Unmasking Political Question Evasions

SemEval 2026 Challenge

Team members:

- Aldea Andrei
- Burcă Theodor
- Duluță George

1. Introduction

1.1. Problem Definition: Political Equivocation and Response Clarity

Evasion is not only the concept of lying but also a phenomenon that describes a non-straightforward type of communication, which is furthermore characterised by lack of clarity. This also includes speech acts such as contradictions, inconsistencies, subject switches, incomplete sentences, misunderstandings, obscure mannerism of speech. All of this makes political speech susceptible to multiple interpretations in the public eye. On this idea, an automated system task would be to classify the relation between a Q&A pair in 3 categories:

- **Clear Reply**: replies that admit only one interpretation
- **Ambivalent**: response is given in the form of a valid answer but allows for multiple interpretations
- **Clear Non-Reply**: containing responses where the answerer openly refuses to share information

1.2. Distinction from Standard Question Answering and Intent Detection

It is crucial to distinguish the task of Response Clarity Evaluation from standard Question Answering (QA) or Deceptive Intent Detection. Traditional QA tasks often focus on the "answerability" of a question given a text span or detecting the responder's subjective intent to deceive. However, assessing intent is inherently subjective and prone to annotator bias. The proposed approach in this project deviates from analyzing intent and instead focuses exclusively on the informational alignment between the question and the response. This shifts the classification objective from a subjective "valid/invalid" judgment to a more objective assessment of whether a response can be interpreted unambiguously. By focusing

on the Clarity/Ambiguity dimension rather than Truthfulness, the task creates a deterministic framework suitable for NLP Models.

1.3. Linguistic Challenges and Multi-Part

Political interviews often contain multi-part questions, queries containing multiple distinct subquestions. A politician might answer one part of the question clearly while evading another, making the classification of the entire response as a single unit problematic. To address this, the problem definition requires decomposing these complex interactions into singular QA pairs (sQAs) to retain fine-grained information. Furthermore, accurately determining clarity requires the model to encode and reason over long contexts, a capability that has only recently become feasible with the advent of Large Language Models.

2. State of the Art

2.1. Traditional Methods

2.1.1 Text Representation

The standard method to transform text into numeric format was the Bag-of-Words model, where the word order is ignored, the text becoming a set of unordered frequencies. Furthermore, to evaluate the relevance of an answer to a question, the method needed to compute the TF-IDF vectors. If the question and the answer had sparse common words, the similitude score was big. The limitation of this method is that it cannot distinguish between negations and nuances. Phrases like "I will not raise taxes" and "I will raise taxes" look almost identical to a BoW algorithm, making evasion detection extremely difficult.

2.1.2. Classification Algorithms

Before the widespread adoption of deep neural networks, detecting relationships between texts (such as question-answer relationships) relied on supervised learning algorithms that used handcrafted features.

Naive Bayes: A probabilistic algorithm made on top of the Bayes theorem, used frequently for text classification (e.g., spam vs. non-spam). While fast, Naive Bayes is making the assumption that each word is independent in relation with the others, a false assumption in political speech, where context is crucial.

SVM: The algorithm tries to find a hyperplane that optimally separates the classes in a multidimensional space. Although SVM performed better than Naive Bayes on short texts, its performance dropped drastically when evasion was based on irony or subtle topic switching, as it could not capture complex semantic relationships.

Logistic Regression: Estimates the probability of a class membership using the logistic (sigmoid) function. It serves as a robust and interpretable baseline, particularly effective when the distinction between classes is linearly separable. However, similar to Naive Bayes, it struggles with the complexity of political discourse; as a linear classifier, it fails to capture non-linear relationships and the deep sequential context required to distinguish between a genuine answer and a nuanced evasion.

