Sentiment Syndicate at SemEval-2026 Task 6: Reframing Political Question–Answer Interactions via Natural Language Inference for Clarity Level Classification

Rafi Rafsan
Rajshahi University of Engineering & Technology

Abstract

This paper presents the Sentiment Syndicate team's submission to SemEval-2026 Task 6, Subtask 1 (CLARITY: Unmasking Political Question Evasions), which focuses on classifying the clarity level of political question–answer interactions. We investigate three modeling strategies: (1) fine-tuning a RoBERTa-based classifier, (2) reformulating the task as a Natural Language Inference (NLI) problem, and (3) leveraging large language models (LLMs) for classification. All approaches are evaluated using macro F1-score on the official dataset. Experimental results demonstrate that the NLI-based formulation outperforms the other strategies, highlighting the effectiveness of modeling semantic alignment between questions and answers. Our best-performing system achieves an F1-score of 0.67 on the test set. [1]

1 Introduction

In the era of rapid mass information dissemination, the ability to critically evaluate the quality and reliability of information has become increasingly important. Political discourse, in particular, often includes evasive or ambiguous responses to direct questions, making it difficult for citizens to interpret intent and make informed decisions (Bull and Mayer, 1993; Rasiah, 2010). The prevalence of unclear or indirect answers highlights the need for computational models that can automatically assess the clarity of responses in political question–answer interactions (Thomas et al., 2024). SemEval-2026 Task 6 (Thomas et al., 2026) directly addresses this challenge. In this work, we focus on Subtask 1, which formulates clarity assessment as a classification problem. Given a political question and its corresponding answer, the objective is to predict the clarity level of the response. The dataset consists of annotated question–answer pairs collected from interviews of the presidents of USA, each labeled according to its clarity category (Thomas et al., 2024). This task is particularly challenging because clarity is not determined solely by topical relevance. Instead, it requires modeling nuanced discourse phenomena such as evasiveness, ambiguity, indirectness, and partial answering (Bull and Mayer, 1993). Successfully distinguishing between clear replies, ambivalent responses, and clear non-replies demands a deep understanding of semantic alignment between the question and the answer. To address this problem, we explore three modeling strategies spanning both traditional fine-tuning approaches and LLMs. Specifically, we investigate: (1) fine-tuning a RoBERTa-based classifier (Liu et al., 2019), (2) reformulating the task as a Natural Language Inference (NLI) problem to explicitly model semantic relationships between questions and answers (Yin et al., 2019; Schick and Schütze, 2021), and (3) leveraging LLMs for classification through advanced prompting techniques (Wei et al., 2022; Kojima et al., 2022). We evaluate all approaches using macro F1-score and compare their effectiveness on the official dataset provided by the shared task organizers (Thomas et al., 2024).

2 Background

2.1 Related Work

Computational analysis of political discourse has gained increasing attention, particularly in the study of misinformation, stance detection, and argumentative structures (Preotiuc-Pietro et al., 2017). Prior work has investigated evasive communication strategies in political interviews and debates, aiming to identify indirect answers, topic shifts, and rhetorical avoidance. Bull and Mayer (1993) and Rasiah (2010) established foundational frameworks for analyzing how politicians evade questions through various linguistic maneuvers. Early approaches relied on rule-based systems and

[1] Our code is available at https://github.com/bit-wander/system-syndicate.git

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2779–2786

discourse-level linguistic features, but with the advancement of deep learning, supervised models using contextualized embeddings have significantly improved the ability to capture semantic relationships (Pak and et al., 2024).

Transformer-based architectures such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have become the standard backbone for such tasks. For instances involving long textual inputs, models such as Longformer (Beltagy et al., 2020) extend the transformer architecture to efficiently process extended sequences using sparse attention mechanisms. These are particularly suitable for political data where answers can be lengthy and contain complex rhetorical structures.

