PolERo: Studying Political Evasion in Romanian

Anonymous ACL submission

Abstract

Political evasion refers to responses that engage with a question while withholding the requested information. Recent NLP work frames political evasion as a classification task using a two-level taxonomy of response clarity and fine-grained evasion strategies. Existing work on response clarity and evasion classification is limited to English, leaving open whether the taxonomy and model behavior transfer across languages and political contexts. We introduce **PolERo**, a dataset of 3,574 human-annotated question–answer pairs extracted from official transcripts of five Romanian presidents, labeled under the same taxonomy and annotation protocol as the English corpus. We evaluate multiple classification approaches on both datasets under matched conditions, including TF-IDF baselines, fine-tuned encoder models, a proposed sliding-window encoder, and zero/few-shot LLM prompting. We study cross-lingual transfer through joint bilingual training and machine-translation-based data augmentation. Our results indicate that open-weight LLMs outperform fine-tuned encoders in both languages, cross-lingual transfer is asymmetric, and ambivalent evasion categories involving pragmatic cues remain the main challenge across all model families.

1 Introduction

Political interviews often contain responses that are relevant to the interviewer's questions while at the same time remaining non-responsive. Such responses may involve implicit answers, partial replies, topic shifts (Msagalla and Visser, 2021), deflections, or strategic ambiguity (Bach et al., 2025), making it difficult to determine whether the requested information was actually provided. In political science and discourse analysis, these phenomena are commonly studied under the notions of **equivocation** and **evasion** (Bavelas et al., 1988; Bull, 1994; Clayman, 2001; Bull, 2003; Rasiah, 2010; Romaniuk, 2013; Bull and Strawson, 2020).

Beyond political communication, response clarity is also relevant to broader NLP problems including dialogue understanding, question answering, and implicitly for creating better user-facing technologies. Despite recent progress in political discourse processing (Glavaš et al., 2019; Huguet Cabot et al., 2020; Afli et al., 2024), response clarity remains largely unexplored in NLP. Existing work has focused primarily on question answerability (Rajpurkar et al., 2018; Min et al., 2020), responder intent (Ferracane et al., 2021), or broader discourse phenomena in political speech (Subramanian et al., 2019; Reinig et al., 2024; Verma et al., 2026), while leaving open the problem of determining whether a response adequately addresses a question. Unlike traditional QA settings (Rajpurkar et al., 2018; Min et al., 2020; Clark et al., 2020a), response clarity requires reasoning about pragmatic alignment rather than factual correctness or semantic similarity. Consequently, response clarity requires modeling information sufficiency (Jain and Garimella, 2026) and discourse structure (Mann, 1984), extending beyond standard answerability formulations.

Recently, Thomas et al. (2024) introduced the first NLP dataset and taxonomy of 3,400 question-answer (QA) pairs from U.S. presidential interviews. The dataset served as a reference for SemEval-2026 Task 6: CLARITY (Thomas et al., 2026) a task that aims to detect and classify response ambiguity in political discourse. However, evidence for response clarity classification currently remains restricted to English U.S. political interviews. Furthermore, response clarity is not a purely linguistic phenomenon: it is also shaped by institutional context, interaction format, and political communication practices. Currently, no resource targets response clarity or evasion classification in Romanian political speech. To the best of our knowledge, such resources are also absent for non-English languages more broadly.

To address this, we introduce PolERo (Political Evasion in Romanian), a dataset of 3,574 human-annotated question-answer pairs extracted from official transcripts covering five Romanian presidents. PolERo adopts the exact annotation protocol and two-level taxonomy proposed by Thomas et al. (2024), enabling controlled comparison between English and Romanian political interviews.

Romanian is an Eastern Romance language with historically limited NLP resources. Recent efforts have significantly expanded this landscape, introducing general evaluation benchmarks and foundational corpora (Dumitrescu and Avram, 2020; Dumitrescu et al., 2021; Rogoz et al., 2026), alongside Romanian-adapted language models (Masala et al., 2020; Dumitrescu et al., 2020; Dima et al., 2024; Masala et al., 2024). Furthermore, Romanian political communication has attracted sustained interest in discourse studies and political linguistics with a focus on political opinion mining (Gîfu and Cristea, 2013; Vasilescu et al., 2024) and hate-speech discourse (Gheorghiu and Praisler, 2022; Dinu et al., 2025).

Nevertheless, despite this progress, no prior datasets or analyses target response clarity or evasion classification in any type of speech. This makes PolERo, to the best of our knowledge, the first such resource for a non-English language.

This paper makes several contributions:

- The release of PolERo,[1] a human-annotated dataset for response clarity and evasion classification in a Romanian political corpus.
- Several classification results for both PolERo and the English dataset of Thomas et al. (2024) under matched conditions, covering several types of encoder models, a multi-head encoder for long text classification, and different LLM prompting strategies.
- A cross-lingual transfer analysis examining the consistency of evasion patterns and model behavior across the two corpora, considering low-data scenarios and machine-translation-based data augmentation.

[1]The dataset and code are released under the CC BY 4.0 License and are currently available as supplementary materials. They will be made available in a public repository upon publication.

2 Dataset and Annotation

2.1 Taxonomy

PolERo adopts the two-level response clarity taxonomy of Thomas et al. (2024) focused on the distinction between clarity and ambiguity in political responses (see. Figure 1). The first level assigns each response one of three clarity labels: *Clear Reply* – the information is given in the requested form, *Ambivalent* – a valid answer is given but admits multiple interpretations, or *Clear Non-Reply* – the speaker openly refuses to share the information. The second level divides these classes into nine fine-grained evasion categories. *Clear Reply* contains only *Explicit* information stated in the requested form.

The *Ambivalent* category includes the cases where 1. *Implicit* information is conveyed without being explicitly stated, 2. *General* information is provided, lacking the requested specificity, 3. *Partial* half-answers are provided, 4. *Dodging* or ignoring the question altogether, and 5. *Deflection* of the topic by shifting focus to a different point.

The final category named *Clear Non-Reply* includes *Declining to Answer* in an explicit manner while acknowledging the question, *Ignorance* when respondents admit not knowing the answer, and *Clarification* when respondents ask for more details.

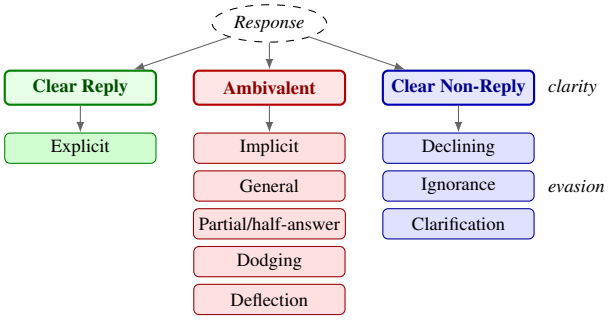

Figure 1: Two-level taxonomy mapping.

An expert in linguistics, working professionally as an authorized translator, adapted the initial English taxonomy for Romanian, fixing category definitions and boundary cases. Appendix A reproduces the Romanian annotation guidelines.

2.2 Annotation Protocol

We build a screening quiz of 15 QA pairs spanning ambiguous evasion categories, where each item is labeled by an expert to serve as a benchmark reference. The quiz is used as a screening step to

ensure that the three annotators understand the taxonomy and can consistently handle borderline and difficult cases. All annotators are native Romanian women aged 22–24 with undergraduate training in psychology.

Annotators complete a training phase that covers all nine categories of taxonomy (§2.1) using worked examples. Each instance requires a single annotation decision: assigning one of the nine evasion labels to the response. The three-way clarity label is then inferred by traversing the taxonomy upward. The annotation process spanned five weeks. During this period, we held three group review sessions with the expert to resolve disagreements and clarify boundary cases. Mean annotation time was 69 seconds per instance, totaling approximately 27 hours per annotator. All annotators were compensated for their work (see section 8).

We divide the dataset into a training set and a test set of 296 examples. A single annotator labels each training instance, with the workload evenly distributed between the three. All three independently label the test split to support inter-annotator agreement measurement and evaluation using multiple annotations. For 2.4% of test instances where all three annotators assigned different evasion labels, the expert assigned the final label.

2.3 Data Collection and Preprocessing

We collect official transcripts from the Romanian presidential websites,[2] covering official press conferences, interviews, and declarations from five Romanian presidents: Ion Iliescu, Traian Băsescu, Klaus Iohannis, Ilie Bolojan, and Nicușor Dan. The corpus spans 25 years between 2001–2026.

In such contexts, it is common for a single interviewer question to contain multiple sub-questions. The initial pool of extracted data contained 6,234 raw question–response candidates, comprising 8,233 sub-questions in total. For the current task, we enforce a one-to-one question–response mapping, and we use GPT-5.4 to flag the set of instances where more than one question is asked in a single instance. As such, we remove 2,053 multi-part entries and 607 entries containing no extractable question. During the annotation process, the evaluators were instructed to flag any remaining double-barreled questions in the labeling platform (see the appendix B.1). No such instances were reported, indicating that no remaining multi-part questions were present in the retained corpus. The retained corpus consists of **3,574** single-part QA pairs. The final dataset consists of 3,278 training and 296 test instances.

[2] https://www.presidency.ro and https://archive.presidency.ro

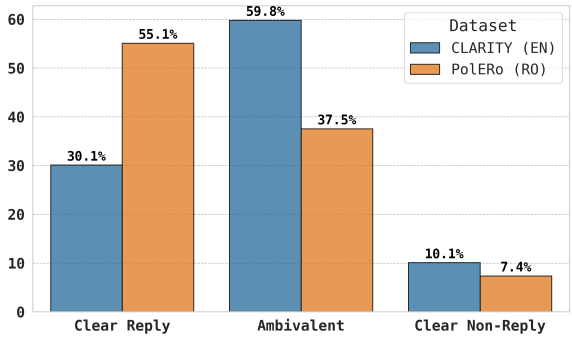

Figure 2: Overall clarity distribution for PolERo compared to the English CLARITY dataset.

2.4 Dataset Statistics

Label distributions are similar across splits. In the training set, *Clear Reply* accounts for 54.9% of instances, *Ambivalent* for 37.7%, and *Clear Non-Reply* for 7.4%. The test set shows a comparable distribution (56.8%, 36.1%, 7.1%).

Figure 2 compares the overall clarity label distribution of PolERo with the English CLARITY dataset (Thomas et al., 2024). *Clear Reply* is the majority class in the Romanian dataset (55.09%), while the English corpus is primarily composed of ambivalent *Ambivalent* responses (59.80%). An exploratory analysis is provided in the Appendix C. The analysis reveals variation in response clarity across speakers, communicative settings, and topics. Unlike the English CLARITY corpus, Romanian presidential discourse is dominated by *Explicit* answers (55.09%), although this proportion decreases substantially in more formal contexts such as declarations and press conferences. Evasion strategies also vary across presidents and topics: earlier speakers rely more heavily on *Dodging*, whereas post-2014 presidents exhibit more balanced distributions with higher rates of *General*, *Deflection*, and *Declining to answer*. Finally, the length of the answer aligns with established theories of verbose evasion (Bavelas et al., 1988; Rasiah, 2010; Bull and Strawson, 2020), with indirect strategies such as *General*, *Deflection*, and *Partial* responses tending to be substantially longer than direct responses.

2.5 Annotation Quality

To verify the performance of the three annotators, the expert independently labeled a benchmark set of 85 instances. Against this gold standard, the annotators achieved accuracies of 95.3%, 95.3%, and 94.1% on the clarity task, and 89.4%, 88.2%, and 84.7% on the evasion task (Appendix B.3).

Inter-annotator agreement on the test set, measured using Fleiss's κ (Fleiss, 1971), is 0.843 for clarity and 0.678 for evasion, corresponding to near-perfect and substantial agreement on the scale of Landis and Koch (1977). At the category level, agreement is highest for *Clarification* ($\kappa = 1.000$), *Claims Ignorance* ($\kappa = 0.881$), and *Explicit* ($\kappa = 0.858$), and lowest for *Deflection* ($\kappa = 0.364$) and *General* ($\kappa = 0.464$). A detailed analysis of the annotation is provided in Appendix B.