2.2. The RNN/LSTM Era: from sequential models to deep contextual understanding

Before the Transformer architecture reshaped modern NLP, recurrent models dominated nearly every task involving ordered text. The intuition was straightforward: language unfolds over time, so a model that processes information sequentially should, at least in theory, possess the right inductive bias. Classical RNNs attempted to encode this evolving structure by updating a hidden state at each timestep. Formally, the recurrence takes the form:

$$h_t = f(W_x x_t + W_h h_{t-1} + b)$$

where h_t represents the hidden state at time *t*, x_t is the input word embedding, and *f* is a nonlinear activation function. The main limitation of this architecture, already recognized in the early 2000s, was its vulnerability to the vanishing and exploding gradient problem. This effectively restricted the model's ability to remember long-range dependencies—an especially damaging weakness for political discourse, where evasive strategies can involve topic shifts, subtle detours, or delayed answers occurring many tokens after the original question.

2.2.1. Word Embeddings: Word2Vec and GloVe

The rise of distributed representations provided the first major leap. Instead of relying on sparse, brittle bag-of-words features, models began using word embeddings to capture semantic and syntactic regularities. Word2Vec's skip-gram objective, for example, maximized the probability of observing a context window around a given word:

$$max_\Theta \sum_{t=1}^{T} \sum_{w_c \in C_t} log\, p(w_c | w_t)$$

where C_t is the context of token w_t. These embeddings allowed models to grasp analogical relations and thematic similarities. For evasion detection, this meant being able to recognize when a reply drifted toward generalities ("prosperity", "leadership") rather than addressing the concrete substance of a question.

However, embeddings were static: the vector for bank could not adapt to distinguish a financial institution from the side of a river. In political interviews, where ambiguity is strategic, this lack of contextual sensitivity was a serious bottleneck.

2.2.2. LSTM and BiLSTM Models

LSTMs were introduced to mitigate the deficiencies of standard RNNs. Their gating mechanisms—input, forget, and output gates—regulated the flow of information through time. The cell state update can be summarized as:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

where f_t and i_t are the forget and input gates, respectively. The output gate o_t does not contribute to the cell-state update; instead, it determines how much of the internal memory becomes externally visible through the hidden state, computed as:

$$h_t = o_t \odot \tanh(c_t)$$

This separation between cell-state dynamics and output gating allows gradients to propagate more effectively across long sequences while maintaining a stable internal memory representation.

Bi-directional variants (BiLSTMs) further improved performance by processing sequences from both directions. This is particularly relevant in political interviews: the meaning of a candidate's reply often becomes clearer only after reading later clauses that subtly contradict or reframe earlier statements.

Despite these advantages, LSTMs had fundamental limitations. Their sequential nature made them slow to train and evaluate, and their capacity to track discourse-level phenomena remained inherently constrained. As answers grew longer or more rhetorically complex, the models' performance plateaued.

Although these architectures were initially developed for broad NLP tasks, they became directly relevant when Thomas et al. (2024) introduced the “I Never Said That” dataset and the associated taxonomy for response clarity in political interviews. The dataset formalizes the phenomenon of equivocation by labeling each question–answer pair along two axes: a coarse distinction between clear replies, ambivalent replies, and clear non-replies, and a fine-grained set of nine evasion strategies. Early attempts to model this task relied on sequential encoders such as LSTMs or BiLSTMs, which could capture local coherence but often failed to track the deeper rhetorical maneuvers that unfold across several clauses. This gap between linguistic behavior and model capacity ultimately motivated the shift toward architectures capable of richer contextual reasoning.

2.3. Transformers and DeBERTa: The current State of the Art

The clarity-classification problem introduced by Thomas et al. is inherently relational: an answer cannot be judged in isolation, but must be evaluated with respect to the expectations set by the question. This makes the task particularly well aligned with the strengths of Transformer architectures, which excel at modeling long-range dependencies and subtle mismatches between linguistic segments. In the CLARITY dataset, many instances of evasion are not marked by explicit refusal but by gradual drift, selective omission, or reframing—patterns that require the model to attend jointly to both sides of the interaction. Transformers, through their self-attention mechanism, offer precisely the representational flexibility needed to capture these dynamics.