Natural Language Inference (NLI) has also emerged as a powerful framework for modeling semantic relationships between paired texts. Traditionally, NLI predicts entailment, contradiction, or neutrality between a premise and a hypothesis (Bowman et al., 2015). Yin et al. (2019) successfully adapted the NLI formulation to various downstream tasks, showing that reframing classification problems as semantic inference tasks allows for effective zero-shot and few-shot classification. This approach was further refined by Schick and Schütze (2021), demonstrating that cloze-style questions and NLI-based modeling enable models to explicitly reason about semantic alignment.

In parallel, instruction-tuned Large Language Models (LLMs) have demonstrated strong zero-shot and few-shot capabilities across diverse NLP tasks (Kojima et al., 2022), often benefiting from techniques like Chain-of-Thought prompting (Wei et al., 2022). As discussed by Ziems et al. (2024), comparing LLM-based classification with supervised transformer fine-tuning provides insight into the relative effectiveness of these paradigms for computational social science.

2.2 Task Description and Dataset

The dataset used in this study was released as part of the SemEval-2026 Task 6 shared task (Thomas et al., 2026), based on the QEvasion dataset and taxonomy introduced by (Thomas et al., 2024). It consists of question–answer pairs extracted from televised presidential interviews of United States presidents conducted between 2006 and 2023. Each instance contains a political question and its corresponding full interview answer, annotated with a clarity label.

The public release of the QEvasion dataset provided 3,448 training instances and 308 test instances. For our modeling experiments, we merged these public splits into a single pool of 3,756 instances to maximize the available training signal. We then applied a stratified 80/20 split to this combined pool, yielding our final training set (3,004 instances) and an internal development set (752 instances) for model validation. Final evaluation was conducted on the official hidden test set provided via CodaBench, which contains 237 instances. Table 1 presents the distribution of instances across clarity labels in our newly defined training and development sets, as well as the hidden test set.

Label	Train	Dev	Test
Clear Reply	905	226	85
Ambivalent Reply	1796	450	117
Clear Non-Reply	303	76	35
Total	**3,004**	**752**	**237**

Table 1: Distribution of instances across clarity labels.

As noted by Thomas et al. (2024) and the task organizers Thomas et al. (2026), the dataset exhibits noticeable class imbalance. The *Ambivalent Reply* category constitutes approximately 60% of the total instances, making it the majority class across all splits. In contrast, *Clear Non-Reply* is the least represented category, accounting for roughly 10% of the data. As highlighted by Johnson and Khoshgoftaar (2019), such imbalance poses additional challenges for model training and motivates the use of strategies to mitigate bias toward majority classes.

3 System Overview

This study investigates three modeling strategies for detecting clarity and evasion in political interviews under the SemEval-2026 Task 6 framework (Thomas et al., 2026). The objective is to classify responses into three categories: *Clear Reply*, *Ambivalent*, and *Clear Non-Reply*.

Rather than relying on a single modeling paradigm, we compare (1) standard sequence classification, (2) hypothesis-augmented classification inspired by Natural Language Inference (NLI), and (3) prompting-based large language model (LLM) classification.

3.1 Direct Encoder-Based Sequence Classification (DSC)

The first strategy adopts a standard discriminative sequence classification framework using a fine-tuned transformer encoder. We employ RoBERTa-base (Liu et al., 2019) as the backbone architecture due to its strong performance on natural language understanding tasks.

Each input instance is constructed by concatenating the available textual components into a single sequence:

```
Question: {context}
Answer: {answer}
Subquestion:
{sub-question}
```

The combined sequence is passed through the encoder, and a linear classification head predicts one of the three clarity labels. Fine-tuning is performed using the Hugging Face Transformers library (Wolf et al., 2020).

To analyze the effect of input length constraints, we explicitly track whether samples are truncated during batching. This allows us to examine performance differences between truncated and non-truncated inputs.

