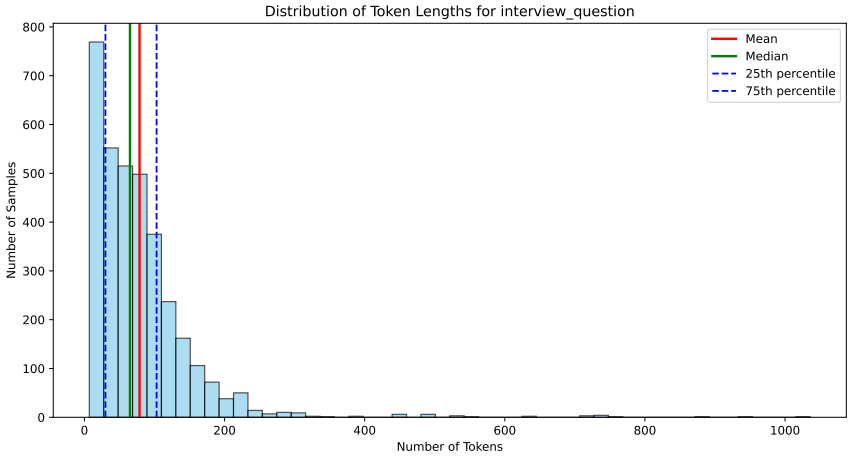

Figure 2: Token Length Distribution: Interview Questions. Questions tend to be shorter than answers, with most falling in the 5–25 token range.

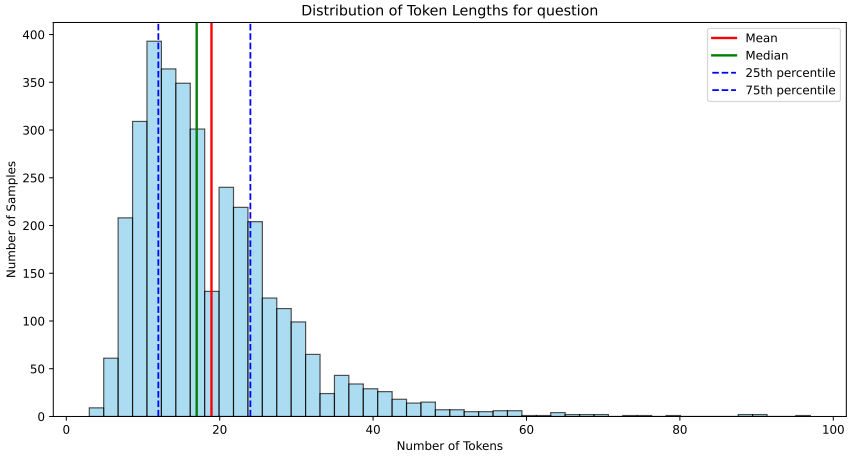

Figure 3: Token Length Distribution: General Questions. Most questions are concise, though a few unusually long questions exist.

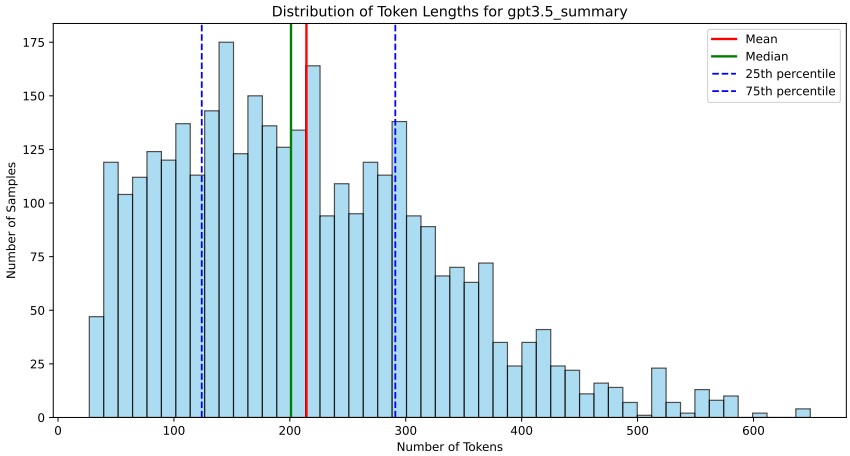

Figure 4: Token Length Distribution: GPT-3.5 Summaries. Summaries are generally shorter than full answers, showing consistency in length.