The introduction of the Transformer architecture marked a decisive shift in NLP. Instead of relying on recurrence, Transformers use attention mechanisms to relate any two

positions in a sequence directly. This allows them to capture global interactions, making them ideally suited for analyzing question–answer pairs in political discourse.

2.3.1. The Self-Attention mechanism

Self-attention computes contextualized representations for each token by comparing it to all others in the sequence. Given queries (Q), keys (K), and values (V), the mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

This formulation allows the model to identify when an answer aligns with a question, when it avoids the topic altogether, and when it introduces unrelated information as a distraction. Such relational patterns are the hallmark of evasive strategies.

BERT demonstrated that bidirectional attention is critical for tasks involving deep semantic inference. Because the meaning of a political answer depends on both what is said and what is deliberately omitted, the bidirectional nature of Transformers is particularly advantageous.

2.3.2. DeBERTa and Disentangled Attention

DeBERTa introduced an important refinement: disentangling content and positional information. While BERT fuses both types of information in a single embedding, DeBERTa treats them separately, enabling more nuanced attention scoring. The core mechanism can be written as:

$$\alpha_{ij} = (q_i^c + q_i^p)(k_j^c + k_j^p)^\top$$

where q_i^c and k_j^c encode content, while k_i^p and k_j^p represent relative positional offsets.

This design allows the model to reason more effectively about discourse structure. For example, political answers often shift from direct responses to broad commentary. Being able to separate position from meaning helps the model detect when the speaker is drifting away from the informational target of the question.

2.3.3. Cross-encoder architectures for answer classification

For classification tasks involving question–answer pairs, the most effective Transformer-based design is the cross-encoder. Instead of encoding the question and answer independently, the model ingests them jointly as a single sequence:

[CLS] Question [SEP] Answer [SEP]

The [CLS] token ("classification") is a synthetic token inserted at the beginning of every input sequence; its final hidden representation is used by the classifier as a summary embedding of the entire question–answer pair.

The [SEP] token ("separator") is used to mark boundaries between segments, allowing the model to distinguish where the question ends and the answer begins.

This allows every token in the question to attend to every token in the answer, and vice versa. The resulting representation captures deviations, inconsistencies, and evasive patterns much more reliably than separate encoders could. Empirically, cross-encoders consistently outperform bi-encoders in tasks requiring relational understanding, which is precisely the core requirement of clarity and evasion classification.