3 Experimental Setup

3.1 Task Formulation

We report macro-averaged F1 on the test split of each dataset as the primary metric across all experiments. We evaluate the models on two classification tasks derived from the two-level taxonomy (§2.1): (1) *clarity classification*, a three-class task over the coarse-grained labels; and (2) *evasion classification*, a nine-class task over the fine-grained labels. For the three-class clarity task, we also compute a macro-F1 score by deterministically (3) *mapping* each predicted evasion label to its parent clarity class. Thus, we can directly compare models trained specifically for the three-class task against those whose coarse-grained labels are inferred through the fine-grained objective.

3.2 Models

We establish baselines across four model families on both the English dataset of Thomas et al. (2024) (CLARITY) and PolERo, training and evaluating each language independently. Input formats, hyperparameters, and prompt templates are reported in Appendix D.2.

Traditional classifier. A TF-IDF logistic regression classifier serves as a non-neural lower bound.

Single-head (SH) encoder fine-tuning. We fine-tune several basic encoder configurations, each trained independently per task. On English we evaluate RoBERTa-large (Liu et al., 2019), ModernBERT-large (Warner et al., 2025), ELECTRA-large (Clark et al., 2020b), BERT-large-cased (Devlin et al., 2019) and XLM-RoBERTa-large (Conneau et al., 2020). The same six architectures are fine-tuned on PolERo, with the addition of **Ro**BERT-large (Masala et al., 2020),[3] and Romanian BERT (Dumitrescu et al., 2020), both pretrained on Romanian corpora, and multilingual mBERT-base (Devlin et al., 2019).

LLM prompting. We evaluate six LLMs under zero-shot and few-shot prompting without fine-tuning: GPT-5.4, Llama-3.3-70B-Instruct, Qwen3.6-35B-A3B, Gemma-4-31B-it, DeepSeek-V4-Pro, and gpt-oss-120b.

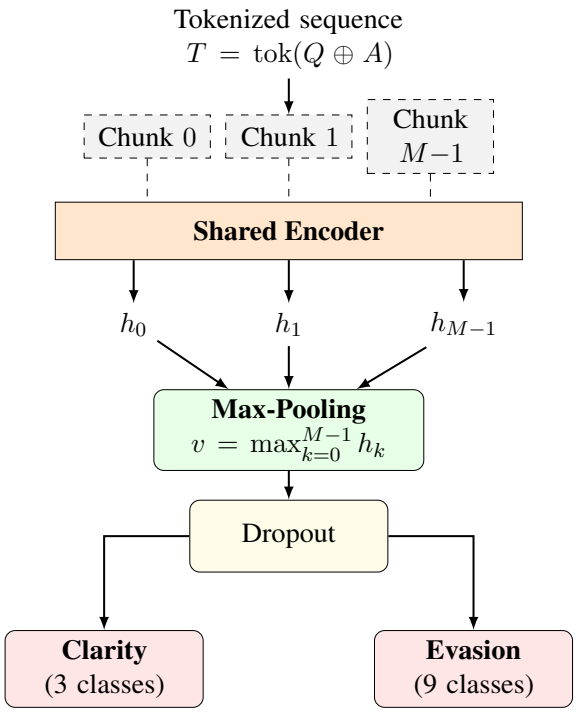

Figure 3: Multi-head architecture. The tokenized concatenated input ($Q \oplus A$) is split into overlapping chunks, each chunk is encoded by a shared encoder, chunk representations are aggregated via element-wise max-pooling, and two task-specific heads predict clarity (3 classes) and evasion (9 classes).

Multi-head (MH) encoder architecture. We additionally evaluate a custom multi-head architecture designed specifically for this task, featuring a shared encoder and two linear heads that predict clarity and evasion jointly. A representation of the architecture is shown in Figure 3. To process long question–answer pairs without truncation,

[3] We emphasize the first two letters in name of Romanian BERT large: **Ro**BERT-large to increase readability and distinguish it more easily from the RoBERTa-large model, both used extensively in this work.

sequences are segmented into overlapping chunks and encoded independently. The chunk representations are aggregated via element-wise max-pooling into a single response vector, which is passed to a 3-class clarity head and a 9-class evasion head. The training objective is the unweighted sum of the two cross-entropy losses. Models are trained with 7-fold stratified cross-validation, and fold predictions are ensembled by probability averaging at inference time.

We train RoBERTa-large (matching the original configuration), XLM-RoBERTa-large (added to support the cross-lingual experiments below), and **Ro**BERT-large, a Romanian-only encoder model. Full architectural details and mathematical formulations together with component ablations isolating the contribution of chunked pooling, joint supervision, and ensembling are reported in Appendix E.

The results in this work are reported on the English dataset using the *development* set from the SemEval 2026 CLARITY Shared Task, since the official test of the task is not yet publicly available. The multi-head RoBERTa-large achieves one of the strongest scores (0.8) on the SemEval 2026 CLARITY task (Thomas et al., 2026) among encoder-only models, especially given the size under 400 million parameters.[4] However, on the development set, the same model obtains a smaller .713 score for the 3-way clarity task prediction.

4 Results and Discussion

Table 1 reports the macro-F1 scores for the strongest models from each approach. A complete set of results for all models, cross-lingual setups, and prompting strategies is provided in Appendix F.

Traditional classifiers. TF-IDF + logistic regression reaches macro-F1 .422/0.453 on English/Romanian 3-class clarity detection, confirming that surface-level lexical features alone are insufficient. Additional per-class error patterns are reported in Appendix D.

Monolingual encoder fine-tuning. RoBERTa-large is the strongest single-head English encoder model for the 3-class clarity (C) identification task (.580) and the 9-class fine-grained evasion (E) detection task (.481). The mapped (M) fine-grained task significantly outperforms (.656) the simple clarity identification task showing that it is better to train on a more fine-grained taxonomy and merge than to train directly on fewer un-balanced classes.

Unsurprisingly, on Romanian, this model performs significantly worse than encoder models pre-trained on Romanian, such as **Ro**BERT-large (Masala et al., 2020). The latter is the strongest single-head model on PolERo (.504 evasion, .701 mapped). Detailed encoder comparisons are available in Appendix F.

	EN (CLARITY)			RO (PolERo)		
Model	C	E	M	C	E	M
TF-IDF + LR	.422	.301	.441	.453	.312	.472
Best Encoders (Fine-Tuned)						
RoBERTa (SH)	.580	.481	.656	.646	.428	.613
RoBERTa (MH)	**.713**	.568	**.731**	.696	.501	.652
XLM-R. (SH)	.579	.419	.564	.680	.365	.686
XLM-R. (MH)	.665	.475	.644	.694	.508	.706
RoBERT (SH)	–	–	–	.688	.504	.701
RoBERT (MH)	–	–	–	**.729**	.545	**.763**
Cross-Lingual Transfer (XLM-R, MH)						
EN → RO	–	–	–	.635	.399	.655
RO → EN	.551	.398	.527	–	–	–
Joint Training (EN+RO)	.679	483	.642	**.723**	**.570**	.717
Best LLMs (Few-Shot Prompting)						
Gemma-4-31B-it	**.721**	.583	.662	.672	.619	.738
Qwen3.6-35B-A3B	.669	.544	.747	.746	**.682**	.760
GPT-5.4	.703	**.626**	**.761**	**.763**	.666	**.763**

Table 1: Summary of test-set macro-F1 for the highest-performing models across all tested approaches. **C**: Clarity (direct 3-class prediction). **E**: Evasion (direct 9-class prediction). **M**: Mapped (3-class clarity prediction obtained by mapping the predicted 9-class evasion label upward through the taxonomy). **SH**: Single-head, **MH**: Multi-head. Best scores are highlighted in **bold**. LLMs obtain the strongest results for the fine-grained Evasion prediction, while smaller-sized multi-head encoders remain competitive for predicting response clarity.

Multi-head fine-tuning. The custom multi-head model architecture choice improves performance across all metrics compared to the corresponding single-head model on both datasets. Multi-head RoBERTa-large is the strongest English encoder configuration, outperforming its single-head counterpart by 13 macro-F1 points on direct clarity and 9 on evasion, and surpassing GPT-5.4 on direct clarity prediction. On Romanian, the best non-LLM result in the table is the multi-head **Ro**BERT-large, which improves over its single-head Romanian-specific counterpart by 4 points on clarity, 4 on evasion, and 6 on mapped clarity. These gains are consistent with the dual clarity+evasion tasks act-

[4] According to the official report by Thomas et al. (2026), all top-scoring systems in the range of 0.82–0.89 relied on proprietary LLMs.

ing as a regularizer: gradients from the coarser clarity supervision help stabilize predictions for rare evasion categories, and vice-versa.

Multi-head XLM-RoBERTa-large shows smaller gains on both languages than the language-specialized backbones, showing that multilingual pretraining sacrifices some within-language performance for broader cross-language coverage. As expected, models pretrained specifically for a given language consistently outperform multilingual models on that language: **Ro**BERT-large achieves higher scores than XLM-RoBERTa-large on all Romanian evaluation metrics, while RoBERTa-large similarly outperforms XLM-RoBERTa-large on the English benchmarks.

Mapped vs. direct clarity. The two-step strategy of predicting the nine fine-grained evasion categories and then projecting upward through the taxonomy outperforms the direct clarity head for RoBERTa-large on English, for **Ro**BERT-large on Romanian, and for nearly every LLM configuration on both datasets. This is consistent with the results reported at the Shared Task (Thomas et al., 2026), where *the winning reverse derivation strategy proved to be the single most effective approach across the entire competition*.

Four encoders do not benefit from this approach: mBERT-base, BERT-large-cased, ELECTRA-large, and ModernBERT-large. In each case, the evasion head assigns most predictions to *Explicit* and *General*, causing the projection step to overpredict *Clear Reply* at the expense of *Ambivalent* and *Clear Non-Reply*. ELECTRA-large on the Romanian split is the clearest example: the model predicts *Explicit* for every instance, yielding a macro-F1 of .084, and the mapped clarity score drops to .24 because *Ambivalent* and *Clear Non-Reply* are never predicted. ModernBERT-large shows the same pattern, though less severely, with two categories still receiving zero recall on Romanian (*Claims ignorance* and *Partial/half-answer*). The two-step strategy therefore depends on the backbone being able to assign a meaningful probability mass to rare categories. Otherwise, the projection step amplifies the model's bias toward the majority class rather than correcting it.

LLM prompting. LLMs achieve the best results across all metrics on both datasets, though their ranking varies by task. Gemma-4-31B-it leads few-shot direct clarity on English, Qwen3.6-35B-A3B leads few-shot evasion on Romanian, and GPT-5.4 achieves the best mapped clarity scores on both. Few-shot examples improve the mapped metric for most LLMs by helping the models better distinguish between closely related categories such as *Dodging/Deflection*. Llama-3.3-70B is the exception: its mapped clarity score drops under few-shot prompting on English, suggesting that the additional examples disrupt rather than improve its internal taxonomy mapping.

LLM reasoning and data contamination. The performance gap between LLMs and fine-tuned encoders likely stems from multiple factors. Firstly, LLMs benefit from broad parametric world knowledge stored in large billion-parameter sized networks, which helps to resolve entity-dependent evasion. Secondly, all evaluated LLMs (except Llama-3.3-70B) are prompted with reasoning enabled, allowing explicit deliberation over pragmatic intent. Notably, Llama-3.3-70B lacking extended reasoning is also the weakest LLM in both languages. Thirdly, we cannot rule out that these models might have been exposed to the data (which is publicly available) during their pre-training phase. To assess whether the advantage of LLMs is influenced by data contamination, we explore the CoDeC framework, Contamination Detection via Completion (Zawalski et al., 2026), on three open-source LLMs: Qwen3.6-35B-A3B, DeepSeek-V4-Pro, and Gemma-4-31B-it. CoDeC scores are extracted solely from the text, yielding contamination percentages on the English test set of 37%, 38%, and 45%; while on the Romanian test set, the corresponding scores are 43%, 41%, and 39%. All six values fall below the 80% threshold that would indicate a clear case of data contamination, according to the authors. This can be regarded as an indicator that LLMs possess a genuine expressivity in extracting the pragmatic information required for this task on both Romanian and English political interviews.