3.2 Hypothesis-Augmented Classification (HAC)

The second strategy reformulates the task as a **hypothesis-augmented classification problem**, inspired by the structure of Natural Language Inference (NLI), but trained fully in a supervised manner. We refer to this approach as *hypothesis-augmented classification (HAC)* to distinguish it from classical NLI setups that rely on entailment-pretrained models.

Instead of directly classifying the concatenated input, we construct a pair of texts:

- **Premise:** the interview answer
- **Hypothesis:** a templated statement describing the relationship between the answer and the sub-question

Specifically, we use the following template:

"The answer clearly addresses the question: {sub-question}"

The model receives the premise–hypothesis pair as input and is trained to predict one of the three clarity labels. While this setup is structurally similar to NLI (premise–hypothesis reasoning), it does **not rely on an NLI-pretrained checkpoint**. Instead, the model learns the relationship between answers and hypotheses directly from task-specific supervision.

We implement this approach using Longformer-base (Beltagy et al., 2020) to better handle long input sequences typical of political discourse. The extended context window allows the model to capture long-range dependencies that may be truncated in standard transformer models.

This formulation encourages the model to reason about whether the answer semantically satisfies the hypothesis, which implicitly captures degrees of clarity, partial relevance, and evasion.

3.3 Generative Classification via LLM

The third strategy frames clarity detection as a reasoning task using a large language model. We evaluate a generative model (`gpt-oss:120b`) (OpenAI, 2025) in a few-shot prompting setup.

A structured prompt (see Appendix A for the full template) is designed to include:

- Definitions of clarity categories
- Illustrative examples
- Instructions to output predictions in JSON format

Each prediction includes:

- `classification`: the predicted label
- `reasoning`: a natural language justification

This approach emphasizes interpretability, as the model produces explicit reasoning alongside predictions. However, unlike supervised fine-tuning, its performance depends heavily on prompt design and example selection.

4 Experimental Setup

Our systems were implemented using the HuggingFace Transformers library (Wolf et al., 2020) and PyTorch (Paszke et al., 2019). All experiments were conducted on Google Colab utilizing GPU acceleration (NVIDIA T4).

To maximize the signal for model convergence, we consolidated the original training and development partitions provided in the `ailsntua/QEvasion` dataset into a single

pool. This combined dataset was subsequently repartitioned using an 80:20 split, resulting in 3,004 samples for training and 752 samples for internal validation.

Preprocessing involved standard tokenization via the respective pretrained tokenizers. For the hypothesis-augmented approach (referred to as NLI for formatting parity), the interview answer was treated as the *premise* and the reformulated question as the *hypothesis*. We utilized the `allenai/longformer-base-4096` model, setting the maximum sequence length to 1024 tokens. While Longformer supports up to 4096 tokens, severe VRAM constraints on the Google Colab T4 (16GB) GPU necessitated this cap to maintain a stable batch size. In the Direct Sequence Classification (DSC) baseline configurations, the question and answer were concatenated as a single input sequence with a maximum length of 512 tokens for RoBERTa-based models.

The models were fine-tuned using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $2e^{-5}$ for the NLI approach and $1e^{-5}$ for the DSC approach. Dropout from the pretrained architectures was retained to reduce overfitting. To address the class imbalance observed in political dodging, we applied a custom weighted cross-entropy loss, where class weights were computed inversely proportional to class frequencies in the training set. Training was conducted for 3 epochs, with early stopping based on the development set macro F1-score, and the best-performing checkpoint was selected for final evaluation.

5 Results

In this section, we present the quantitative evaluation of the three proposed systems on the QEvasion test set ($N = 237$). Following the official evaluation protocol, we prioritize the **Macro F1 score** as the primary metric due to the substantial class imbalance between the majority class (*Ambivalent*) and the minority class (*Clear Non-Reply*).