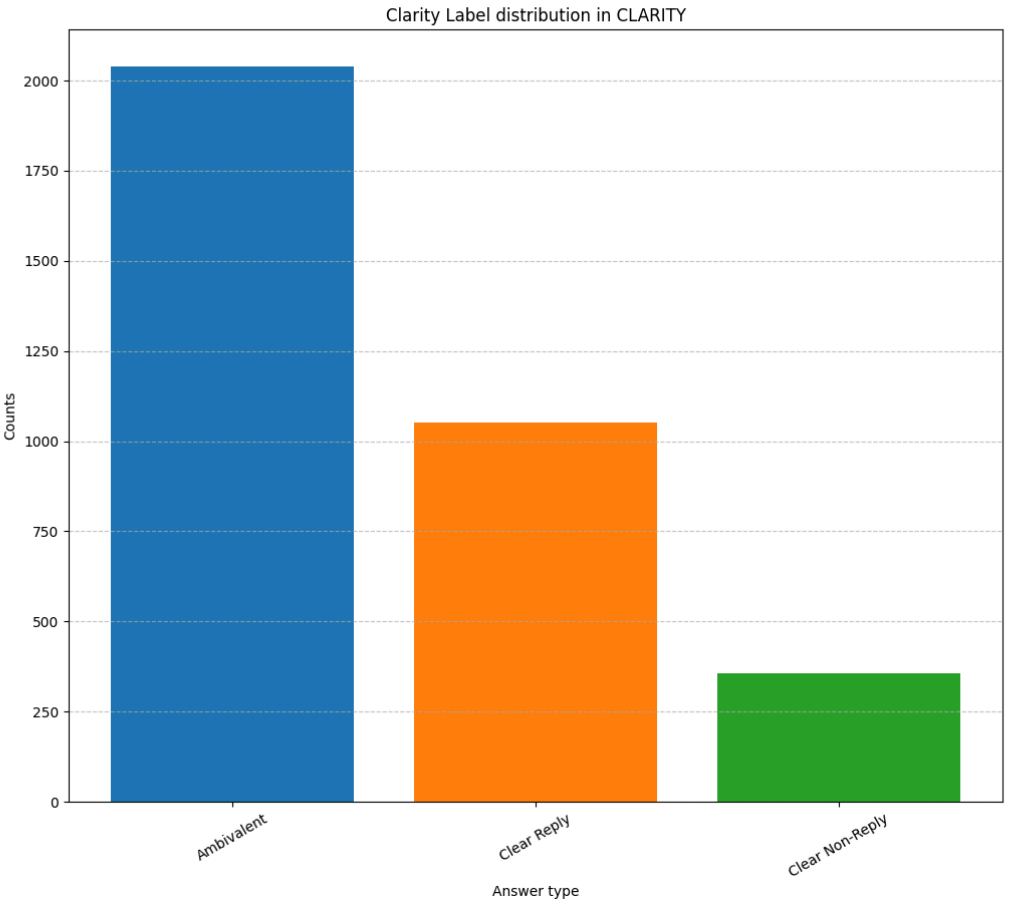
This joint encoding is essential for the CLARITY taxonomy, because the distinction between a Clear Non-Reply and an Ambivalent Reply, or between Deflection and General Evasion, hinges on how the content of the answer diverges from the informational target of the question. A cross-encoder does not merely encode the answer; it implicitly learns the rhetorical relationship between the two segments, mirroring the judgment process followed by annotators in the dataset.

2.4. Clarity Dataset

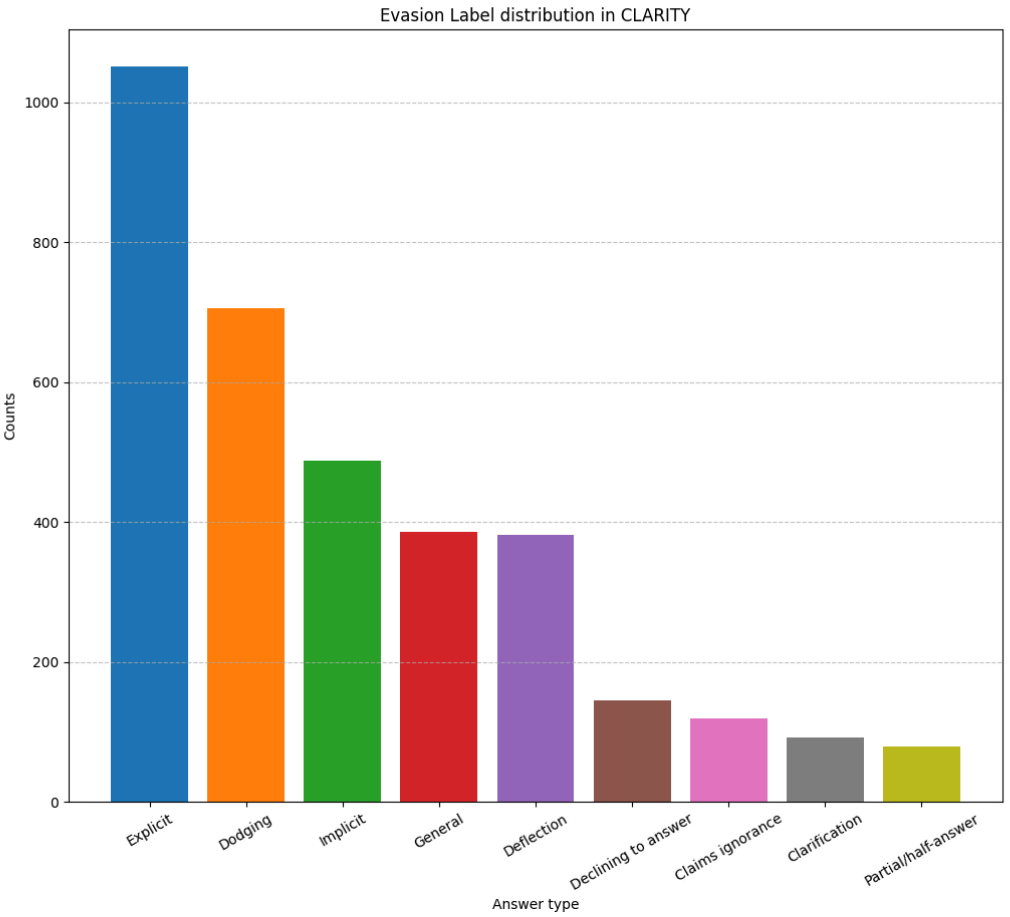
To train and evaluate the proposed model, this paper uses the CLARITY dataset, introduced by Thomas et al. (2024). It is an essential benchmark in the field of NLP for political discourse analysis, specifically designed to distinguish between responses that directly address a question and those that are evasive or irrelevant.