5 Cross-Lingual Transfer

We assess cross-lingual generalization using *only the multi-head (MH) architecture*, instantiated with both RoBERTa-large and XLM-RoBERTa-large. In addition to the monolingual setup, each model is trained in five cross-lingual configurations:

- **EN** $\rightarrow$ **RO**: trained on English, evaluated on Romanian.
- **RO** $\rightarrow$ **EN**: trained on Romanian, evaluated on English.

- **EN + RO → RO**: trained on the concatenation of both datasets, evaluated only on Romanian;
- **EN + RO → EN**: trained on the concatenation of both datasets, evaluated only on English.

The first two configurations explore the ability to generalize across the two languages and political spheres (U.S. vs. Romania), while the remaining configurations explore the possibility of increasing the predictive power using multilingual data augmentation.

Generalizability. Training on English data and predicting on Romanian yields a significant 6 to 7 point drop (.694 → .635) using both XLM- and RoBERTa-large models. Training in the opposite direction, in Romanian → English, results in a worse performance degradation of 11 to 14 points, the highest affected model being RoBERTa (.713 → .572). In both scenarios the performance degradation makes the MH models similar to other single-head models such as mBERT-base, BERT-large-cased, ELECTRA-large, or ModernBERT (see Table 8 containing complete experimental results in the Appendix).

The Romanian-trained model strongly overpredicts *Clear Reply*, misclassifying 54% of English *Ambivalent* answers as *Clear Reply*, and similarly maps most English *Implicit* and *Deflection* responses to *Explicit*. These results indicate that cross-lingual predictions do not lead to catastrophic losses and that the task may benefit from additional multilingual data.

Multilingual augmentation. As shown in Table 1, combining the two datasets and predicting on each test set does not lead to any clear improvement: in Romanian, the results remain comparable to the MH approach, with slightly better evasion detection abilities.

One of the best non-LLM configurations for Romanian is joint EN+RO training with XLM-RoBERTa-large, which achieves the highest Romanian evasion score among fine-tuned models (0.570). Two factors may contribute to this improvement. Firstly, joint training regularizes rare categories: Romanian examples such as *Clarification* or *Claims ignorance* are reinforced by similar English instances, reducing the data scarcity that causes failures. For example, *Declining to answer* reaches perfect recall on the Romanian test split under joint training. Secondly, joint training encourages the shared encoder to model evasion in a cross-lingual way rather than relying on specific language lexical cues. The English dataset contains more *Ambivalent* evasive answers, where evasion is expressed through topic shifts and intent rather than specific words. This provides additional supervision for learning pragmatic behavior. When trained jointly, the model may transfer this ability to Romanian, improving its ability to detect evasion without clear lexical signals. As a result, Romanian evasion scores increase by 0.062 over the in-language XLM-RoBERTa-large baseline, whereas the English scores show only an insignificant improvement (+0.008).

RoBERTa-large does not benefit in the same way. Its tokenizer and pretraining are optimized for English, so adding Romanian data introduces more noise than useful signal, causing English evasion performance to drop from 0.568 to 0.503. Overall, the English performance drops consistently across all criteria when doing joint training.

5.1 Machine Translation

As a secondary analysis, we extend the cross-lingual setup with machine-translated training and test data in order to reduce the biases of the multilingual encoders. All translations are generated with NLLB-200-distilled-600M (NLLB Team et al., 2022). Appendix F.5 reports a semi-automatic analysis on the quality estimation of the translations using XCOMET-XL (Guerreiro et al., 2024). The model indicates higher MT quality for RO→EN than EN→RO, although manual inspection suggests many EN→RO translations flagged as errors are acceptable, with pragmatic categories such as *Deflection* remaining the most error-prone in both translation directions.

Translating the English data into Romanian partially recovers cross-lingual clarity performance but provides limited gains for evasion detection, according to the results presented in Table 2. Pragmatically driven categories such as *Deflection* and *Dodging* suffer the largest degradation under MT augmentation and also receive the lowest translation quality scores, whereas more formulaic categories such as *Clarification* and *Declining to answer* translate reliably.

The transfer asymmetry between English and Romanian persists under MT, indicating that surface-form mismatch is not the main bottleneck. Instead, differences in class distributions and different discourse structures shaped by historical practices, geopolitical contexts, and topics of interest, make

Model	EN (CLARITY)			RO (PolERo)		
	C	E	M	C	E	M
MT Cross-Lingual Transfer (XLM-R, MH)						
Low-data scenario; only test set available						
EN → RO	–	–	–	.635	.399	.655
EN2RO → RO	–	–	–	.665	.412	.706
RO → EN	.551	.398	.527	–	–	–
RO2EN → EN	.477	.358	.439	–	–	–
Multilingual vs MT augmentation						
RO + EN2RO → RO	–	–	–	.708	.492	.681
EN + RO2EN → EN	.621	.460	.647	–	–	–
Joint Training (EN+RO)	.679	483	.642	**.723**	**.570**	.717
Best LLMs (Few-Shot Prompting)						
Gemma-4-31B-it	**.721**	.583	.662	.672	.619	.738
Qwen3.6-35B-A3B	.669	.544	.747	.746	**.682**	.760
GPT-5.4	.703	**.626**	**.761**	**.763**	.666	**.763**

Table 2: Summary of test-set scores for the cross-lingual models augmented with machine translated (MT) data. In low-data scenarios it is best to use few shot prompting and LLMs. When data is available, joint multi-lingual training achieves better results than monolingual training using machine translated data.

the Romanian data less useful for detecting evasion in higher-resourced languages such as English. Consequently, English-to-Romanian transfer is stronger, and LLMs outperform encoder models most clearly on evasion tasks, suggesting that pragmatic intent rather than language-specific form remains the main challenge.

5.2 Remarks

Based on our Romanian case study, we propose the following takeaways to guide future assessment of political evasion in new languages:

1. In low-resource settings, when creating a dataset for a new language such as Romanian, it is advisable to first benchmark the dataset using few-shot prompting with available LLMs, preferably open-weights models.
2. If LLMs perform poorly for a particular language, translating existing datasets with NLLB-200 or similar tools and training cross-lingual encoders may be a viable alternative, although factors related to data representativeness, geopolitical factors, and local histories must be carefully taken into account.
3. If a sufficiently large training set can be constructed for the target language (i.e., at least 3,000 samples) and existing monolingual pre-trained encoders are available, then a strong strategy is to use the multi-head approach described in Section 3; else, if monolingual encoders are not available, then joint multi-lingual training may be more beneficial than augmenting the data with machine-translated examples.
4. The strong performance of LLMs must be weighed against cost and efficiency; our proposed multi-head encoder proved competitive with LLMs on the clarity classification task while using fewer than 500M parameters; at the same time, the encoders have limitations in terms of fine-grained evasion detection.
5. The distribution of ambiguous responses can have a significant effect on the generalizability of cross-lingual models.

6 Conclusions

In this work we introduce *PolERo*, a novel Romanian dataset of 3,574 question–answer pairs from presidential transcripts annotated using a two-level taxonomy of evasion. We compare multiple models on an existing English data and on our proposed Romanian dataset under matched conditions, covering traditional classifiers, standard encoder fine-tuning, a custom multi-head architecture for joint clarity and evasion prediction, and large language models under zero-shot and few-shot prompting.

Our results corroborate previous findings (Thomas et al., 2026) that prompting modern LLMs outperforms fine-tuned encoders and further show that this also holds for a medium-resource language such as Romanian using open-weights models. We propose a sliding-window and max-pooling approach with two classification heads that significantly increases the predictive power of encoder models, making them competitive with LLMs for the three-class clarity classification task, while using only a fraction of the memory footprint. However, for the nine-class fine-grained evasion task, larger models achieve better predictive performance.

Finally, we investigate cross-lingual transfer from the perspective of multilingual training versus monolingual training with machine-translated texts. Our analyses consistently reveal an asymmetric relation: English-trained models generalize better to Romanian than the reverse. Joint multilingual training yields good cross-lingual performance for Romanian, while ambivalent categories remain the main challenge across all model families, as they depend on pragmatic intent rather than surface form.

7 Limitations

PolERo covers a single non-English language. Whether the taxonomy, annotation difficulty, and model behavior observed here generalize to other languages and political contexts remains an open question.

The annotation protocol assigns a single annotator per training instance, with triple annotation reserved for the test split. Evasion classification requires interpreting communicative intent, which can vary across annotators. Our annotators have undergraduate training in psychology rather than political science or linguistics, which may affect interpretation of borderline cases. Full triple annotation with domain experts would likely reduce this uncertainty, but at substantially higher cost.

8 Ethical Considerations

Source data consists of official transcripts published on the public archives of the Romanian presidency. The speakers are public officials whose statements were issued in their official capacity; no private communications, personal data, or non-public records are included.

The three annotators were each paid the equivalent of 80 EUR for approximately 27 hours of labeling work distributed over five weeks. The linguistic expert contributed voluntarily. Before participation, annotators were informed of the task purpose, the political nature of the content, and the option to withdraw at any point without penalty.

Clarity and evasion classification of political speech can be used to selectively frame or attack individual speakers. We release PolERo for research on response clarity in political discourse and discourage its deployment as a standalone instrument of public accountability. Model predictions are not ground truth: as the inter-annotator agreement statistics show, even trained human annotators disagree on a fraction of instances. Automated labels should not be treated as definitive characterizations of any individual speaker's behavior.

Finally, we acknowledge the computational cost associated with the experimental pipeline presented in this work. All fine-tuning and inference experiments were conducted on a single NVIDIA H100 80GB GPU, with a total estimated budget of approximately 180 GPU hours.

Acknowledgments

References

Haithem Afli, Houda Bouamor, Cristina Blasi Casagran, and Sahar Ghannay, editors. 2024. *Proceedings of the Second Workshop on Natural Language Processing for Political Sciences @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.

Parker Bach, Carolyn E Schmitt, and Shannon C McGregor. 2025. Let me be perfectly unclear: strategic ambiguity in political communication. *Communication Theory*, 35(2):96–106.

Janet Beavin Bavelas, Alex Black, Lisa Bryson, and Jennifer Mullett. 1988. Political equivocation: A situational explanation. *Journal of Language and Social psychology*, 7(2):137–145.

Peter Bull. 1994. On identifying questions, replies, and non-replies in political interviews. *Journal of language and social psychology*, 13(2):115–131.

Peter Bull. 2003. *The Microanalysis of Political Communication: Claptrap and Ambiguity*. Routledge, London and New York.

Peter Bull and Will Strawson. 2020. Can't answer? won't answer? an analysis of equivocal responses by theresa may in prime minister's questions. *Parliamentary Affairs*, 73(2):429–449.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020a. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020b. Electra: Pre-training text encoders as discriminators rather than generators. *Preprint*, arXiv:2003.10555.

Steven E Clayman. 2001. Answers and evasions. *Language in society*, 30(3):403–442.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George-Andrei Dima, Andrei-Marius Avram, Cristian-George Craciun, and Dumitru-Clementin Cercel. 2024. RoQLlama: A lightweight Romanian adapted language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4531–4541, Miami, Florida, USA. Association for Computational Linguistics.

Anca Dinu, Andreea C. Moldovan, and Adina Marincea. 2025. AntiSemRO: Studying the Romanian expression of antisemitism. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 291–298, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online. Association for Computational Linguistics.

Stefan Daniel Dumitrescu and Andrei-Marius Avram. 2020. Introducing RONEC - the Romanian named entity corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4436–4443, Marseille, France. European Language Resources Association.

Stefan Daniel Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, Luciana Morogan, George Dima, Gabriel Marchidan, Traian Rebedea, Madalina Chitez, Dani Yogatama, Sebastian Ruder, Radu Tudor Ionescu, Razvan Pascanu, and Viorica Patraucean. 2021. Liro: Benchmark and leaderboard for romanian language tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. Did they answer? subjective acts and intents in conversational discourse. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1644, Online. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378–382.

