UFG-Semantic at SemEval-2026 Task 6: CLARITY - Unmasking Political Question Evasions

Aline Hamano, Beatriz Felicio, Henrique Galvão and **Nádia Felix**
Institute of Informatics
Federal University of Goiás
Goiás, GO, Brazil
{aline_soma, beatrizfelicio, henriquegalvao}@discente.ufg.br, nadia.felix@ufg.br

Abstract

We propose an approach for *Task 6: CLARITY - Unmasking Political Question Evasions*. We make use of data augmentation, supervised fine-tuning, and model benchmarking to detect and classify response ambiguity in political discourse. Building on well-founded theory on equivocation and leveraging recent advancements in language modeling, our system was structured based on question/answer (QA) pairs extracted from presidential interviews, and it was evaluated exclusively in Evasion-level Classification (Subtask 2).

1 Introduction

The SemEval-2026 Task 6: CLARITY addresses the complex dynamics of political discourse through two distinct challenges: Clarity-level Classification (Subtask 1) and Evasion-level Classification (Subtask 2). Although the overall task encompasses both concepts, this paper focuses exclusively on solving Subtask 2. We delimit our scope to evasion because detecting political deflection requires analyzing deep semantic nuances and pragmatic shifts, rather than acoustic or transcript-level speech clarity (Thomas et al., 2024, 2026).

In contexts such as presidential speeches and interviews, evasion often manifests through ambiguity, where words or phrases have multiple conflicting interpretations (Liu et al., 2023). This makes it difficult to distinguish between straightforward answers and calculated dodges, as language contains a variety of double meanings, irony, and other invisible factors that complicate overall interpretation (Eisenberg, 1984). Politicians can consciously use these mechanisms to improve evasion techniques when answering sensitive questions (Lihong and Weijie, 2018). Consequently, our focus is on presenting computational solutions to classify the level of evasiveness, given that politicians wield considerable influence and misinterpretations can be dangerous.

To address Subtask 2, we apply a three-stage pipeline: (i) Data Augmentation via PromDA (Prompt-based Data Augmentation) (Wang et al., 2022), which leverages prompt-driven generation to synthetically expand the training set; (ii) the LAQDA architecture with few-shot learning (Liu et al., 2024), which adapts the model to the target classification task from a limited number of labeled examples[1]; and (iii) Error Analysis to investigate the generalization gap between augmented and natural discourse.

2 Related Work

Alvarez and Morrier (2025) propose a self-supervised approach to assess answer quality in parliamentary Q&A by training a Sentence-BERT biencoder on 58,343 Canadian Question Period exchanges. Answer quality is inferred from question–answer embedding distance, modeling evasion as a single continuous dimension, which obscures fine-grained distinctions between different evasive behaviors.

Another structurally related line of work investigates answerability in reading comprehension; that is, whether a given question can be answered based on a provided passage. Rajpurkar et al. (2018) introduced SQuAD 2.0, combining over 100,000 answerable questions with more than 50,000 adversarially crafted unanswerable ones designed to closely resemble answerable questions, requiring models to both extract answers and abstain when no answer is supported by the context. This framing established the binary distinction between answerable and unanswerable as a first-class NLP problem.

In contrast, the present work adopts a supervised classification framework grounded in an explicit two-level evasion taxonomy (Thomas et al., 2024), enabling the model to distinguish between specific

[1]All resources are available at https://github.com/henriquegalva0/clarify-data-augmentation

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2384–2393

discourse strategies, such as implicit replies, political deflections, and outright refusals that lie in the same region of semantic space but differ in their pragmatic function. The cost of this granularity is the need for human-labeled data, and the challenge is further compounded by the class imbalance inherent to the evasion dataset, where certain evasion classes are substantially underrepresented relative to others. This directly motivates our use of two complementary strategies to operate effectively under annotation scarcity: LLM-based data augmentation via PromDA (Prompt-based Data Augmentation) (Wang et al., 2022), which leverages prompt-driven generation to synthetically expand the training set, and the use of the LAQDA with few-shot (Liu et al., 2024), which adapts the model to the target classification task from a limited number of labeled examples.