2.4.1. Data Composition and Source

CLARITY contains 3445 question and response pairs from 287 unique presidential interviews of US Presidents spanning from 2006 until 2023. This diversity is crucial to avoid ideological bias and to train a model capable of generalizing regardless of the speaker's political orientation.

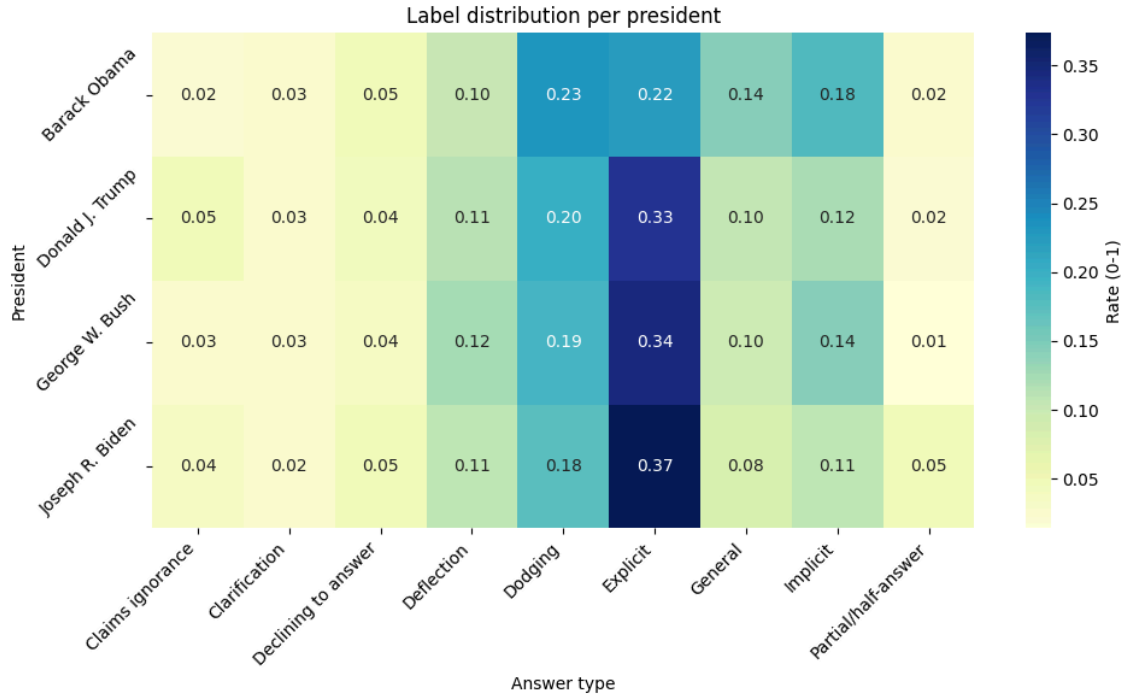


It is clearly observed that the "Ambivalent" class is dominant, with over 2000 instances, which is approximately double the "Clear Reply" class (~1050). The "Clear Non-Reply" class is the least represented (< 400 instances). From this stats we can draw the hypothesis that politicians prefer to give a vague answer ("Ambivalent") than to explicitly refuse to answer ("Clear Non-Reply"), because direct refusal can be negatively charged by the public.