Oana Celia Gheorghiu and Alexandru Praisler. 2022. Hate speech revisited in romanian political discourse: from the legion of the archangel michael (1927–1941) to aur (2020–present day). *Humanities and social sciences communications*, 9(1):235.

Daniela Gîfu and Dan Cristea. 2013. Towards an automated semiotic analysis of the romanian political discourse. *Computer Science Journal of Moldova*, 61(1):36–64.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2019. Computational analysis of political texts: Bridging research efforts across communities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 18–23, Florence, Italy. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4479–4488, Online. Association for Computational Linguistics.

Akriti Jain and Aparna Garimella. 2026. Knowing what's missing: Assessing information sufficiency in question answering. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4163–4174, Rabat, Morocco. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

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. *Preprint*, arXiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

William C. Mann. 1984. Discourse structures for text generation. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 367–375, Stanford, California, USA. Association for Computational Linguistics.

Mihai Masala, Denis C. Ilie-Ablachim, Dragos Corlatescu, Miruna Zavelca, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. Openllm-ro – technical report on open-source romanian llms. *Preprint*, arXiv:2405.07703.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. RoBERT – a Romanian BERT model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637, Barcelona,

Spain (Online). International Committee on Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Brighton Msagalla and Marianna Visser. 2021. Agenda-setting through topic shift in tanzanian parliamentary debate: The derailment of strategic manoeuvring. *Southern African Linguistics and Applied Language Studies*, 39(4):327–337.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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.

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

Ines Reinig, Ines Rehbein, and Simone Paolo Ponzetto. 2024. How to do politics with words: Investigating speech acts in parliamentary debates. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8287–8300, Torino, Italia. ELRA and ICCL.

Ana-Cristina Rogoz, Radu Tudor Ionescu, Alexandra-Valentina Anghel, Ionut-Lucian Antone-Iordache, Simona Coniac, and Andreea Iuliana Ionescu. 2026. A large-scale benchmark for evaluating large language models on medical question answering in romanian. *Preprint*, arXiv:2508.16390.

Tanya Romaniuk. 2013. Pursuing answers to questions in broadcast journalism. *Research on Language & Social Interaction*, 46(2):144–164.

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2019. Target based speech act classification in political campaign text. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 273–282, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Andra Vasilescu, Mihaela-Viorica Constantinescu, Ariadna Ștefănescu, and Șerban Hartular. 2024. *Insights into Romanian political discourse*. Cambridge Scholars Publishing.

Bhuvanesh Verma, Mounika Marreddy, and Alexander Mehler. 2026. Predicting convincingness in political speech: How emotional tone shapes persuasive strength. In *The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2026)*, pages 37–51, Rabat, Morocco. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Michał Zawalski, Meriem Boubdir, Klaudia Bałazy, Besmira Nushi, and Pablo Ribalta. 2026. Detecting data contamination in llms via in-context learning. In *ICLR 2026*.

A Taxonomy and Annotation Guidelines

A.1 Romanian Annotation Guidelines

Label names. We retained the nine original English label names (*Explicit*, *Implicit*, *General*, *Partial/half-answer*, *Dodging*, *Deflection*, *Declining to answer*, *Claims ignorance*, *Clarification*) to ensure our dataset remains directly comparable with the English corpus.

Descriptions and examples. We translated all category descriptions into Romanian and replaced the English examples of Thomas et al. (2024) with Romanian examples drawn from political discourse data in our dataset. Each example pairs an answer (R) with a question (Î) and a short rationale identifying the features that trigger the label.

A Romanian linguistic expert assisted with example selection to ensure that each category is illustrated by a clear and representative instance. Table 3 reproduces the guidelines, with English descriptions and bilingual examples (Romanian as supplied to annotators; English translation in italics for readability).

A.2 Summaries and Quality Control

Annotators were required to classify the evasion strategy based on the original text. In addition, the annotators received small GPT-5.4-generated summaries and were asked to verify the validity of the summary. To ensure that the annotators were reading the source text, we inserted 60 intentionally incorrect summaries. The annotators correctly flagged 58 of these (96.67%), producing only two false negatives. None of the other summaries was flagged as inconsistent.

A.3 Worked Example

Î: *"Domnule Președinte, pădurea este o chestiune de siguranță națională?"*
R: *"Eu cred că pădurile sunt o chestiune de suflet național și de aceea sunt foarte încrezător că vom reuși să schimbăm lucrurile în bine."*

English:
Q: *"Mr. President, are forests a matter of national security?"*
A: *"I believe forests are a matter of national spirit, and for that reason I am very confident that we will succeed in changing things for the better."*

The response receives clarity label *Ambivalent* and evasion label *Deflection*. The answer engages with the topic introduced in the question (forests as a national concern), but does not address whether forests constitute a national security issue. It reframes the discussion and shifts the attention away from the original premise without explicitly rejecting it.

B Annotation Protocol

B.1 Labeling Platform

We used Labelbox[5] to manage the annotation workflow. Each item displayed the original interview question and answer verbatim, with the LLM-generated question and answer summaries attached as side panels for reference (Figure 4). The annotator completed two tasks per item: (i) assign one of the nine evasion labels (*Clasificare Evaziune*) from the Level-2 taxonomy, and (ii) verify whether the answer summary correctly reflects the original.

The platform provided the complete annotation guidelines, which included the taxonomy table (Table 3) and an extended set of examples for borderline cases covering categories with closely related semantics (e.g. *Dodging* vs. *Deflection*, *General* vs. *Partial/half-answer*). Annotators could consult the guidelines at any point during a labeling session.

Annotators could additionally tag any instance with one of five issue categories:

- *Parsing Error* – the QA pair was malformed during extraction;
- *Not a Question* – the journalist's statement does not solicit any information;
- *Multiple Questions* – a single journalist's question contains several sub-questions;
- *Insufficient Context* – the response cannot be accurately classified without additional background information;
- *Borderline / Hard to Decide* – the response is difficult to classify and may fall between two evasion categories;

Annotators could also create custom issue categories to report any additional problems not covered by the predefined taxonomy.
Items tagged with the first three issues were filtered from the released corpus. Items tagged *Insufficient Context* or *Borderline / Hard to Decide* were escalated to the expert, who either adjudicated a final label or removed the instance.

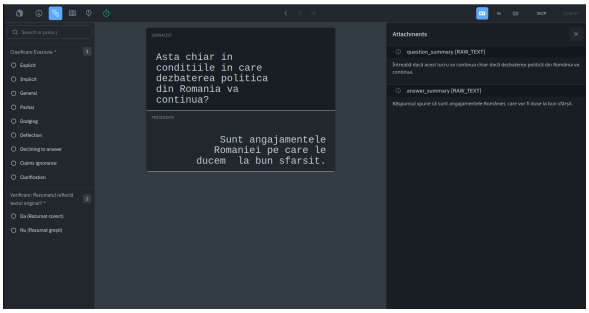

Figure 4: Labelbox annotation interface. Left panel: evasion classification and summary validity check. Centre: original question and answer. Right panel: LLM-generated question and answer summaries as attachments.

[5] https://labelbox.com/

B.2 Inter-Annotator Agreement

Raw annotator agreement on the test set is high, especially for clarity. All three annotators agree on 87.2% of clarity items, while the remaining 12.8% receive a two-thirds majority, leaving no items without consensus. Agreement on evasion is slightly lower but still substantial: 69.6% of items receive full three-way agreement, 28.0% receive a two-thirds majority, and only 2.4% result in complete disagreement between all three annotators.

Figure 5 shows pairwise Fleiss' κ across the three clarity labels. All pairwise values are at least 0.85, with the lowest agreement at the *Clear Reply–Ambivalent* boundary ($\kappa = 0.85$). Disagreement is mainly found between *Clear Reply* and *Ambivalent* answers rather than between the two non-ambiguous labels (*Clear Reply* vs. *Clear Non-Reply*: $\kappa = 0.98$).

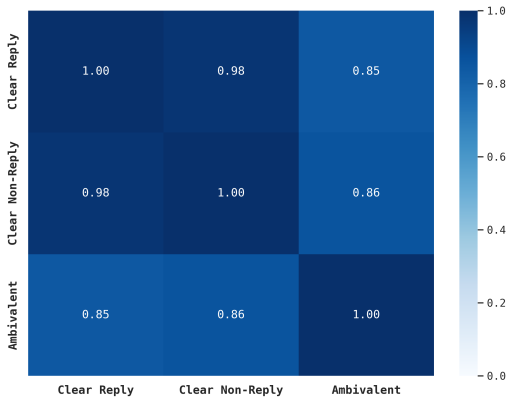

Figure 5: Pairwise Fleiss' κ between clarity categories.

Figure 6 shows the same matrix for the nine evasion categories. The lowest pairwise κ values occur for *Deflection*, in particular *Deflection–Dodging* ($\kappa = 0.26$) and *Deflection–General* ($\kappa = 0.38$). These confusions reflect how close the categories are in the taxonomy. *Dodging* and *Deflection* both involve leaving the question unanswered. They differ in whether the answer acknowledges the question before shifting away. *General* and *Deflection* differ in whether the answer remains on topic at all, which depends on how broadly the topic is interpreted.

Figure 7 shows the per-category one-vs-rest Fleiss' κ. Four categories cross the substantial-agreement threshold of $\kappa = 0.6$ (*Clarification*, *Claims ignorance*, *Explicit*, *Declining to answer*); three fall between fair and substantial (*Partial/half-answer*, *Dodging*, *Implicit*); two fall in the fair range (*General*, *Deflection*).

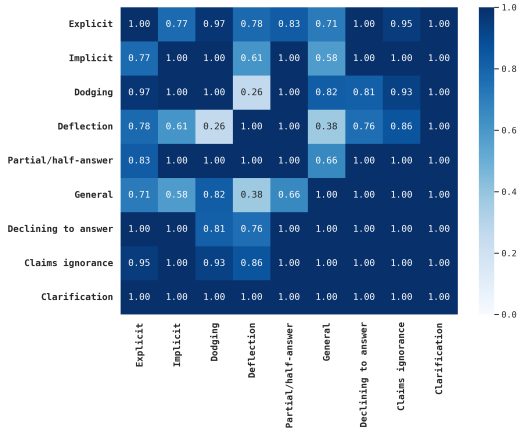

Figure 6: Pairwise Fleiss' κ between evasion categories.

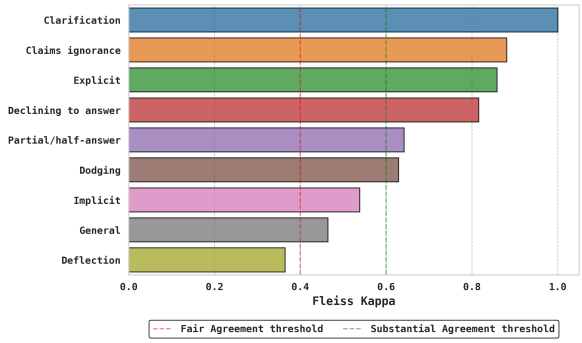

Figure 7: One-vs-rest Fleiss' κ per evasion category.

B.3 Expert Benchmark Evaluation

The linguistic expert independently labeled a benchmark of 85 instances drawn from the test split. Figure 8 reports per-annotator accuracy against the expert: 95.3%, 95.3%, 94.1% on clarity and 89.4%, 88.2%, 84.7% on evasion. Clarity accuracy is 5–10 percentage points higher than evasion accuracy for each annotator, which is consistent with the greater difficulty of the nine-class task compared to the three-class task.