3 Our Approach

Our approach follows a three-stage pipeline (see Figure 1): (1) data augmentation via PromDA using Gemini 2.0 Flash [2] (Gemini Team et al., 2025), employing specialized prompts informed by a statistical analysis of response length, sentiment, and linguistic fillers to mitigate class imbalance; (2) the LAQDA (Label-Adapter and Query-Data-Augmenter) architecture, which utilizes a BAAI/bge-m3 encoder (Chen et al., 2025) and cross-attention to capture semantic interactions between evasion labels and political discourse (Vaswani et al., 2023); and (3) prototypical refinement, using a transductive sampler (Liu et al., 2019) to dynamically shift class centers based on query-set similarity. The system is optimized through 9-way 55-shot (Vinyals et al., 2017) episodic training with a combined cross-entropy and regularization loss function, designed to handle the pragmatic nuances of political evasion.

3.1 Data Augmentation

The original dataset consists of 20 features, including metadata (*title, date, president, url*), interview content (*interview_question, interview_answer, question, question_order*), automated processing outputs (*gpt3.5_summary, gpt3.5_prediction*), and human annotations (*annotator_id, annotator1, annotator2, annotator3*). Additionally, binary flags for *inaudible*, *multiple_questions*, and *affirmative_questions* were included, alongside the primary classification targets: *clarity_label* and *evasion_label* (Table 1).

Considering that evasion labels are subgroups of clarity labels and that most of the data provided by the original dataset (Thomas et al., 2024) is not useful for training the model, we selected only *interview_question, interview_answer*, and *evasion_label* as our main columns for the filtered dataset.

Table 1: Representation of the filtered dataset with random example samples.

interview_ question	interview_ answer	evasion_ label
Q. What is your...	Our should be...	General
Q. But to follow...	I think they...	Deflection
Q. Mr. President...	You know, I...	Dodging
...	...	...
Q. How will you...	We will focus on...	Implicit

The original dataset shows a significant class imbalance. While the most frequent label, *Explicit*, contains 1,052 samples, there are other classes such as *Clarification* and *Partial/half-answer* that have fewer than 100 samples each (Table 2). This distribution is a challenge for model training, since underfitting on minority classes may lead to overall imprecision and bias (Bashir et al., 2020). To mitigate this, we explore data augmentation techniques in the following section.

Table 2: Label distribution in the original training set.

Evasion Label	Count	Evasion Label	Count
Explicit	1,052	Deflection	381
Dodging	706	Declining to answer	145
Implicit	488	Claims ignorance	119
General	386	Clarification	92
Partial/half-answer	79		
		Total	**3,448**

3.1.1 Few-shot Prompt-based Data Augmentation (PromDA)

In this approach, we designed a poly-specialist prompt to generate synthetic data across all categories (Wang et al., 2022). A *few-shot* prompting strategy was used, incorporating five examples of `question, answer, label` triplets to ensure structure and semantic alignment (Brown et al., 2020) with the original dataset (Thomas et al., 2024). In details, we instructed the model to act as an expert (White et al., 2023) in political discourse analysis with the mission of generating three variations for each input sample. The generation process was constrained by: (i) semantic preservation of the

[2]Free model avaiable in `https://aistudio.google.com`.