Although ambiguity dominates overall, taken individually, the "Explicit" label is the most common single strategy (>1000). This means that when they don't use it, politicians are clear. Preferred Evasion Strategies: When politicians decide to be evasive, the preferred strategies are, in order:

- Dodging: Ignoring the question and moving on to another topic (~700).
- Implicit: Answers that suggest something without saying it directly (~500).
- General & Deflection: Generalizing or deflecting attention (~400 each).



Who is the most direct?

- Joe Biden seems to be the most direct, with the highest score on Explicit (0.37), followed by George W. Bush (0.34).

Who is the most evasive/sophisticated?

- Barack Obama has the lowest score on Explicit responses (0.22 - Explicit). Instead, he uses the Implicit (0.18) and Dodging (0.23) strategies the most. This suggests a complex rhetorical style, based on nuances and subtexts, which is harder to decode by a simple algorithm.

Donald Trump: He has a mixed style. Although he has a relatively high score on Explicit (0.33), he uses a lot of Dodging (0.20) but very little Implicit (0.12) compared to Obama.

2.5. Comparative Summary and Conclusions

The baselines reported in the original “I Never Said That” study reinforce this developmental arc. Traditional linear models trained on handcrafted features achieved only modest performance, especially on the nine-category evasion task. LSTM-based models improved the situation slightly but struggled with the nuanced boundary cases that define political communication. The most substantial gains emerged with Transformer-based models, particularly cross-encoders, whose performance provided the first strong baselines for both levels of the taxonomy.

The evolution of models for response clarity classification reflects broader trends in NLP. Early approaches depended heavily on feature engineering and statistical classifiers such as SVMs and Naive Bayes. Their inability to model semantic nuance or discourse coherence made them ill-suited for detecting political evasion.

RNNs and LSTMs extended capability by enabling sequential modeling and capturing patterns over time. They were a significant improvement but still struggled with context length and global reasoning. Political discourse frequently involves rhetorical devices whose interpretation depends on distant parts of the answer, something LSTMs inherently handle poorly.

Transformers, with their global self-attention operations, finally broke through these limitations. Their ability to model long-range dependencies, combined with deeper contextual understanding, made them the natural choice for tasks requiring fine-grained interpretation of question–answer relationships. Architectures like DeBERTa further enhanced this capability through improved attention mechanisms that separate content and positional information.

Taken together, these observations show that the CLARITY task is not merely a classification problem but a test of a model's ability to track pragmatic intent across a dialogue turn. The dataset captures a spectrum of evasive behaviors that cannot be detected through surface-level similarity alone, and the steady progression from RNNs to modern Transformer architectures illustrates how representational depth translates directly into performance on such subtle, discourse-driven tasks.

A synthetic comparison of model families can be summarized as follows:

Model Family	Strengths	Limitations	Typical Performance Range
Classical ML (SVM, NB)	Fast, interpretable, easy to train	No semantic depth, fragile to paraphrasing	~40–55% F1
RNN / LSTM	Sequential modeling, better semantics	Limited long-range reasoning, slow	~55–65% F1
Transformers (BERT)	Strong contextual learning	High computational cost	~70–78% F1
Advanced Transformers (DeBERTa)	Best-in-class reasoning, fine-grained discrimination	Heavy inference cost	80%+ F1

Overall, the State of the Art clearly favors Transformer-based cross-encoder architectures, especially for tasks where subtle deviations between question and answer reveal the speaker's intent. The CLARITY taxonomy, with its hierarchical structure, aligns closely with the multi-layer representations learned by these models, making modern Transformers not just suitable but essential for high-performance evasion classification.