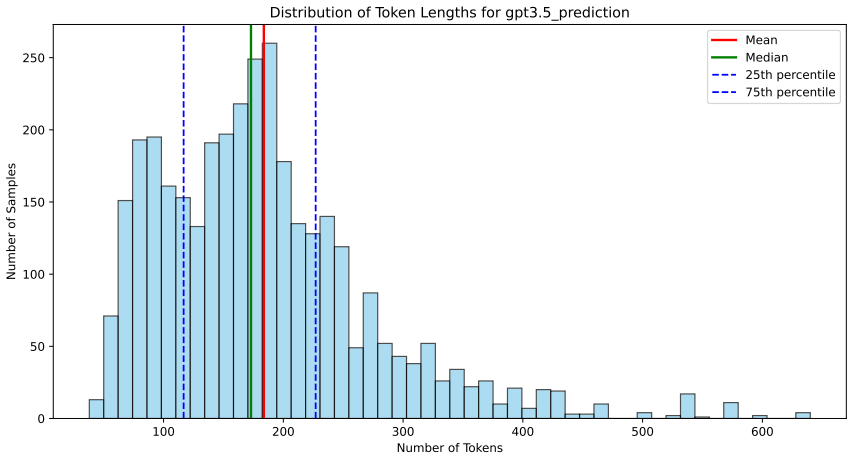

Figure 5: Token Length Distribution: GPT-3.5 Predictions. Predictions are usually short phrases or single words, resulting in a narrow length distribution.

4.2 Label Distribution

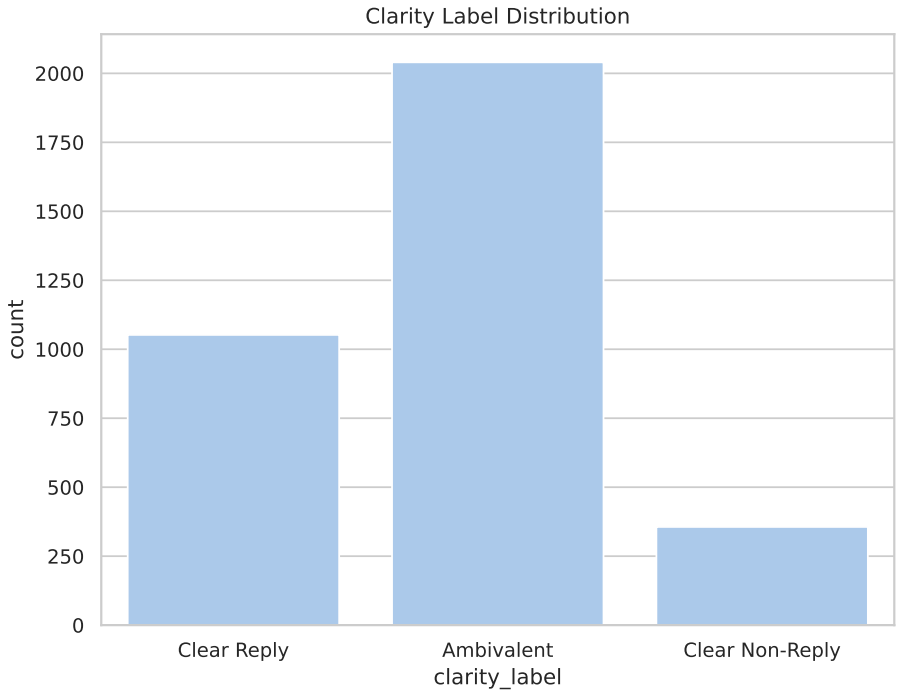

Figure 6: Clarity Label Distribution. The dataset is slightly imbalanced toward clear responses, though ambiguous answers are also frequent.