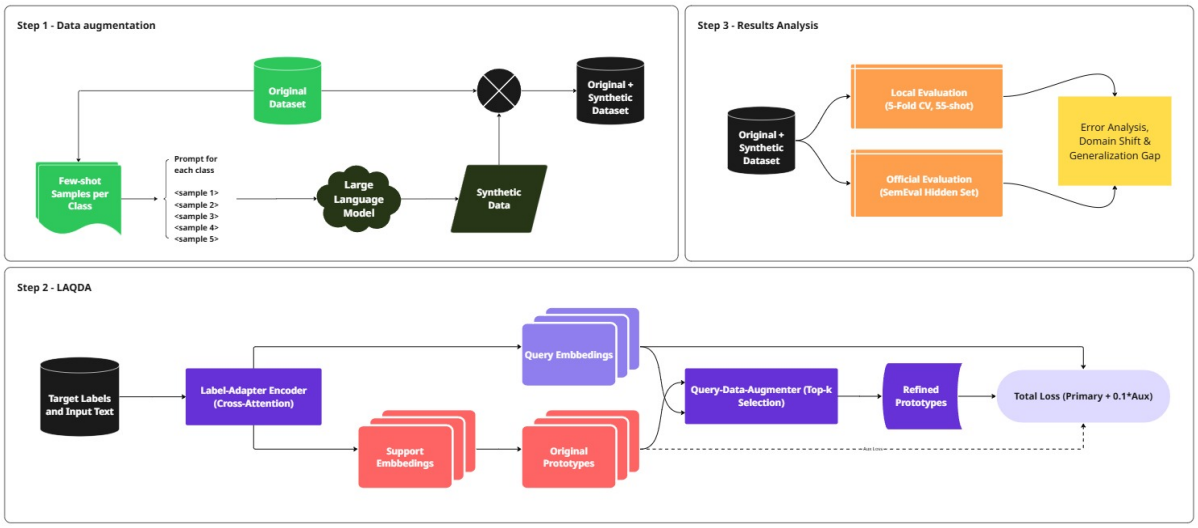


Figure 1: Pipeline describing our full approach to the task. Step 1: Data augmentation 3.1. Step 2: *LAQDA* (Label-Adapter and Query-Data-Augmenter) architecture 3.2. Step 3: Error Analysis and Discussion 5.2.

original label, (ii) maintenance of the core political intent, and (iii) strict adherence to a JSON output format for systematic parsing.

3.1.2 Few-shot and LLM-based Generation Techniques

In the second approach, we developed nine distinct prompts, each tailored to the specific linguistic and nuances of the target evasion labels (Table 3). Unlike the single-prompt strategy, this method allowed more control over the generation process by incorporating label-specific constraints. Specifically, each prompt defined: (i) the core semantic mechanism of the evasion technique (e.g., claiming ignorance, declining to answer, or implicit responses), (ii) precise character length constraints, and (iii) target sentiment polarity (negative or neutral). Furthermore, to increase the structural complexity of the synthetic samples, we introduced a variable percentage of linguistic fillers and hedges - ranging from 20% to 70% (Nikolić and Nikolić, 2022) depending on the label's typical characteristics. This granularity ensures that the synthetic data not only preserves the label's intent but also mimics the natural rhetoric found in political discourses.

The data augmentation was made by the `gemini-2.0-flash` model[3] (Gemini Team et al., 2025). The implementation used the configuration described in Listing 1. To promote accessibility, we employed Gemini 2.0 Flash, leveraging its robust free-tier API to generate high-quality synthetic samples for all classes.

[3]More information about the model in https://gemini.google.com/

Table 3: Short representation of parameter variations across specialized prompts.

Target Label	Length Limit	Fillers/ Hedges	Sentiment
Explicit	1500 chars	20%	Pos/Neutral
Clarification	500 chars	40%	Neg/Neutral
Implicit	2200 chars	70%	Neutral
...	...	...	...
General	2000 chars	60%	Neutral

Listing 1: LLM configuration for data augmentation.

```
llm = ChatGoogleGenerativeAI(
    model="gemini-2.0-flash",
    temperature=0.45,
    google_api_key=API_KEY
)
```

3.1.3 Dataset Analysis for Prompt Engineering

The values selected to build the previously mentioned nine prompts were obtained through a detailed analysis of the original dataset. The following *prompt template* is the main prompt structure.

Listing 2: Fraction of the prompt template.

```
* Keep the original evasion technique <evasion
    technique> seen in the few-shot examples.
* The idea of <evasion technique> is <definition
    of the evasion technique>.
* Preserve the core political intent and meaning,
    but rephrase creatively without losing the
    evasion technique used.
* Introduce diverse speaking styles (formal,
    colloquial, rhetorical, defensive, assertive
    , vague, etc.).
```

```
* Maintain an answer length <length percentage
    per technique> above or less <characters
    value per technique> characters.
* Ensure logical consistency between interviewer
     and president.
* Develop a mostly <sentiment per technique>
    sentiment returning the output.
* On <fillers percentage per technique> of the
    generated samples try adding <number of
    fillers per technique> sentence usual
    fillers and hedges.
* Avoid factual errors, hallucinations or *
    unrelated topics* while building the
    questions and answers.
```

3.1.3.1 Answer length analysis

The length of the answers to each evasion technique is important to capture meaning and readability (Matthews and Folivi, 2023). Therefore, we developed a length analysis over the president's answers to ensure efficiency while generating data with the prompts.

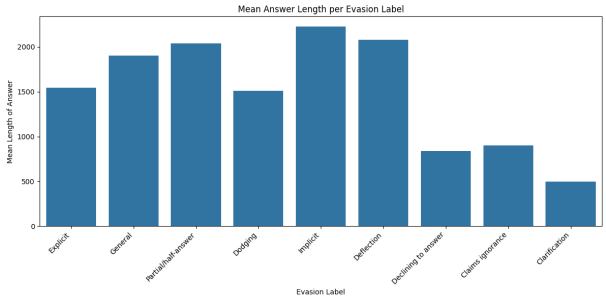


Figure 2: Average length of answers to interviewer questions.