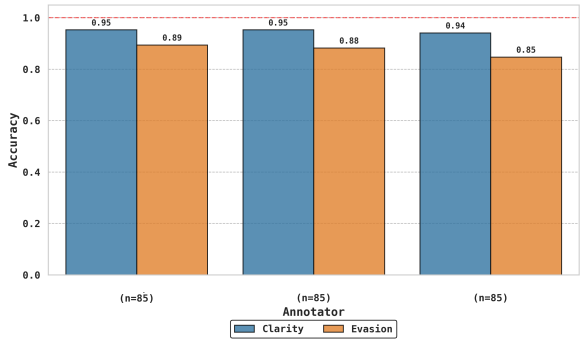

Figure 8: Per-annotator clarity and evasion accuracy against expert benchmark.

C Exploratory Data Analysis

This appendix reports descriptive statistics over the 3,574 question–answer pairs in the dataset (3,278 train; 296 test). The statistics are computed over the full dataset and use the majority-vote labels on the test set for both clarity and evasion categories.

C.1 Label Distribution

Figures 9 and 10 show the label distributions for clarity and evasion. Explicit answers make up the majority of the dataset (55.09%). The other eight evasion strategies are much less common, with Clarification occurring in only 1.09% of the question-answer pairs.

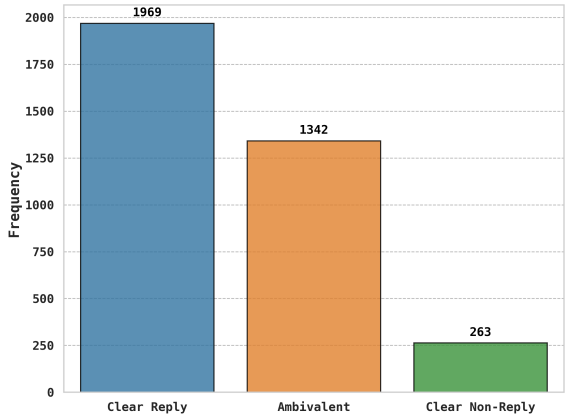

Figure 9: Overall label distribution for Level 1 (Clarity).

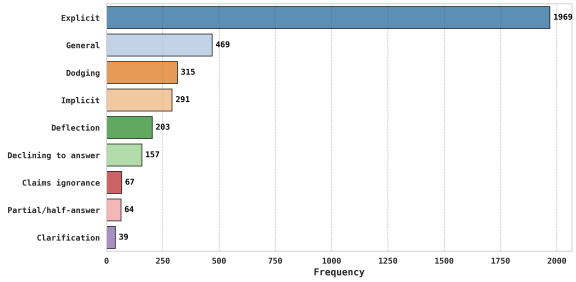

Figure 10: Overall label distribution for Level 2 (Evasion).

C.2 Distribution by Speaker

The corpus covers five Romanian presidents (Figure 11). Coverage reflects data availability: Băsescu (1,214) and Iohannis (1,100) account for 64.8% of the corpus; Bolojan (56) is much less frequent, as he has been an interim president for 100 days.

Figure 12 shows the clarity rates for each speaker. Iliescu and Băsescu give the most direct answers, with Clear Reply rates of 62.06% and 62.60%. In contrast, Bolojan gives the most Ambivalent answers (53.57%), and Iohannis has the highest rate of Clear Non-Replies (11.55%).

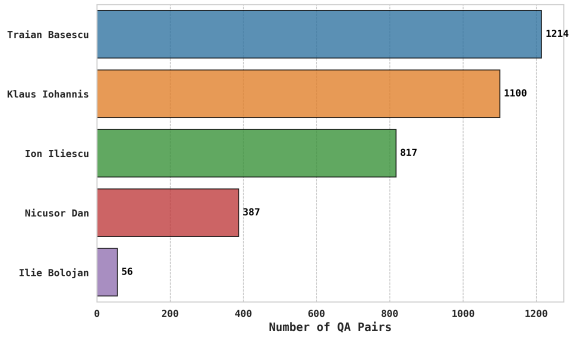

Figure 11: Total number of question-answer pairs per speaker.

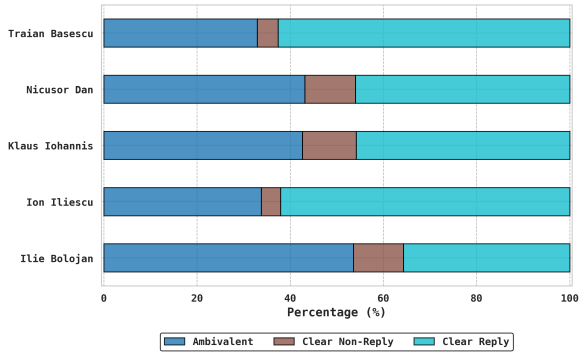

Figure 12: Clarity label distribution per speaker.

Fine-grained evasion distributions per speaker are shown in Figure 13. Dodging is most frequent for Iliescu (11.87%) and Băsescu (9.06%). Deflection is most common for Bolojan (12.50%) and Dan (10.08%). Declining to answer is highest for Iohannis (9.09%) and Bolojan (8.93%). The post-2014 speakers (Iohannis, Bolojan, and Dan) show a more balanced distribution of evasion types, with higher proportions of General, Deflection, and Declining responses. In contrast, Iliescu and Băsescu rely more heavily on Dodging, while still remaining largely dominated by Explicit answers.

C.3 Distribution by Communicative Context

We assigned a context label to each source transcript from event titles using keywords associated with each category. Interviews make up the majority of the dataset (1,904 pairs), followed by declarations (1,213 pairs) and press conferences (457 pairs).

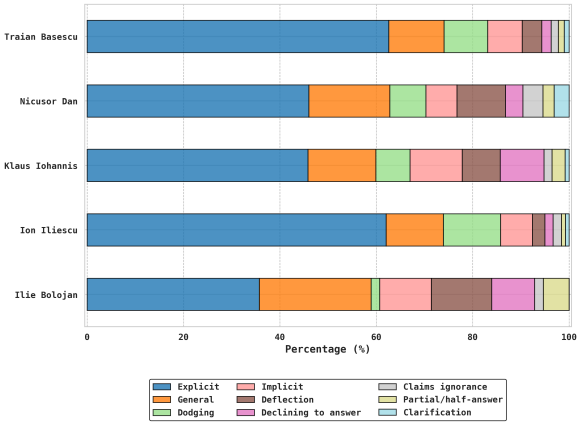

Figure 13: Evasion label distribution per speaker.

Figure 14 reports clarity rates across the three resulting contexts. Interviews have the highest rate of Clear Replies (63.03%) and the lowest rate of Clear Non-Replies (3.62%). Declarations and press conferences show very similar clarity distributions.

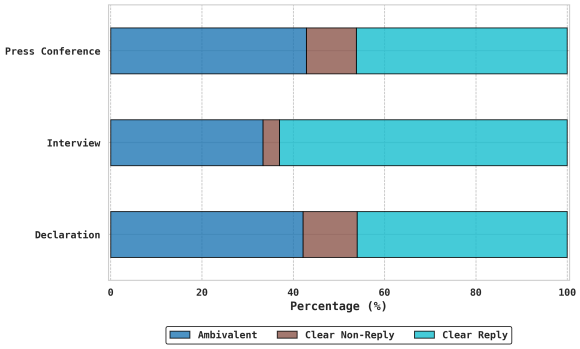

Figure 14: Clarity distribution per communicative context.

Part of the higher clarity rate in interviews is linked to differences in topic distribution. Personal & Others makes up 38.39% of interview questions, compared to only 12.94% in declarations and 8.75% in press conferences (§C.4). This category also has the highest Clear Reply rate of any topic (65.62%). Press conferences contain more Justice & Anti-corruption and Domestic Policy questions, which have the lowest Clear Reply rates.

Evasion distributions per context are shown in Figure 15. Interviews show a higher percentage of Dodging (10.45%), but low levels of the other evasion types. Declarations and press conferences have roughly twice the rate of Declining to answer (8.41% and 6.13%) and higher Deflection (7.50% and 10.72%) compared to interviews (1.42% and 3.31%). Press conferences have the highest Deflection rate overall. If we exclude interviews and consider only the more formal settings of declarations and press conferences (n=1,670), the overall proportion of direct answers drops significantly: the *Explicit* rate falls to 46.05%, while evasive categories such as *Deflection* (8.38%) and *Declining to answer* (7.79%) account for a larger share of responses.

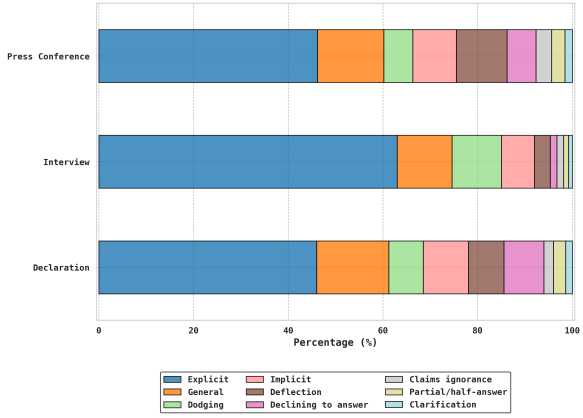

Figure 15: Evasion label distribution per communicative context.

Figure 16 crosses speaker with context. Iliescu and Băsescu interviews (n=754, n=1,150) explain the high interview Explicit rate (62.47%, 63.39%) and also account for most Dodging (12.47%, 9.13%) in interviews. Iohannis declarations (n=852) account for the bulk of Declining to answer in the declaration row (9.74%). Dan declarations show the highest General share (17.20%) and Bolojan press conferences (n=36) the highest combined non-Explicit share (72.22%).

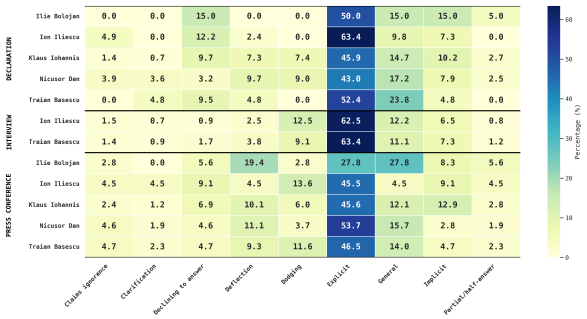

Figure 16: Evasion distribution per speaker × context cell.

C.4 Distribution by Topic Category

Categories were defined by inspecting question content to identify nine main topics, then a single label was assigned to each question by GPT-5.4. Figure 17 shows the overall distribution of these

topics.

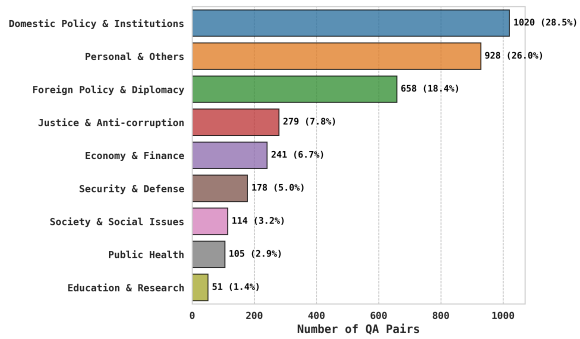

Figure 17: Total number of question-answer pairs per topic category.

Figure 18 shows the clarity rates for each topic. Even though Justice & Anti-corruption questions make up only 7.81% of the data, they trigger the highest rate of Clear Non-Replies (15.05%).

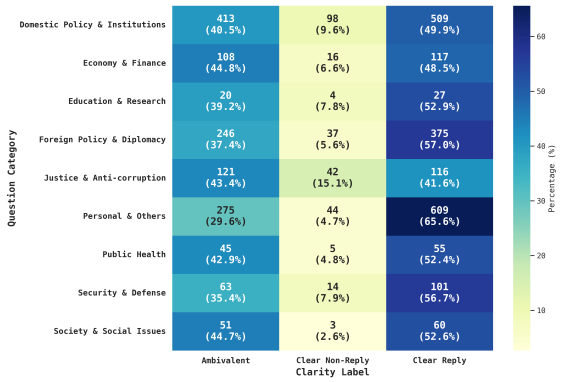

Figure 18: Clarity label distribution per topic category.