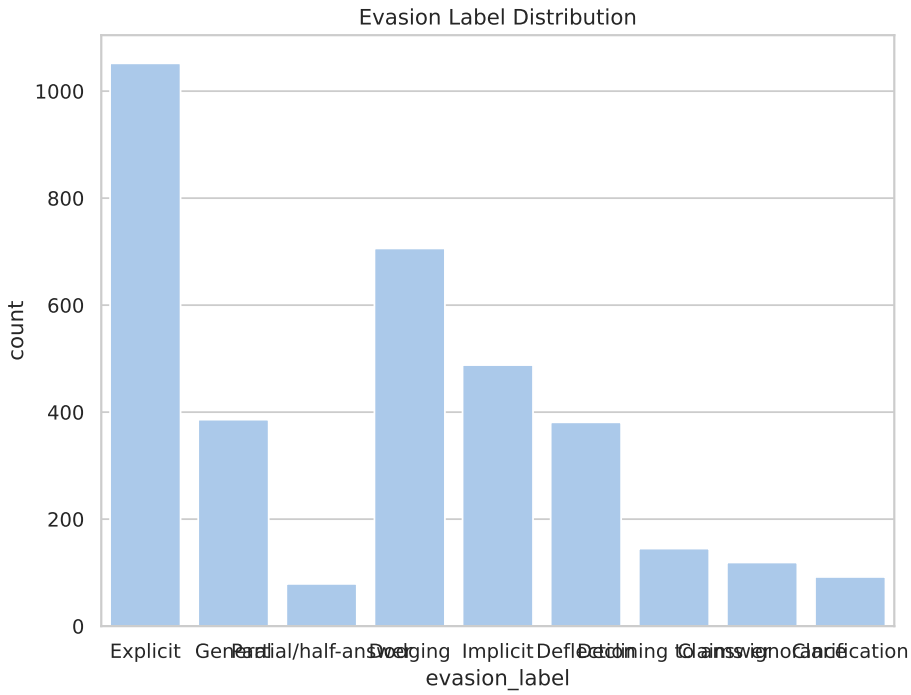

Figure 7: Evasion Technique Distribution. Several categories have very few examples, highlighting challenges for rare-class prediction.

4.3 Correlation and Cross-Modal Patterns

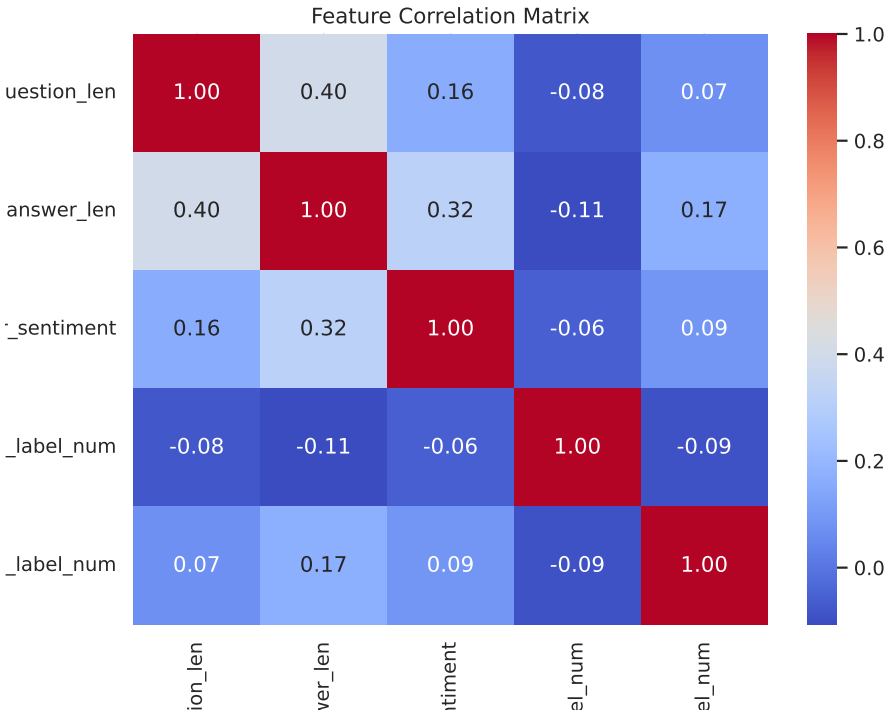

Figure 8: Feature Correlation Matrix. Weak correlations suggest that token length or sentiment alone cannot reliably predict clarity.