3.1.3.2 Fillers and hedges analysis

Excessive use of fillers in scientific presentations can reduce the credibility of the (Villar and Castillo, 2016) speaker as well as impair the comprehension of the speaker's message by the audience (Seals and Coppock, 2022). To optimize data generation and ensure linguistic precision, we conducted a quantitative analysis of fillers and hedges within the president's responses.

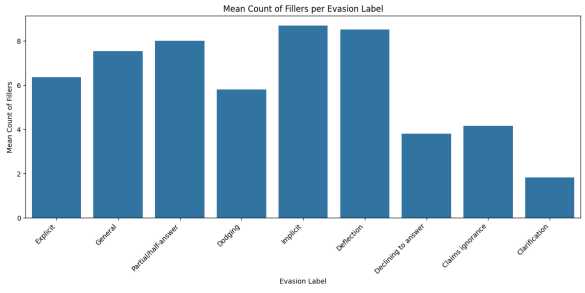


Figure 3: Average count of fillers and mitigating expressions in respondents' answers to interviewer questions.

3.1.3.3 Sentiment analysis

Beyond lexical choices, the emotional value of a response - whether positive, negative, or neutral - serves as a proxy for the speaker's strategic stance (Gennaro and Ash, 2021). High levels of negativity may indicate defensive evasion or "attacking the questioner," while neutral or positive tones can signal attempts to de-escalate tension or bypass critical inquiries through diplomatic phrasing. By quantifying these emotional markers, we can correlate sentiment patterns with specific evasion techniques, offering a more robust understanding of theoretical effectiveness and potential bias in the generated data.

The analysis was first developed on the interviewer questions to understand its disparity when compared to the responses' sentiments (see Figure 5 in Appendix C for the full overview of the training set).

3.2 LAQDA

To address the challenges of classifying political question evasions, especially given the imbalanced and nuanced nature of the dataset, we propose a Few-Shot Learning (FSL) framework based on Prototypical Networks. Specifically, our system adapts the LAQDA (Label-Adapter and Query-Data-Augmenter) architecture (Liu et al., 2024), which enhances traditional FSL through two main modules: a Label-Adapter text encoder and a Query-Data-Augmenter transductive sampler.

3.2.1 Label-Adapter Encoder

Traditional text classification models often encode the input text independently of the target labels. In our system, we employ a Label-Adapter encoding mechanism. We utilize a generic Transformer model to process both the input text (a concatenation of the interview question and answer) and the textual description of the target classes.

Formally, the model tokenizes the input text and the task classes to extract their respective embeddings. The label embeddings are expanded to match the batch dimension and concatenated with the sentence embeddings. A cross-attention layer is then applied, allowing the model to learn the semantic interaction between the input discourse and the specific evasion categories. A residual connection is added to preserve the original sentence representation, yielding the final task-adaptive embedding.

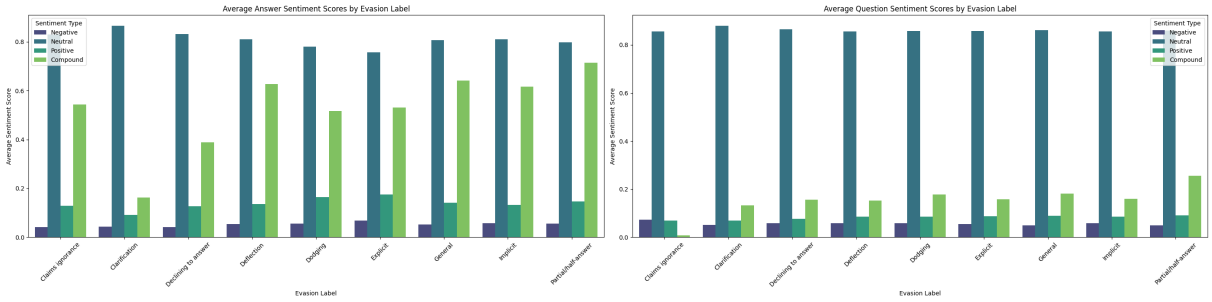


Figure 4: Average sentiment analysis on each label from the original dataset.