Per-category evasion distributions are shown in Figure 19. Personal & Others has the highest Dodging share (11.21%), although most responses in this category are still Explicit. Justice & Anti-corruption has the highest rates of Declining to answer (9.32%) and Claims ignorance (5.02%). Society & Social Issues shows the highest share of Implicit (11.40%), followed by Domestic Policy & Institutions (10.29%). General responses are most frequent for Society & Social Issues (18.42%) and Public Health (18.10%).

C.5 Temporal Coverage

QA pairs span 2001–2026 in two main periods: 2001–2007 (Iliescu and Băsescu, 1,935 pairs) and 2015–2026 (Iohannis, Bolojan, and Dan, 1,543 pairs). The intervening 2008–2014 window contains 96 pairs, due to limited archival availability.

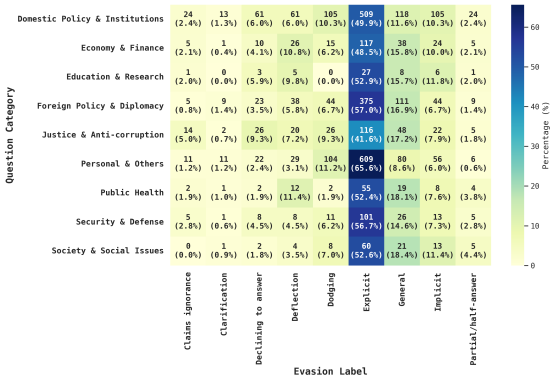

Figure 19: Evasion label distribution per topic category.

Per-year volumes and per-speaker timelines are shown in Figures 20 and 21.

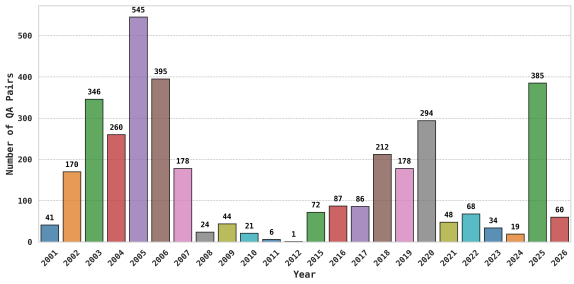

Figure 20: Yearly volume of QA pairs, 2001–2026.

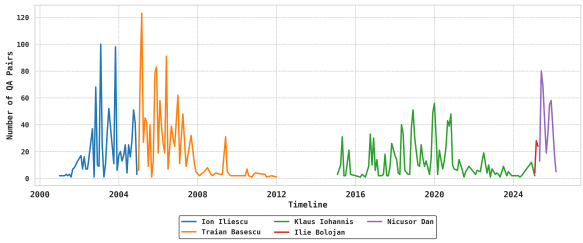

Figure 21: Yearly QA volume per speaker.

Year-level clarity composition is shown in Figure 22. Clear Non-Reply rates remain below 6% in 2001–2007, and increase to around 10–20% in 2016–2026, with Ambivalent responses increasing in parallel.

C.6 Length and Verbosity

Mean question length is 30.7 words, while the median is 21 words; mean answer length is 76.9 words, while the median is 45 words. Overall length distributions are shown in Figure 23.

Answer length differs substantially by label (Figures 24 and 25). Clear Replies average 88.4 words, Ambivalent answers 67.4, and Clear Non-Replies 39.7.

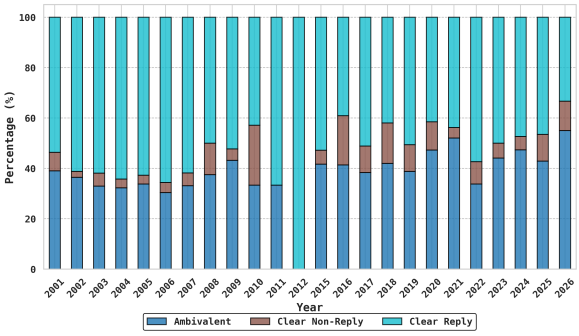

Figure 22: Clarity label composition per year, 2001–2026.

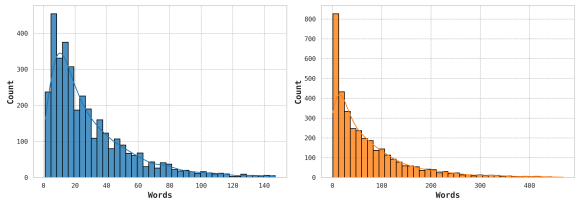

Figure 23: Question and answer word-count distributions.

Looking at the specific evasion strategies, Clarification answers are the shortest (mean 6.4, median 4) followed by Dodging (mean 27.6, median 13). In contrast, General, Deflection, and Partial/half-answer responses are the longest non-Explicit categories (averaging 89.7, 88.6, and 87.3 words). This aligns with verbose evasion tactics, where politicians surround the question with extra, related material instead of answering it directly.

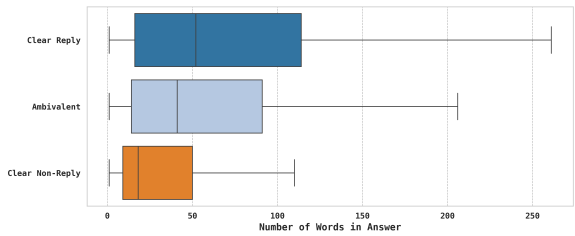

Figure 24: Answer word-count distribution per clarity label.

Per-speaker mean answer length ranges from 52.9 words (Dan) to 110.6 (Bolojan, n=56), with Iohannis at 73.1, Băsescu at 69.3 and Iliescu at 102.6. The full per-speaker answer-length distribution is shown in Figure 26. Speaker-level verbosity does not align with evasion: Iliescu is among the most verbose speakers and also one of the most *Explicit*, while Dan is the most concise and is also predominantly *Explicit*.

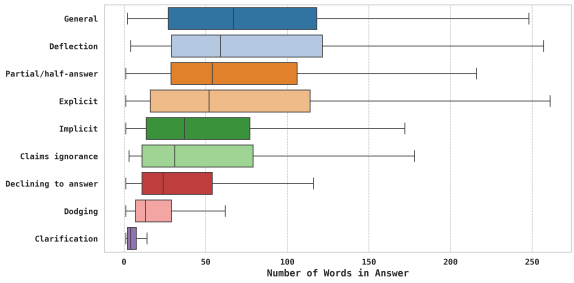

Figure 25: Answer word-count distribution per evasion label.

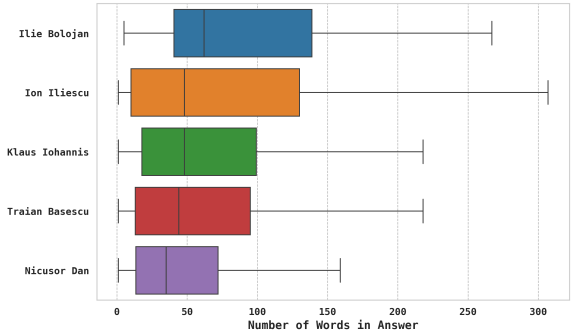

Figure 26: Answer word-count distribution per speaker.

D Experiment Details

This appendix expands the experimental setup and analysis summarised in Sections 3 and 4.

D.1 Evaluation

Across all experiments, we evaluate model performance using macro-averaged F1 as the primary metric. The macro-F1 is calculated as the unweighted mean of the class-specific F1 scores:

$$F1_{\text{macro}} = \frac{1}{|C|} \sum_{i=1}^{|C|} F1_i \tag{1}$$

where $|C|$ is the total number of classes (3 for clarity, 9 for evasion), and $F1_i$ is the harmonic mean of precision (P_i) and recall (R_i) for class i:

$$F1_i = 2 \cdot \frac{P_i \cdot R_i}{P_i + R_i} \tag{2}$$

To compute precision and recall, model predictions are compared against human annotations. To ensure our results are directly comparable with Thomas et al. (2024), we adopt their exact evaluation methodology.

For the 3-class clarity task, predictions are evaluated strictly against a single ground-truth label

obtained via majority vote. A prediction is considered correct only if it exactly matches this label.

For the 9-class evasion task, we use an "any-of" rule on the test set to reflect variation in annotator labels. A prediction is counted as correct if it matches the label of any of the three annotators. If it does not match any annotator label, it is treated as incorrect and compared against the majority-vote label.

D.2 Methodology

TF-IDF baseline. We extract word unigram TF-IDF features and tune $\min_df \in \{1, 2, 3\}$ and $\max_df \in \{0.9, 0.95, 1.0\}$. We use an L2-regularised logistic regression classifier with class-balanced weights. We perform a grid search over $C \in \{0.25, 0.5, 1, 2, 3, 5, 10\}$ and select the best model based on development set macro-F1.

For CLARITY, the best configuration is $\min_df$=2, $\max_df$=0.95, C=0.5 for clarity, and $\min_df$=2, $\max_df$=0.9, C=2 for evasion. For PolERo, the best parameters are $\min_df$=3, $\max_df$=0.9, C=5 for clarity and $\min_df$=1, $\max_df$=0.9, C=2 for evasion.

Encoder Fine-Tuning

Input formatting. Encoder models use the input format "`Question: {q}\nAnswer: {a}`". On CLARITY, q is the GPT-decomposed atomic `question` field paired with the `interview_answer`. On PolERo, q is the `interview_question` field paired with the same `interview_answer`, since the Romanian dataset already contains single-question inputs.

Single-head encoders. Each of the nine single-head encoder configurations is fine-tuned separately for each task for up to 20 epochs using AdamW (Loshchilov and Hutter, 2019), with a batch size of 64. The maximum sequence length is set to 512 for all models except ModernBERT-large, which utilizes a maximum sequence length of 8192. The best checkpoint is selected on an 85/15 train/validation split of the official training set and evaluated once on the test set. Learning rates vary by architecture: ModernBERT-large, and ELECTRA-large; 1×10^{-5} for RoBERTa-large and XLM-RoBERTa-large; and 2×10^{-5} for BERT-large-cased, mBERT-base, **Ro**BERT-large, and Romanian BERT.

Multi-head architecture. We apply this architecture to three encoder configurations: RoBERTa-large, XLM-RoBERTa-large, and **Ro**BERT-large. All multi-head models use 7-fold cross-validation stratified by the clarity label. Models are trained for 20 epochs using AdamW with a learning rate of 1×10^{-5} and a batch size of 8. We use early stopping based on the combined validation macro-F1 of both heads, with a patience of 5. At inference time, we ensemble all 7 folds by averaging their predicted class probabilities and selecting the most probable class for each classification head.

Large Language Models.

Open-source LLMs were run locally, while proprietary models were accessed via the OpenAI API platform[6]. All models are queried with default settings and reasoning configurations, using a restricted output schema. Zero-shot prompts include the task description, the full set of category names, and one-line definitions. Few-shot prompts add one labeled example per category. Prompts are run separately for the clarity and evasion tasks and use the dataset's language.

E Multi-Head Architecture Details.

This appendix expands the multi-head architecture introduced in Section 3 and isolates the contribution of each component on the English dataset. All ablation numbers are 7-fold cross-validation means and standard deviations of validation macro-F1, computed on the training split with folds stratified by the clarity label. Test-set scores for the full configuration are reported in Table 8.

E.1 Hierarchical Input Processing

Responses in both datasets are often longer than the 512-token limit of standard pretrained encoders. Because evasion cues can appear anywhere in a response, simple truncation may remove the relevant span. To avoid this, we segment each tokenized question–answer sequence T into overlapping windows of length L=512 with stride S=256:

$$C_k = T[kS : kS + L] \quad \text{for } k = 0, \ldots, M - 1, \tag{3}$$

where $M = \lceil \max(|T| - L, 0)/S \rceil + 1$. The final chunk extends to $|T|$ and is zero-padded if shorter. The window size L is fixed by the encoder's positional embedding capacity. The 50% overlap ensures that every interior token is included in at least two consecutive windows, reducing the chance that important evasion cues are separated across chunks.