5.1 Quantitative Performance

Table 2 summarizes the overall system performance. Among the three strategies, the **NLI-based Longformer** model achieved the highest Macro F1 score (0.67) and Accuracy (0.68). This result indicates that combining long-context modeling with class-weighted optimization provides the most robust performance for clarity classification. However, we acknowledge that comparing the DSC and NLI pipelines inherently introduces confounding variables—specifically, simultaneous changes to the model architecture (RoBERTa vs. Longformer), context window size, and the loss function. Future work should isolate these variables through controlled ablation studies to pinpoint the exact source of performance gains.

Pipeline	Acc.	Prec.	Recall	F1
NLI	**0.68**	0.70	0.66	**0.67**
DSC	0.67	**0.73**	**0.68**	0.65
LLM	0.65	0.63	**0.68**	0.62

Table 2: Comparison of macro-average performance across different pipelines.

5.2 Comparison with Shared Task Baselines

While our internal evaluation highlights the hypothesis-augmented Longformer as our strongest methodology, it is essential to contextualize these results within the broader competitive landscape of the SemEval-2026 shared task. The official revised baseline provided by the task organizers for the hidden evaluation set achieved a Macro F1 score of 0.82. Furthermore, top-performing systems in the shared task demonstrated that even higher absolute scores are achievable. For instance, the top-ranked system (TeleAI) achieved a Macro F1 of 0.89, and the second-ranked system (AsymVerify) achieved 0.85, likely through extensive ensembles and more advanced multi-step verification strategies.

While our approach (0.67 Macro F1) does not exceed these upper-bound benchmarks, the systematic comparison provides valuable insight: reformulating the task via hypothesis-augmented classification closes the performance gap for resource-constrained encoder models.

To better understand model behavior, Table 3 presents the F1-score breakdown for each clarity category, while Figure 1 provides a detailed visualization of the specific misclassification patterns for our best-performing NLI pipeline.

Category	DSC	NLI	LLM
Ambivalent	**0.73**	0.71	0.56
Clear Reply	0.69	0.59	**0.78**
Clear Non-Reply	0.53	**0.71**	0.53

Table 3: F1-score breakdown per clarity category.

Detecting the *Clear Non-Reply* category proved most challenging for discriminative models. The

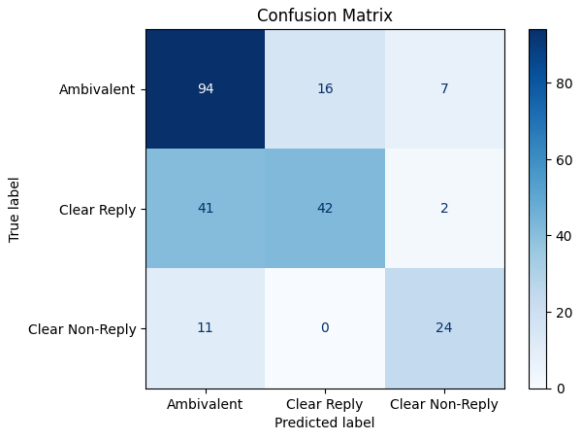

Figure 1: Confusion matrix for the NLI Pipeline (Longformer) on the official hidden test set.

baseline DSC model achieved only 0.53 F1 for this class. However, by incorporating class-weighted loss in Strategy 2, the NLI model improved performance to 0.71 F1, representing an 18-point increase. As shown in the confusion matrix (Figure 1), the NLI model effectively identifies the majority of non-replies, although some instances are still misclassified as ambivalent responses.

This substantial improvement suggests that the primary bottleneck in evasion detection stems from majority-class bias rather than insufficient linguistic signal. When the imbalance is explicitly addressed through weighted optimization, the model becomes significantly more sensitive to minority evasive behaviors.

The LLM-based system exhibited a distinct performance pattern. It outperformed the fine-tuned encoders in identifying *Clear Replies*, achieving 0.78 F1. This suggests strong semantic sensitivity to direct and explicit answers.

However, the LLM struggled to differentiate between *Ambivalent* and *Clear Non-Reply*. It frequently over-generalized nuanced evasions and displayed weaker performance on the minority class. This indicates that while large generative models possess broad semantic competence, they may lack the domain-specific calibration that supervised fine-tuning provides.