3.2.2 Query-Data-Augmenter (Transductive Sampler)

In standard Prototypical Networks, a prototype for each class is calculated as the mean of its support set embeddings. However, in low-resource scenarios, these prototypes can be biased. To mitigate this, we implement a Query-Data-Augmenter, acting as a transductive sampler.

This module dynamically refines the prototypes by estimating them together with query set samples. It calculates the cosine similarity between the support and query embeddings. By selecting the top-k most similar query samples for each class, the system transductively incorporates this unlabeled data to recalculate and shift the prototypes closer to the true class centers.

3.2.3 Loss Formulation

Our network is optimized using a combined loss function. The primary loss is the Cross-Entropy over the query set, computed using the Euclidean distance between the query embeddings and the *refined* prototypes (where shorter distances yield higher log-probabilities).

To ensure the stability of the transductive sampler, we introduce an auxiliary regularization loss. This secondary loss penalizes the Cross-Entropy of the pseudo-labeled sampled data against the *original* (unrefined) prototypes. The overall objective function is the sum of the primary loss and the auxiliary loss scaled by a factor of 0.1.

4 Experimental Setup

4.1 Data Processing

To format the data for the FSL episodic training, we merged the corresponding columns into a single string using a separator: `Question [SEP] Answer`. Duplicate entries were removed to prevent data leakage. The dataset was then stratified and partitioned using a 5-fold cross-validation scheme; within each fold, samples were divided into 80% for training, 10% for validation, and 10% for testing, preserving the class distribution across all splits. Crucially, to strictly prevent data leakage and optimistic local evaluation, the prompt-based data augmentation (described in the Data Augmentation, Section 3.1) was performed *exclusively* on the training subset of each fold *after* the splitting process. The validation and test sets remained completely unaltered, consisting solely of original human discourse. The subsets were saved in JSONL format.

4.2 Training Configurations

Our experiments were structured around an N-way K-shot learning paradigm. Specifically, we configured the tasks as 9-way and 55-shot, meaning each episode contained 9 classes with 55 support samples each, alongside 3 query samples per class.

We employed `BAAI/bge-m3` as our base generic encoder. To retain general semantic knowledge while adapting to the political discourse domain, we froze the first 4 layers of the model.

The system was optimized using the AdamW optimizer with a learning rate of 2×10^{-5} and a linear learning rate scheduler with warmup steps. We trained the model for up to 45 epochs, implementing an early stopping mechanism with a patience of 20 epochs based on the macro-Accuracy metric evaluated on the validation set. To ensure robustness, experiments were conducted using a 5-fold cross-validation approach.

4.3 Evaluation Metrics

Following the task guidelines, the performance of the model is evaluated using primarily macro-averaged metrics to account for the severe class imbalance. We track macro-F1 and Accuracy.

5 Results

5.1 Main Quantitative Findings

Due to the severe class imbalance in the original dataset, our system's training phase utilized the **augmented training subsets** to ensure a balanced distribution of original samples and their LLM-generated variations. We emphasize that all local evaluations (validation and test splits) within our 5-fold cross-validation relied *exclusively* on unaugmented, original data to prevent any synthetic variants from leaking across partitions. For these experiments, we configured the LAQDA system in a 55-shot setting.

As shown in Table 4, the model achieved high and stable performance during the local 5-fold cross-validation phase, yielding an average macro-F1 score of 0.886 and an average accuracy of 0.881.

Fold	Accuracy	Macro-F1
Fold 01	0.864	0.874
Fold 02	0.881	0.885
Fold 03	0.892	0.894
Fold 04	0.878	0.888
Fold 05	0.890	0.888
Average	**0.881**	**0.886**

Table 4: Local 5-fold cross-validation results on the augmented manifold (55-shot setting).

However, when evaluated on the official SemEval hidden test set—which consists exclusively of original, unaugmented human discourse—the system achieved a macro-F1 score of 0.33.

5.2 Error Analysis and Discussion

The discrepancy between our local cross-validation (0.88) and the official test results (0.33) highlights a significant **generalization gap** between synthetic and natural political discourse.

The local performance suggests that the LAQDA encoder successfully mastered the linguistic manifold and rhetorical structures present within the augmented training set, generalizing well to the local unaugmented test folds. However, the severe drop in official metrics indicates an **over-specialization** on the patterns introduced by the generative model during training. While LLM-based augmentation (PromDA) provides a balanced environment, it may produce "canonical" versions of evasion that lack the extreme pragmatic variance found in the fully unseen, real-world presidential interviews of the official test set.