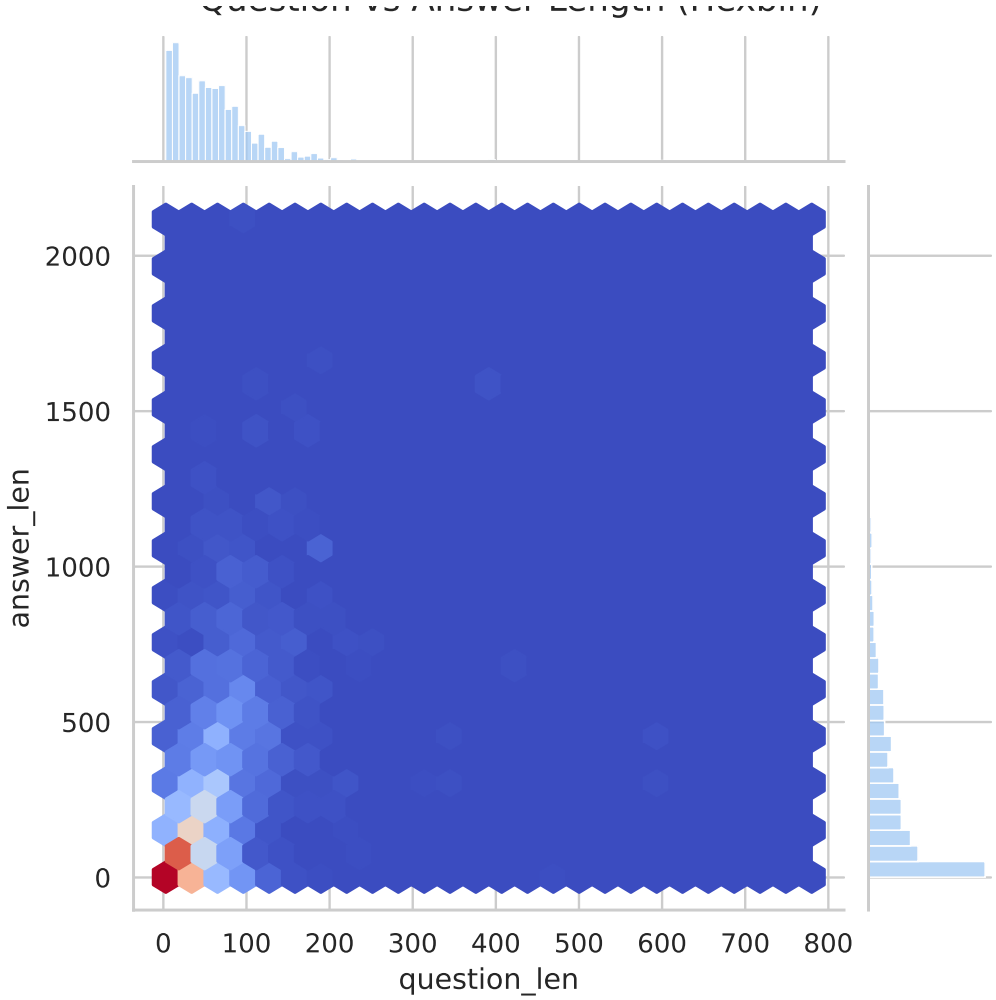

Figure 9: Hexbin Plot: Question vs. Answer Lengths. Most answers are longer than their corresponding questions; dense clusters show typical lengths.

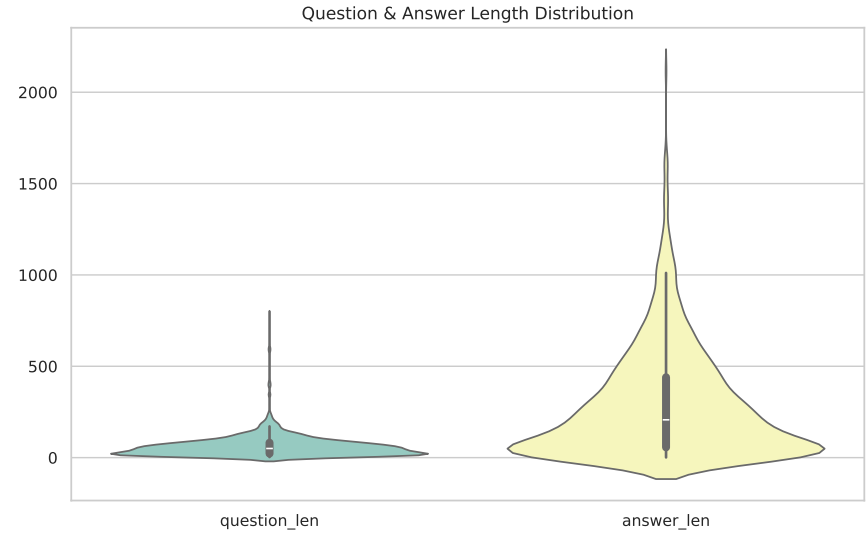

Figure 10: Joint Distribution: Q/A Length Relationship. Scatter plot highlights that very long answers often correspond to average-length questions.