5.3 Impact of Context Length

A key advantage of Strategy 2 lies in its use of the Longformer architecture, supporting a maximum sequence length of 1024 tokens during our experiments, compared to 512 tokens in the DSC baseline.

To rigorously evaluate the impact of sequence length on performance, we stratified the official test set into non-truncated ($\leq$ 500 tokens, 213 instances) and truncated ($>$ 500 tokens, 24 instances) subsets. The empirical results demonstrate that the DSC baseline (RoBERTa) experienced a severe performance degradation on long sequences, with its macro F1-score dropping from 0.62 on short sequences down to 0.52 on long sequences. In contrast, the Longformer NLI pipeline demonstrated significantly higher overall effectiveness and resilience. While it also exhibited a performance drop on the hardest, longest sequences (dropping from 0.71 on short sequences to 0.59 on long sequences), its long-sequence performance (0.59) remained highly competitive with the baseline's short-sequence performance (0.62). These results are visually summarized in Figure 2.

This length-stratified analysis confirms that evasion detection in political discourse often depends on long-range dependencies, and models restricted by limited context windows suffer disproportionate information loss when parsing extended answers.

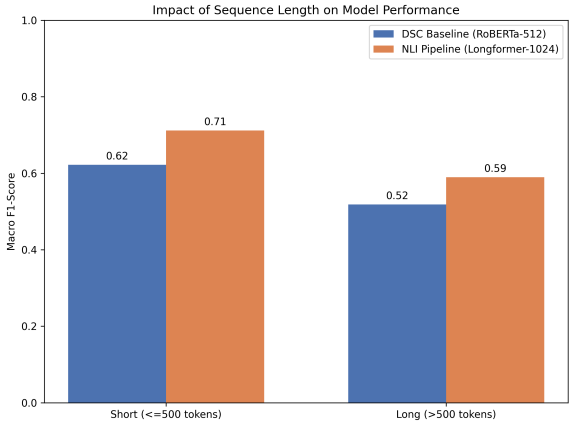

Figure 2: Impact of Sequence Length on Macro F1-Score (DSC Baseline vs. NLI Pipeline). The Longformer-based NLI pipeline exhibits significantly higher resilience on sequences exceeding 500 tokens.

6 Conclusion

This paper presented the Sentiment Syndicate team's approach to SemEval-2026 Task 6, Subtask 1, which focuses on classifying the clarity level of political question–answer interactions. We explored three modeling strategies: a RoBERTa-based sequence classification pipeline, a NLI reformulation, and a prompting-based LLM approach.

Our experimental results demonstrate that reformulating the task via hypothesis-augmented su-

pervised classification provides a more balanced modeling framework for capturing semantic alignment between questions and answers. Among our proposed methods, the hypothesis-augmented pipeline achieved the strongest and most stable performance, outperforming standard fine-tuning and most prompting-based baselines. The results suggest that explicitly framing the input as a premise-hypothesis pair helps the system better detect clarity, ambiguity, and non-replies, even without prelearned entailment geometry.

Limitations

Despite promising results, our approach has several limitations. First, the dataset size is relatively small for training transformer-based models, particularly for minority classes such as *Clear Non-Reply*. Although weighted loss was applied to mitigate class imbalance, the model may still exhibit bias toward the majority *Ambivalent* class.

Second, the NLI reformulation relies on a manually designed hypothesis template ("The answer clearly addresses the question: {subquestion}"). Our setup maps the standard NLI outputs (Entailment, Neutral, Contradiction) directly to the three clarity labels (Clear Reply, Ambivalent, Clear Non-Reply) using this single hypothesis. This 1-to-1 mapping assumes that clarity and responsiveness are perfectly aligned with standard entailment geometry. The effectiveness of this approach may depend on the phrasing of the template, and future work may split the hypotheses using distinct templates for clear replies, ambivalence, and explicit refusals to obtain richer semantic labels and performance outcomes.