When faced with the official test set, the prototypes refined by the Query-Data-Augmenter (Section 3.2.2) struggled to align with genuine human speech, which is characterized by spontaneous ambiguity, idiosyncratic speaking styles, and high-stakes diplomatic nuances. This confirms that our model became highly optimized for the *synthetic distribution*, but faced a severe domain shift when transitioning to unconstrained real-world data. Specifically, the extreme performance drop indicates that the model learned the "Gemini Pattern"—predictable linguistic artifacts, standardized lengths, and repetitive syntactic structures introduced by the generative LLM—rather than the underlying pragmatic intent of human evasion.

This analysis underscores a critical challenge in political NLP: synthetic data can effectively solve the data scarcity problem mathematically, but it may introduce a **representation bias** that masks the true complexity of human evasion tactics. Future efforts should focus on increasing the adversarial diversity of synthetic samples to better bridge the gap between generated and natural political rhetoric.

6 Conclusion

In this paper, we presented an approach for classifying political question evasions using a Few-Shot Learning architecture (LAQDA) combined with prompt-based data augmentation (PromDA). While our system achieved strong local performance on the augmented dataset, its significantly lower score on the official SemEval test set highlighted the vulnerabilities of LLM-generated synthetic data to domain shift and overfitting. Our findings emphasize that while synthetic data can effectively alleviate class imbalance mathematically, it often struggles to capture the pragmatic subtleties and spontaneous nuances of genuine political discourse. The severe performance drop on the official test set confirms that our model overfitted to the synthetic training data. Rather than learning true human evasion tactics, the model memorized the generative artifacts and structural repetitions of the LLM. Future work must focus on developing more adversarial and highly diverse data augmentation techniques to mitigate this representation bias and better capture the pragmatic variance of genuine political rhetoric.

7 Acknowledgments

We would like to thank the Institute of Informatics (INF) at the Federal University of Goiás (UFG) for providing the infrastructure and academic support necessary to conduct this research, Brazilian Research Agencies FAPEG and CNPq, and also extend our gratitude to the SemEval-2026 Task 6 organizers for providing the dataset and the platform for this evaluation.

References

R. Michael Alvarez and Jacob Morrier. 2025. Measuring the quality of answers in political Q&As with large language models.

Daniel Bashir, George D. Montañez, Sonia Sehra, Pedro Sandoval Segura, and Julius Lauw. 2020. An information-theoretic perspective on overfitting and underfitting. In *AI 2020: Advances in Artificial Intelligence (AJCAI 2020)*, pages 319–331. Springer International Publishing.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2025. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Eric M. Eisenberg. 1984. Ambiguity as strategy in organizational communication. *Communication Monographs*, 51(3):227–242.

Gemini Team et al. 2025. Gemini: A family of highly capable multimodal models.

Gloria Gennaro and Elliott Ash. 2021. Emotion and reason in political language. *The Economic Journal*, 132(643):1037–1059.

Wang Lihong and Gou Weijie. 2018. Analysis of ambiguity. *Advances in Social Science, Education and Humanities Research, volume 132*.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.

Xinyue Liu, Yunlong Gao, Linlin Zong, and Bo Xu. 2024. Improve meta-learning for few-shot text classification with all you can acquire from the tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. 2019. Learning to propagate labels: Transductive propagation network for few-shot learning.

Ning Matthews and Folly Folivi. 2023. Omit needless words: Sentence length perception. *PLoS One*, 18(2):e0282146.

Melina Nikolić and Maja Nikolić. 2022. Multiple hedging in the political interview. *Reci Beograd*, 14:24–49.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Douglas R. Seals and McKinley E. Coppock. 2022. We, um, have, like, a problem: excessive use of fillers in scientific speech. *Advances in Physiology Education*, 46(4):615–620. PMID: 36074921.

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.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Gina Villar and Paola Castillo. 2016. The presence of 'um' as a marker of truthfulness in the speech of tv personalities. *Psychiatry, Psychology and Law*, 24:1–12.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2017. Matching networks for one shot learning.

Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang. 2022. PromDA: Prompt-based data augmentation for low-resource NLU tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4242–4255, Dublin, Ireland. Association for Computational Linguistics.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt.