3. Methods

3.1 Classical Machine Learning Baselines

As a first set of baselines, we implement several traditional supervised classifiers: Naive Bayes, Support Vector Machines (SVM), and Logistic Regression. For these models, question–answer pairs are represented using TF–IDF features computed over the concatenation of the question and the corresponding answer.

These approaches serve as lightweight reference points, allowing us to quantify the limitations of surface-level lexical representations and linear decision boundaries in capturing the semantic and pragmatic complexity of political discourse.

3.2 Fine-Tuned Transformer Models

To model deeper contextual relationships between questions and answers, we employ several pre-trained Transformer architectures, including bert-base-uncased, distilbert-base-uncased, roberta-base, albert-base-v2, and gpt2. All Transformer models are implemented as cross-encoders, jointly encoding the question and answer as a single input sequence of the form:

[CLS] Question [SEP] Answer [SEP]

To jointly address clarity-level and evasion-level classification, we wrap the Transformer backbone within a multi-task learning framework, using a shared encoder and separate classification heads for each task. Models are fine-tuned on the training split for three epochs using standard cross-entropy loss, and performance is evaluated on the held-out test split.

3.2 Prompting-Based Large Language Models

In addition to supervised learning approaches, we investigate prompting-based inference using large language models without any task-specific fine-tuning. This setup differs fundamentally from the methods above, as model parameters remain frozen and task adaptation is achieved solely through natural language instructions.

We evaluate GPT-4o-mini in two configurations:

1. Zero-shot prompting, where the model is provided only with task instructions and label definitions.
2. Few-shot prompting, where six labeled examples (two per clarity class), sampled from the training split, are included in the prompt to calibrate the model to the dataset-specific interpretation of clarity labels.

For each test instance, the model is instructed to output exactly one clarity label. Outputs are normalized and mapped to the target label set. This approach allows us to assess the extent to which response clarity distinctions can be recovered through inference-time reasoning alone.

4. Experimental Results

The separation into training and test data was done by Thomas et al. and provided to us via huggingface.

Our implemented methods are described bellow:

1. SVM
2. Naive Bayes
3. Logistic Regression
4. Pretrained transformer

For our pretrained transformers, we wrapped it inside a MultiTaskModel in order to simultaneously take into account the Clarity Label and Evasion Label. For this task we did 3 epochs of additional training on the Training dataset. After that we manually ran through different models, to collect data on their performance.

In the table below we categorize our findings by model and by the results in classifying the clarity label of the question answer pair (the Accuracy CL, Precision CL, F1 Score CL columns) and by the results in classifying the evasion label (the Accuracy EL, Precision EL, F1 Score EL columns)

Model	Accuracy CL	Precision CL	F1 Score CL	Accuracy EL	Precision EL	F1 Score EL
SVM	0.5986	0.5820	0.5708	0.2899	0.3395	0.2959
Naive Bayes	0.3246	0.6445	0.1819	0.3174	0.2627	0.1671
Logistic Regression	0.5652	0.5747	0.5502	0.2696	0.3211	0.2740
bert-base-uncased	0.6435	0.6295	0.6204	0.3507	0.3352	0.2912
distilbert-base-uncased	0.6377	0.6242	0.6052	0.3565	0.3214	0.3045
roberta-base	0.6580	0.6560	0.6262	0.3507	0.2137	0.2575
albert-base-v2	0.5855	0.3428	0.4324	0.2957	0.0874	0.1349
gpt2	0.6609	0.6554	0.6241	0.3594	0.3609	0.3295

In addition to supervised and fine-tuned models, we investigate whether large language models can perform response clarity classification purely through **prompting**, without any task-specific training or parameter updates. This setting differs fundamentally from the approaches above, as the model weights remain frozen and only natural language instructions are used at inference time.