[6] https://platform.openai.com

Each chunk is encoded independently by the shared encoder. We extract the hidden state at position 0 of each chunk, $h_k = H_k[0, :] \in \mathbb{R}^d$, and aggregate the M chunk vectors via element-wise max-pooling:

$$v_j = \max_{k=0}^{M-1} h_{k,j} \quad \text{for } j = 1, \ldots, d. \tag{4}$$

The position-0 token corresponds to the encoder's start-of-sequence token only for the first chunk; for subsequent chunks it is the first content token of the window. We use this position uniformly because the encoder distributes contextual information across all tokens.

E.2 Multi-Task Heads and Training

The pooled vector v is passed through dropout (p=0.1) and fed into two linear classification heads:

$$\hat{y}_c = \text{softmax}(W_c \cdot \text{Dropout}(v) + b_c), \tag{5}$$
$$\hat{y}_e = \text{softmax}(W_e \cdot \text{Dropout}(v) + b_e), \tag{6}$$

with $W_c \in \mathbb{R}^{3\times d}$ and $W_e \in \mathbb{R}^{9\times d}$. The training objective is the unweighted sum of two cross-entropy losses, $\mathcal{L} = \mathcal{L}_c + \mathcal{L}_e$. At inference time, the 7 fold models are ensembled by averaging predicted class probabilities and taking the argmax per head.

E.3 Component Ablations

The ablations below isolate the impact of each architectural choice over single-head encoder fine-tuning.

Pooling strategy. Table 4 compares element-wise max-pooling with mean-pooling and a baseline that keeps only the first chunk representation. Max-pooling performs best. Mean-pooling performs worse than max-pooling, indicating that simple averaging is less effective for aggregating chunk representations. The first-chunk baseline misses any signal beyond the first 512 tokens, which affects 28.8% of English responses.

Pooling	Clarity	Evasion
First chunk only	0.67 ± 0.01	0.42 ± 0.01
Mean-pooling	0.68 ± 0.02	0.43 ± 0.02
Max-pooling	$\mathbf{0.70 \pm 0.02}$	$\mathbf{0.45 \pm 0.02}$

Table 4: Chunk aggregation strategy. 7-fold CV validation macro-F1 (mean $\pm$ std).

Multi-task versus single-task. Table 5 shows that joint training matches single-task performance on clarity and improves single-task evasion by $+0.03$ macro-F1. The coarser clarity supervision acts as a regularizer for rare evasion categories: each evasion label maps deterministically to a clarity label, so gradients from $\mathcal{L}_c$ help stabilize representations of the parent class, even when the evasion head provides limited signal.

Objective	Clarity	Evasion
Single-task (clarity)	$\mathbf{0.70 \pm 0.02}$	–
Single-task (evasion)	–	0.42 ± 0.01
Multi-task	0.70 ± 0.02	$\mathbf{0.45 \pm 0.02}$

Table 5: Effect of joint training. 7-fold CV validation macro-F1 (mean $\pm$ std).

Ensemble size. Table 6 reports performance for $k \in \{3, 5, 7\}$ folds. Increasing k reduces variance and increases the amount of training data per fold model. Improvements are consistent across both subtasks and larger for the more difficult evasion task.

Folds	Clarity	Evasion
3	0.66 ± 0.01	0.42 ± 0.02
5	0.68 ± 0.02	0.43 ± 0.03
7	$\mathbf{0.70 \pm 0.02}$	$\mathbf{0.45 \pm 0.02}$

Table 6: Effect of ensemble size. k-fold CV validation macro-F1 (mean $\pm$ std).

Loss function. Table 7 compares unweighted cross-entropy with inverse-frequency class-weighted cross-entropy and focal loss (γ=2) (Lin et al., 2020). Neither alternative improves macro-F1. Class weighting increases minority-class recall (*Clear Non-Reply* from 0.62 to 0.67, *Partial/half-answer* from 0.00 to 0.10), but this is offset by lower precision on majority classes. Focal loss yields intermediate performance between the two. Most evasion errors occur between pragmatically similar classes (*Implicit*, *Deflection*, *General*, *Dodging*) that share surface features. Reweighting increases gradient emphasis on rare classes but does not resolve this overlap.

Loss	Clarity	Evasion
Cross-entropy (unweighted)	$\mathbf{0.70 \pm 0.02}$	$\mathbf{0.45 \pm 0.02}$
Class-weighted CE	0.66 ± 0.02	0.41 ± 0.02
Focal (γ=2)	0.69 ± 0.02	0.44 ± 0.02

Table 7: Effect of loss function. 7-fold CV validation macro-F1 (mean $\pm$ std).

F Extended Model Comparison and Error Analysis

Table 8 presents the complete test-set macro-F1 results for all baselines, architectural variants, cross-lingual transfer setups, and LLM prompting strategies discussed in this work.

F.1 Traditional classifier

The TF-IDF + logistic regression baseline achieves similar macro-F1 on both languages (0.42 on English, 0.45 on Romanian), but its error patterns show which parts of the task rely on lexical cues. The model correctly identifies most *Clear Reply* instances in Romanian (64% recall), but performs much worse in English (33% recall). It performs well on explicit replies when they contain strong unigram cues such as commitment verbs, numerals, dates, or named entities that match the question.

In contrast, the *Ambivalent* class is often confused with the other two labels. This confusion is especially symmetric in Romanian: 48% of true *Ambivalent* instances are predicted as *Clear Reply*, while 33% of true *Clear Reply* instances are predicted as *Ambivalent*. This suggests that any answer that reuses question vocabulary is often treated as a reply by the baseline, even when it does not clearly commit to one.

Clear Non-Reply performs worst, with an F1 of 0.25 on English and 0.26 on Romanian (and recall as low as 24%). These are typically short, lexically varied refusals (e.g., *Declining to answer*, *Claims ignorance*) that depend on discourse-level cues than lexical overlap, which a bag-of-words model cannot capture.

The slightly higher performance on Romanian is consistent with the higher share of *Explicit* replies in its training distribution (Section C.1): more cases can be solved through surface matching, leaving less room for improvement over a lexical baseline.

F.2 Encoder Models

Best single-head models. Among single-head encoders, RoBERTa-large is the strongest English model on evasion, benefiting from a tokenizer with a language-specific vocabulary and from pretraining objectives that produce stable sentence-pair representations.

RoBERT-large is the strongest model on PolERo. Its parameters are trained specifically for Romanian syntax and semantics, and its tokenizer handles Romanian morphology without the over-segmentation caused by XLM-RoBERTa's shared cross-lingual vocabulary. Although XLM-RoBERTa-large is the only multilingual model competitive with monolingual encoders on Romanian clarity, its larger scale does not fully compensate for these tokenisation issues, leading to weaker performance on the fine-grained evasion task.

Failure modes. ELECTRA-large and ModernBERT-large are the two main failure cases. ELECTRA's replaced-token-detection (RTD) objective encourages strong local, token-level predictions. While this works well in general, on small and highly imbalanced datasets it causes the model to rely on frequent lexical cues from the majority class instead of learning the broader patterns needed for evasion detection. As a result, the model collapses toward the *Explicit* class and, in the Romanian case, predicts a single class for all inputs.

ModernBERT-large shows the same pattern, but less severely. Although its local–global attention design is intended for longer documents, its pretraining data is closer to code and long-form web text than to short-form question-and-answer exchanges. As a result, three rare categories (*Claims ignorance*, *Partial/half-answer*, *Deflection*) receive zero recall on Romanian, and the mapped clarity projection inherits the same majority-class bias.

Multi-head architecture. Multi-head models consistently outperform single-head models across all reported metrics. This improvement comes from three main factors. First, joint multi-task supervision exploits the relationship between clarity and evasion (Section 2.1). The auxiliary clarity signal helps stabilize predictions for rare evasion categories, reducing false positives for *Explicit* by 25% on English RoBERTa-large (from 39 to 29 instances). At the same time, the evasion signal improves separation between *Clear Reply* and *Ambivalent*, lowering cross-boundary errors from 78 to 62. Second, chunked max-pooling prevents truncation of long inputs. Clarity or evasion signals can be found anywhere in a response. While most standard models truncate at 512 tokens, the chun-

ked architecture preserves the full context. Third, 7-fold ensembling reduces variance compared to single-checkpoint evaluation.

The strongest absolute performance is observed for RoBERTa-large (English) and **Ro**BERT-large (Romanian), where the backbone has sufficient capacity and in-language pretraining to benefit from the joint signal. XLM-RoBERTa-large shows larger relative gains, but its multilingual representations trade some within-language precision for cross-lingual coverage, which limits its final performance compared to monolingual encoders.

Error Analysis. While the boundary between *Clear Reply* and *Ambivalent* remains the largest source of errors, the out-of-fold confusion matrices for our strongest models (Figure 27, Figure 28) reveal that the fine-grained residual errors between evasion strategies concentrate along two main semantic axes.

The first axis is *Dodging*↔*Deflection*. Both involve topic shifts. The distinction depends on whether the model detects an initial attempt to address the prompt before shifting to a different topic, or instead an immediate shift away from the question. On English, this axis accounts for 157 errors, with similar numbers in Romanian.

The second main axis is *Implicit*↔*General*, which accounts for 142 cross-misclassifications on English. Both correspond to partial answers, and the distinction depends on whether a broad response can still be linked to the specific question.

Partial/half-answer is the most difficult category for all encoders, with an F1 of 0.066 on English and 0.024 on Romanian. Correctly identifying it requires determining whether only part of a multi-part question is answered while another part is ignored, a relational judgment that standard encoder representations do not capture well.

Error structure. We inspect where the errors of the strongest fine-tuned encoders land relative to annotator agreement on the test split. For multi-head RoBERTa-large on the English evasion task, 43.7% of errors occur on items with full 3/3 annotator agreement and 50.4% on items with 2/3 agreement; only 5.9% involve perfect disagreement. The pattern is more pronounced on multi-head **Ro**BERT-large for Romanian, where 54.7% of evasion errors and 56.6% of clarity errors fall on unanimously-labelled items. Errors concentrate on items with the cleanest annotation signal rather than on the genuinely ambiguous ones. The higher ap-

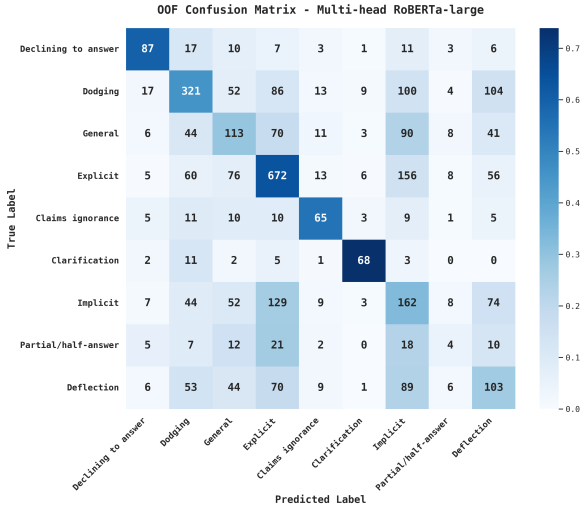

Figure 27: Out-of-fold (OOF) confusion matrix for the 9-class evasion task on CLARITY (EN) using multi-head RoBERTa-large.

parent accuracy on the perfect-disagreement bucket (.758 on English evasion vs. .528 on 3/3 items) is largely an artefact of the any-of evaluation rule (§D.1), which credits a prediction matching any of the three distinct annotator labels.

Calibration. The gap between top-1 probability on correct and incorrect predictions is narrow across all heads. The English RoBERTa-large multi-head averages .605 on correct evasion predictions and .483 on incorrect ones; the clarity head averages .835 vs. .726. The Romanian **Ro**BERT-large multi-head is more confident in both regimes (.923 vs. .822 for clarity; .859 vs. .637 for evasion), consistent with the more skewed Romanian label distribution. Confident errors recur in both languages: the Romanian clarity head assigns probability 1.00 to *Clear Reply* on terse, single-word responses unanimously labelled *Ambivalent*; the English evasion head assigns probability above .85 to *Explicit* on 3/3-agreement *Implicit* and *Deflection* responses. Top-1 probability is therefore not a reliable rejection signal at deployment.