A Prompt Templates for Data Augmentation

For reproducibility, we provide the core prompt structure used to instruct the `gemini-2.0-flash`

model during our data augmentation phase. The variables in brackets (e.g., <evasion technique>) were dynamically replaced according to the target class being generated.

A.1 General Prompt Template

General System Prompt

System Role: You are an expert in political discourse analysis and linguistics. Your task is to generate synthetic question-and-answer pairs simulating a political interview.
Instructions:

- Keep the original evasion technique <evasion technique> seen in the few-shot examples.
- The idea of <evasion technique> is <definition of the evasion technique>.
- Preserve the core political intent and meaning, but rephrase creatively without losing the evasion technique used.
- Introduce diverse speaking styles (formal, colloquial, rhetorical, defensive, assertive, vague, etc.).
- Maintain an answer length <length percentage> above or less <characters value> characters.
- Ensure logical consistency between interviewer and president.
- Develop a mostly <sentiment> sentiment returning the output.
- On <fillers percentage> of the generated samples try adding <number of fillers> sentence usual fillers and hedges.
- Avoid factual errors, hallucinations or unrelated topics.

A.2 Instantiated Example: Dodging Class

Instantiated Prompt — *Dodging*

System Role: You are an expert in political discourse analysis and linguistics. Your task is to generate synthetic question-and-answer pairs simulating a political interview.
Instructions:

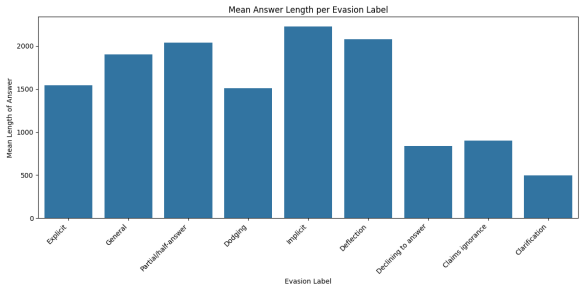
- Keep the original evasion technique *Dodging* seen in the few-shot examples.
- The idea of *Dodging* is *a response in which the speaker deliberately shifts the topic of the question, addressing a related but distinct issue rather than the one explicitly asked, without overtly refusing to answer*.
- Preserve the core political intent and meaning, but rephrase creatively without losing the evasion technique used.
- Introduce diverse speaking styles (formal, colloquial, rhetorical, defensive, assertive, vague, etc.).
- Maintain an answer length of approximately *1800 characters*.
- Ensure logical consistency between interviewer and president.
- Develop a mostly *neutral* sentiment in the returned output.
- On *50%* of the generated samples try adding *2–3* sentence-level fillers and hedges (e.g., "you know," "well," "I mean," "sort of").
- Avoid factual errors, hallucinations or unrelated topics.

Few-shot examples: five *question / answer* pairs labeled as *Dodging*, sampled from the original training partition of the current fold.

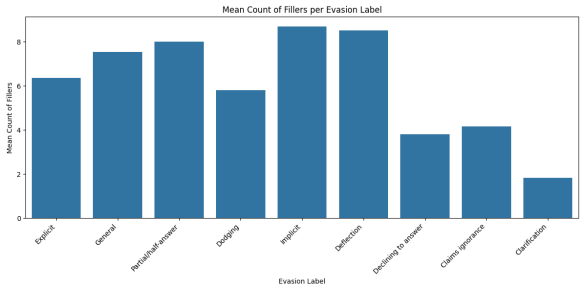
B Hardware and Infrastructure

To ensure efficient training and accommodate the computational demands of the `BAAI/bge-m3` encoder alongside the Transductive QDA Sampler, all experiments, including the 5-fold cross-validation pipeline, were executed on an NVIDIA DGX node equipped with H100 Tensor Core GPUs.

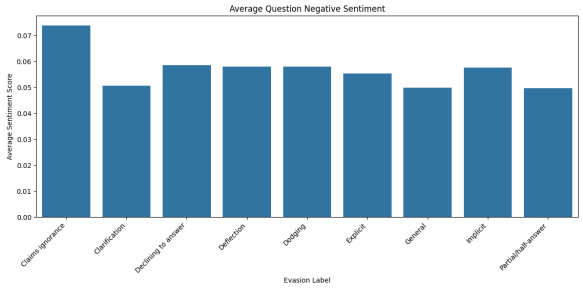
C Sentimental Analysis



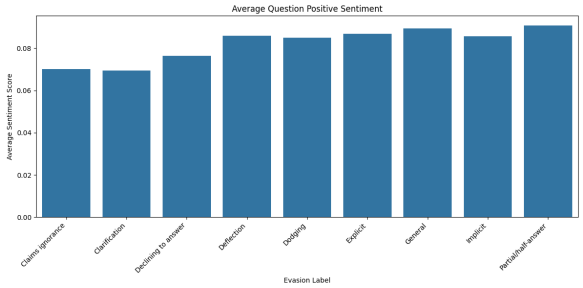
(a) Average answer length.