We evaluate GPT-4o-mini as a representative instruction-following language model in two configurations: zero-shot and few-shot prompting. In the zero-shot setting, the model receives only a task description and label definitions. In the few-shot setting, six labeled examples (two per clarity class) sampled from the training split are included in the prompt to calibrate the model to the dataset-specific interpretation of the clarity labels.

The following table reports the results for clarity-level classification. Zero-shot prompting achieves a Macro-F1 score of 0.44, indicating that the model captures some aspects of response clarity without supervision. Few-shot prompting substantially improves performance, reaching a Macro-F1 of 0.59, corresponding to an improvement of more than 15 absolute points. This gain is primarily driven by better discrimination between Clear Reply and Ambivalent responses, which constitute the most frequent source of confusion in the zero-shot setting.

While fine-tuned Transformer cross-encoders remain the strongest approach in terms of macro-averaged F1 score, prompting-based inference achieves competitive performance without any task-specific training. In particular, few-shot prompting attains higher accuracy than all fine-tuned baselines and approaches their Macro-F1 scores, indicating that large language models encode substantial pragmatic knowledge that can be effectively elicited through carefully designed prompts.

Model	Prompting Strategy	Accuracy (CL)	Precision (Macro)	Recall (Macro)	F1 (Macro)
GPT-4o-mini	Zero-shot	0.604	0.59	0.51	0.44
GPT-4o-mini	Few-shot (6 ex.)	0.708	0.60	0.59	0.59

5. Conclusion

This study assessed various NLP models on the CLARITY dataset for political question evasion classification (Clear Reply, Ambivalent, or Clear Non-Reply) and fine-grained strategy identification. Classical machine learning approaches proved insufficient for this task, achieving poor performance due to their inability to capture semantic nuance and discourse-level reasoning. In contrast, fine-tuned Transformer-based cross-encoders, particularly **bert-base-uncased**, achieved strong results on clarity-level classification (Macro-F1: **0.6204**), outperforming other base models and classical methods, confirming the effectiveness of joint question–answer encoding for modeling political equivocation.

However, performance on fine-grained evasion strategy classification remained substantially lower (Macro-F1: **0.4319**), indicating that while models can reliably detect the presence of evasion, they still struggle to distinguish between specific rhetorical strategies. This highlights the inherent difficulty of fine-grained pragmatic analysis and the impact of class imbalance in the dataset.

In addition to supervised approaches, we explored prompting-based inference using large language models without task-specific fine-tuning. Zero-shot prompting achieved moderate performance (Macro-F1: 0.44), while few-shot prompting substantially improved results (Macro-F1: 0.59), narrowing the gap with fine-tuned models. These findings suggest that large language models encode a significant amount of pragmatic knowledge relevant to response clarity assessment, which can be effectively elicited through carefully designed prompts.

Overall, our results validate Transformer-based models as the current state of the art for response clarity detection, while also demonstrating that prompting-based methods offer a competitive, training-free alternative. Future work should investigate more advanced architectures (e.g., DeBERTa), improved handling of class imbalance, and hybrid approaches that combine fine-tuning with prompt-based reasoning to further advance the automatic analysis of political question evasion.

6. References

- Thomas, K., Filandrianos, G., Lymperaiou, M., Zerva, C., & Stamou, G. (2024). “I Never Said That”: A dataset, taxonomy and baselines on response clarity classification. Findings of EMNLP 2024.
- HuggingFace QEvasion Dataset Card. QEvasion: Political Question Evasion Dataset.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. EMNLP 2014.
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. Summary article: https://colah.github.io/posts/2015-08-Understanding-LSTMs/
- Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. NeurIPS 2017. Illustrated explanation: https://jalammar.github.io/illustrated-transformer/
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT 2019. Documentation summary: https://huggingface.co/docs/transformers/model_doc/bert
- He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. ICLR 2021.
- Graves, A. (2012). Supervised Sequence Labelling with Recurrent Neural Networks. Springer.
- Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. arXiv preprint arXiv:1406.1078.