F.3 LLM Prompting Analysis

LLM Performance Ranking. GPT-5.4 achieves the best mapped scores on both languages and the strongest English evasion performance. Qwen3.6-35B-A3B performs best on Romanian evasion under few-shot prompting, while Gemma-4-31B-it is best on English direct clarity.

DeepSeek-V4-Pro and gpt-oss-120B fall between these extremes. gpt-oss-120B performs par-

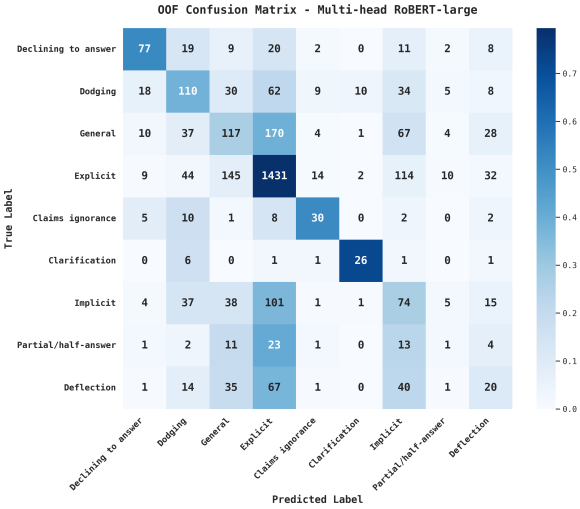

Figure 28: Out-of-fold (OOF) confusion matrix for the 9-class evasion task on PolERo (RO) using multi-head **Ro**BERT-large.

ticularly well on the mapped metric despite weaker direct clarity scores, showing that it often identifies the correct fine-grained evasion category but struggles more when directly predicting the broader three-class clarity labels.

Few-shot prompting. Few-shot prompting improves the mapped metric for most models because the examples help clarify the boundaries between similar categories, especially *Dodging*/*Deflection* and *Implicit*/*General*. Under zero-shot prompting, these are the category pairs that models commonly confuse, and even one example per category substantially improves alignment.

Llama-3.3-70B is the clearest exception. Its mapped clarity score on English drops from 0.624 to 0.564 under few-shot prompting. The confusion matrix shows that the model becomes heavily biased toward *Deflection*, predicting it for 79.5% of all English test instances regardless of the gold label. The same holds for Romanian, where *Deflection* accounts for 44% of all predictions.

LLMs vs. encoders. For both languages, the gap between the strongest LLMs and encoder ensembles is larger on the evasion task than on the clarity task. This reflects the difference between surface-level alignment and pragmatic intent. The clarity task mainly requires checking whether the answer matches the question, which strong encoders already handle well. In contrast, the nine-class evasion task requires distinguishing between closely related rhetorical strategies and benefits from broader discourse and pragmatic knowledge.

Multi-head fine-tuning reduces this gap, but does not eliminate it. Performance on the ambivalent categories remains the main challenge for both model families.

F.4 Machine Translation

Machine Translation Augmentation. Translate-Train recovers part of the mapped clarity gap but yields only small improvements on the evasion task. Machine translation preserves enough propositional content to recover the coarse three-class distinction, but distorts pragmatic cues that separate ambivalent categories.

Deflection and *Dodging* are most affected, as both rely more on discourse structure than lexical cues. As a result, both show substantial recall drops under translation compared to in-language single-head models. On the Romanian test set, recall for *Deflection* decreases from 41% to 10%, while *Dodging* drops from 50% to 30%.

F.5 Machine Translation Quality

We evaluate the quality of the NLLB-200-distilled-600M translations used in the cross-lingual experiments with XCOMET-XL, a model that estimates translation quality without reference translations. Results are reported for both translation directions on the training and test splits.

For RO→EN, translating the Romanian data resulted in mean QE scores of 0.650 (train, n=3,278) and 0.655 (test, n=296), with clear variation across categories. Formulaic categories translate well, including *Clarification* (0.781 train, 0.941 test) and *Declining to answer* (0.763 train, 0.763 test), which obtain the highest scores. Categories driven by pragmatics score lower. *Deflection* has the lowest scores in both splits (0.582 train, 0.473 test). This drop matches the recall degradation for *Deflection* and *Dodging* under MT augmentation.

For EN→RO, translating the English data resulted in mean QE scores of 0.218 (train, n=3,448) and 0.171 (test, n=308). Although these scores are notably low, a manual inspection of the translations shows that many spans flagged as errors are acceptable translations. Despite this noise, the relative ranking across categories is consistent with the RO→EN direction: *Deflection* and *General* score lowest, while *Clarification* and *Claims ignorance* score highest.

Clarity (L1)	Evasion (L2)	Description	Example (Romanian; English in italics)
Clear Reply	Explicit	The requested information is given directly, in the form requested, without ambiguity or required inference.	**Î:** Când va fi gata autostrada? **R:** Lucrările vor fi finalizate în luna martie 2025. *Q: When will the highway be ready? A: Works will be completed in March 2025.*
Ambivalent	Implicit	The requested information is conveyed but not stated in the expected form; the listener must infer the conclusion.	**Î:** Vă veți da demisia din funcția de ministru? **R:** Nu mai pot rămâne o secundă în acest guvern. *Q: Will you resign as minister? A: I cannot stay in this government a second longer.*
	General	The response stays on topic but remains too broad, abstract, or vague to deliver the requested specifics.	**Î:** În ce lună va fi gata autostrada? **R:** Dezvoltarea infrastructurii e o prioritate zero pentru noi și muncim zi de zi pe șantiere. *Q: In which month will the highway be ready? A: Infrastructure development is our top priority and we work every day on site.*
	Partial/half-answer	Covers concretely only one component of the requested information and omits the rest.	**Î:** Ce companii de stat vor fi privatizate anul acesta? **R:** Vă pot spune cu certitudine că Poșta Română se află pe această listă. *Q: Which state-owned companies will be privatised this year? A: I can tell you with certainty that the Romanian Postal Service is on the list.*
	Dodging	Does not engage with the substance of the question; the topic is neither acknowledged nor touched.	**Î:** De ce ați pierdut alegerile locale? **R:** Presa ar trebui să se ocupe de problemele reale ale cetățenilor. *Q: Why did you lose the local elections? A: The press should focus on citizens' real problems.*
	Deflection	Acknowledges the topic of the question but shifts the argument and focus to something other than what is asked.	**Î:** Câți bani ați alocat pentru acest program? **R:** Programul este un proiect la care ținem foarte mult, însă marea realizare de anul acesta este că am reușit să modernizăm peste 50 de școli. *Q: How much money did you allocate to this programme? A: The programme is a project we care deeply about, but this year's big achievement is that we modernised over 50 schools.*
Clear Non-Reply	Declining to answer	A verbal, assumed and explicit refusal to provide the requested information.	**Î:** Când veți publica raportul financiar? **R:** Nu voi face niciun comentariu pe acest subiect până la finalizarea auditului. *Q: When will you publish the financial report? A: I will not comment on this matter until the audit is complete.*
	Claims ignorance	The speaker explicitly admits not holding the requested information at the moment.	**Î:** La ce dată exactă a ordonat guvernul reamenajarea navei HMAS Kanimbla? **R:** Nu știu acea dată. Voi afla și voi informa Camera. *Q: On what exact date did the government order the refit of HMAS Kanimbla? A: I do not know that date. I will find out and inform the Chamber.*
	Clarification	Does not provide the requested information and instead asks for clarification of what the question refers to.	**Î:** A fost decizia dumneavoastră să eliberați fondul? **R:** Vă referiți la fondul public sau cel privat? *Q: Was it your decision to release the fund? A: Do you mean the public fund or the private one?*

Table 3: PolERo taxonomy. Descriptions and examples are taken from the Romanian annotation guidelines used by annotators; English translations are provided in italics for readability.

Model	CLARITY (EN, n=308)			PolERo (RO, n=296)		
	Clarity	Evasion	Mapped	Clarity	Evasion	Mapped
TF-IDF + Logistic Regression	.422	.301	.441	.453	.312	.472
Single-head encoder fine-tuning						
mBERT-base	.529	.347	.517	.601	.328	.502
BERT-large-cased	.587	.361	.581	.475	.282	.501
ELECTRA-large	.607	.231	.524	.500	.084	.241
ModernBERT-large	.575	.360	.524	.440	.246	.443
RoBERTa-large	.580	.481	.656	.646	.428	.613
XLM-RoBERTa-large	.579	.419	.564	.680	.365	.686
Romanian BERT (RO)	–	–	–	.684	.470	.634
RoBERT-large (RO)	–	–	–	.688	.504	.701
Multi-head encoder fine-tuning						
RoBERTa-large	<u>.713</u>	.568	.731	.696	.501	.652
XLM-RoBERTa-large	.665	.475	.644	.694	.508	.706
RoBERT-large (RO)	–	–	–	.729	.545	**.763**
Cross-lingual transfer (XLM-RoBERTa-large, multi-head)						
EN → RO	–	–	–	.635	.399	.655
EN2RO → RO	–	–	–	.665	.412	.706
RO → RO2EN	–	–	–	.681	.494	.702
RO + EN2RO → RO	–	–	–	.708	.492	.681
EN + RO → RO	–	–	–	.723	.570	.717
RO → EN	.551	.398	.527	–	–	–
RO2EN → EN	.477	.358	.439	–	–	–
EN → EN2RO	.616	.448	.612	–	–	–
EN + RO2EN → EN	.621	.460	.647	–	–	–
EN + RO → EN	.679	.483	.642	–	–	–
EN + RO → EN + RO †		Clarity .737	Evasion .529		Mapped .705	
Cross-lingual transfer (RoBERTa-large, multi-head)						
EN → RO	–	–	–	.624	.409	.639
EN2RO → RO	–	–	–	.609	.432	.625
RO → RO2EN	–	–	–	.625	.427	.635
RO + EN2RO → RO	–	–	–	.664	.512	.652
EN + RO → RO	–	–	–	.712	.463	.681
RO → EN	.572	.337	.499	–	–	–
RO2EN → EN	.526	.352	.465	–	–	–
EN → EN2RO	.479	.357	.502	–	–	–
EN + RO2EN → EN	.672	.513	.654	–	–	–
EN + RO → EN	.671	.503	.683	–	–	–
EN + RO → EN + RO †		Clarity .715	Evasion .528		Mapped .718	
LLM prompting (no fine-tuning)						
Llama-3.3-70B (zero-shot)	.474	.377	.624	.479	.354	.446
Llama-3.3-70B (few-shot)	.535	.382	.564	.555	.312	.411
Gemma-4-31B-it (zero-shot)	.655	.546	.619	.657	.560	.713
Gemma-4-31B-it (few-shot)	**.721**	.583	.662	.672	.619	.738
GPT-OSS-120B (zero-shot)	.548	.559	.683	.575	.544	.740
GPT-OSS-120B (few-shot)	.542	.554	.729	.553	.580	.729
DeepSeek-V4-Pro (zero-shot)	.599	.512	.693	.604	.556	.696
DeepSeek-V4-Pro (few-shot)	.601	.545	.670	.627	.594	.751
Qwen3.6-35B-A3B (zero-shot)	.673	.536	.719	.684	.631	.738
Qwen3.6-35B-A3B (few-shot)	.669	.544	<u>.747</u>	<u>.746</u>	**.682**	<u>.760</u>
GPT-5.4 (zero-shot)	.687	<u>.613</u>	.729	.687	.649	.744
GPT-5.4 (few-shot)	.703	**.626**	**.761**	**.763**	<u>.666</u>	**.763**

Table 8: Full test-set macro-F1 across all models. **Clarity**: direct 3-class prediction. **Evasion**: direct 9-class prediction. **Mapped**: 3-class clarity prediction obtained by mapping the predicted 9-class evasion label upward through the taxonomy. For cross-lingual setups, *EN2RO* denotes English data machine-translated into Romanian, and *RO2EN* denotes Romanian data machine-translated into English. Best score per column is **bolded**; second-best is <u>underlined</u>.

† Evaluated on the combined EN+RO test set (n=604).