4.4 Sentiment and Missing Data Analysis

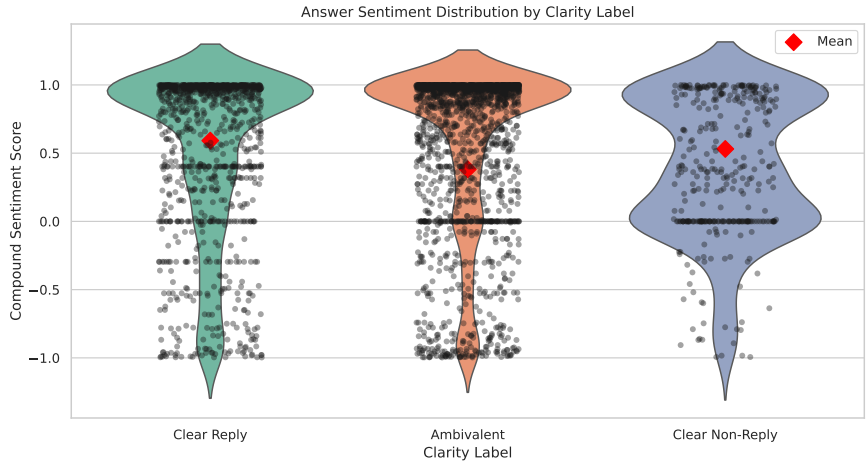

Figure 11: Sentiment Distribution by Clarity Label. Sentiment differs slightly by clarity, but alone is not a strong predictor of ambiguity.

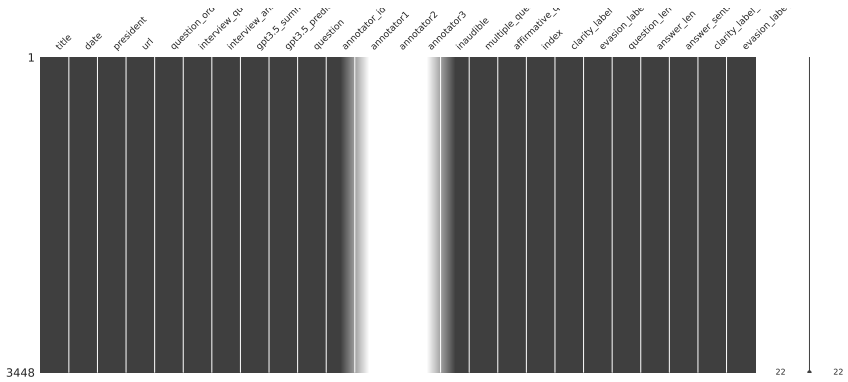

Figure 12: Missing Data Heatmap. Missing GPT outputs and annotations may require preprocessing or imputation for modeling.

5 Problem Formulation

The primary task is a supervised multi-class classification problem:

- **Task 1:** Binary classification of response *clarity* (Clear vs. Ambiguous)
- **Task 2:** Fine-grained multi-class classification across 9 *evasion types*

Inputs:

$$x = (\text{question}, \text{answer})$$

Outputs:

$$y_{\text{clarity}} \in \{\text{Clear}, \text{Ambiguous}\}, \quad y_{\text{evasion}} \in \{1, \ldots, 9\}$$

6 Evaluation Metrics

- Accuracy
- Macro F1-score (handles class imbalance)
- Confusion matrix inspection

Macro-F1 is used as the main evaluation metric for both tasks.

7 Conclusion

This EDA highlights key characteristics of the dataset, including significant variability in token lengths, imbalanced labels, sentiment differences, and missing data patterns. These insights guide preprocessing decisions and model selection for downstream experiments in response clarity classification and evasion detection.