Finally, our experiments were limited by available computational resources (Google Colab GPU environment). More extensive hyperparameter tuning and larger-scale experimentation may further improve the results.

Future work should explore more robust inference formulations, richer discourse-aware representations, and larger annotated datasets to better capture the political question evasiveness.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Peter Bull and Kate Mayer. 1993. How politicians evade questions. *Political Psychology*, 14(4):651–669.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *International Conference on Learning Representations*.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Alexandr Pak and et al. 2024. Word embeddings: A comprehensive survey. *Computación y Sistemas*, 28(4):2005–2029. Epub March 25, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Daniel Preotiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: Political ideology prediction from text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 729–740.

Parameswari Rasiah. 2010. A framework for the analysis of evasion in parliamentary discourse. *Journal of Pragmatics*, 42(3):664–680.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. 2024. "I never said that": A dataset, taxonomy and baselines on response clarity classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5204–5233, Miami, Florida, USA. Association for Computational Linguistics.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. 2026. Semeval-2026 task 6: Clarity – unmasking political question evasions. *Preprint*, arXiv:2603.14027.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei-Fei Li, Denny Zhou, Ed Chi, and Quoc V Le. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 38 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4814–4819.

Caleb Ziems, William Held, Omar Asiedu, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Nature Computational Science*, pages 1–13.

A LLM Prompt Template

The following prompt template was used for the zero-shot and few-shot classification experiments using the `gpt-oss:120b` model. The template includes system instructions, label definitions, and illustrative examples to guide the model's reasoning process.

```
You are an expert content analyzer
    specializing in political
    communication and discourse analysis.

Your task is to analyze an interview
    Answer in the context of the full
    Interview Context and determine if
    it clearly answers a specific Target
     Question.

### Categories

1. **Clear Reply**
  - The interviewee directly answers the
      Target Question.
  - Includes Direct Answers (Yes/No/
    Specifics).
  - Includes Negations/Corrections (e.g
    ., "It's not X, it's Y").
  - Includes replies with necessary
    context, as long as the core
    question is definitively addressed.


2. **Ambivalent** (The "Soft Dodge")
  - The answer is partial, vague, or
    indirect regarding the Target
    Question.
  - Filibustering: The interviewee
    discusses the general topic at
    length but avoids the specific
    constraint of the question.
  - Hedging: Uses non-committal language
     ("I'm not going to predict," "It
    depends," "We are monitoring") to
    avoid a definitive stance.
  - Inference Required: The listener
    must infer the answer; it is not
    stated explicitly.

3. **Clear Non-Reply** (The "Hard Dodge
    ")
  - Explicit Refusal: "No comment," "I
    won't answer that."
  - Total Ignore: Completely changes the
     subject to an unrelated topic.
  - Dismissal: Ends the interview or
    attacks the question/interviewer
    without answering.
  - Stalling: "I'll get back to you," "
    Ask my team" (with no current
    answer).


### Few-Shot Examples

[Representative few-shot examples
    including context, question, answer,
     classification, and chain-of-
    thought reasoning are inserted here.
     Examples are sampled from the
    training set to cover all three
    clarity categories.]


### Input Format
You will be provided with:
- **Context**: The full context of the
    interview turn.
- **Target Question**: The specific
    question to evaluate.
- **Answer**: The response given by the
    interviewee.

### Output Format
You must output strictly in JSON format.
     The output should be a valid JSON
    object with the following keys:
- "classification": The classification
    label (Clear Reply, Ambivalent, or
    Clear Non-Reply).
```

```
- "reasoning": A brief explanation of
    why the classification was chosen.

Example:
{
  "classification": "Clear Reply",
  "reasoning": "The interviewee directly
      addresses..."
}


### Task
Classify the following:

**Context**: {context}
**Target Question**: {target_question}
**Answer**: {answer}
```

Listing 1: System prompt template for clarity classification.