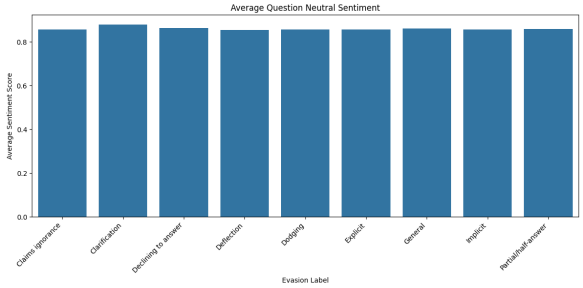
(b) Fillers and hedges.



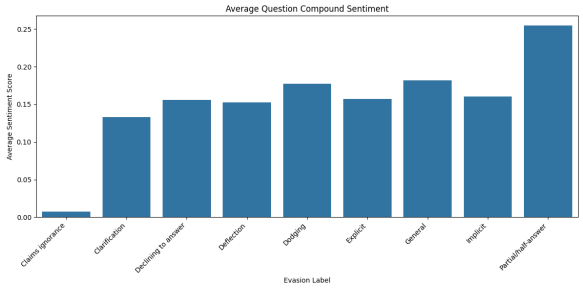
(c) Question negative sentiment.



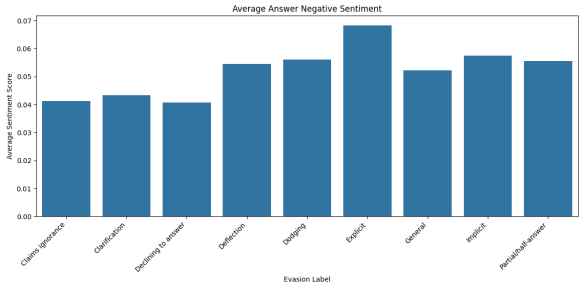
(d) Question positive sentiment.



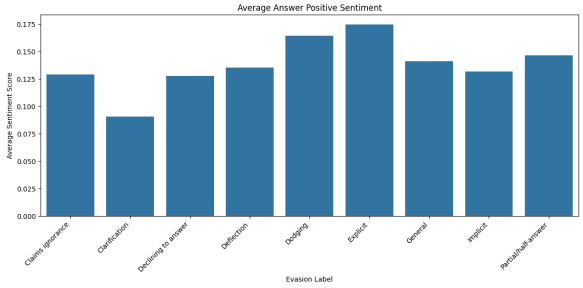
(e) Question neutral sentiment.



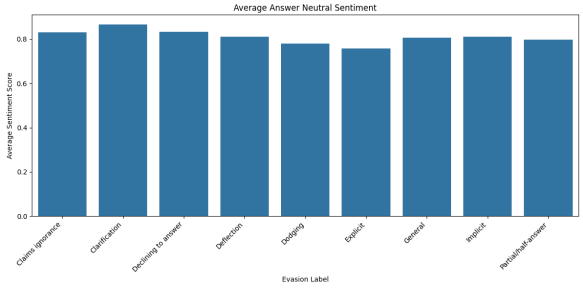
(f) Question compound sentiment.



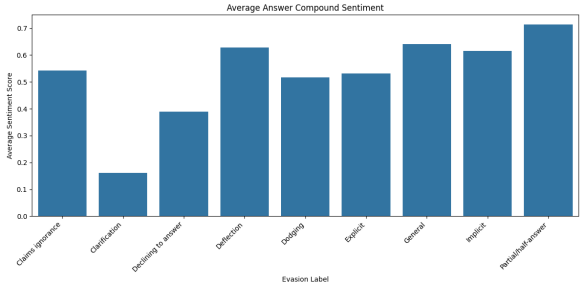
(g) Answer negative sentiment.



(h) Answer positive sentiment.



(i) Answer neutral sentiment.



(j) Answer compound sentiment.

Figure 5: Overview of the training set.