ILab-NLP at SemEval-2026 Task 9: Comparing XLM-RoBERTa and LLaMA-2 for Multilingual Polarization Detection

Declan Booth
Heriot-Watt University
db2042@hw.ac.uk

Gavin Abercrombie
Heriot-Watt University
g.abercrombie@hw.ac.uk

Simona Frenda
Heriot-Watt University
s.frenda@hw.ac.uk

Abstract

Online polarization is a growing concern in digital discourse, and detecting it early can support safer, more inclusive online spaces. We describe our submission to SemEval-2026 Task 9 (POLAR), Subtask 1, which focuses on binary classification of polarized versus non-polarized posts. We compare two approaches on English and Spanish: fine-tuning a multilingual encoder model (XLM-RoBERTa) and prompting a generative large language model (LLaMA-2 7B) to output a binary label. On the official test split, XLM-R achieves higher average performance (0.753 macro-F1 and 0.760 accuracy across English and Spanish) than LLaMA-2 (0.6845 macro-F1 and 0.7018 accuracy). Our analysis shows that LLaMA-2 is strongly biased toward predicting the positive class in Spanish, which increases recall but produces many false positives and poor calibration. We also report efficiency measurements using CodeCarbon, highlighting a practical tradeoff between performance and computational and environmental cost.

1 Introduction

POLAR at SemEval-2026 introduces a multilingual, multicultural, and multi-event shared task focused on detecting online polarization, defined as sharp division and hostility between social, political, or identity groups (Naseem et al., 2026a,b). Polarized discourse often precedes hate speech and social fragmentation, making early detection valuable for research and for building safer online platforms (Naseem et al., 2026a). However, polarization is highly contextual and can vary across languages, cultures, events, and platforms (Naseem et al., 2026b).

In this work, we participate in Subtask 1 (Polarization Detection) and focus on English and Spanish. We compare two common NLP approaches: (1) fine-tuning a multilingual encoder classifier (XLM-RoBERTa) and (2) prompting a generative large language model (LLaMA-2 7B) in a constrained, instruction-style setup. Beyond headline performance, we log per-example confidence signals and use SHAP to analyze which words and phrases drive the encoder model's predictions (our best system).

Our main finding is that fine-tuning XLM-R provides stronger and more stable performance across both languages, while few-shot prompting with LLaMA-2 transfers less reliably to Spanish and shows a strong tendency to over-predict the polarized class. We also quantify practical tradeoffs using CodeCarbon, comparing training and inference time and estimated emissions. The code and scripts for reproducing our experiments are available at https://github.com/Boothu/semeval-2026-task9-polarization-detection.

2 Background and Related Work

SemEval-2026 Task 9 (POLAR) is designed to capture multilingual, multicultural, and multi-event online polarization across a wide range of contexts (Naseem et al., 2026a). The POLAR benchmark includes data from multiple online sources (e.g., X, Facebook, Reddit, Bluesky, Threads, and news or commentary forums) and covers events such as elections, conflicts, migration, and gender rights (Naseem et al., 2026b). The associated POLAR benchmark paper describes a large-scale dataset with polarization annotated along three axes: detection, type, and manifestation (Naseem et al., 2026b).

While the shared task includes three subtasks covering polarization detection, type classification, and manifestation identification (Naseem et al., 2026a), this work focuses exclusively on Subtask 1 (Polarization Detection).

We participate only in Subtask 1 and evaluate English and Spanish. The input is a short post and the output is a binary label (Naseem et al.,

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1462–1468

2026a). Given a post, the system predicts a polarization label where 1 indicates polarized language and 0 indicates non-polarized language. On the test split, English contains 1452 posts and Spanish contains 1488 posts. The official evaluation metric for the shared task is macro-F1 (Naseem et al., 2026a), which we report alongside accuracy for interpretability. For additional analysis, we also report class-1 precision, recall, and F1, together with confidence summaries (mean confidence, confidence on correct predictions, confidence on incorrect predictions) and Brier score for calibration. Since macro-F1 gives equal weight to each class, it is more appropriate than accuracy alone when class distributions differ across languages.

More broadly, NLP research on political polarization has increasingly adopted text-based computational methods (Németh, 2023). Recent studies comparing these architectures find that while large generative models demonstrate strong zero-shot capabilities, smaller fine-tuned encoders like XLM-R consistently achieve superior performance on text classification tasks when sufficient application-specific training data is available (Bucher and Martini, 2024). Furthermore, recent work suggests that LLMs are often more effective when used as synthetic data generators to teach smaller, more efficient encoder models rather than being used as direct classifiers themselves (Pecher et al., 2026).

3 Approach

Our first system fine-tunes `xlm-roberta-base` (Conneau et al., 2020) for binary classification using HuggingFace Transformers.

Our second system uses `meta-llama/Llama-2-7b-chat-hf` (Touvron et al., 2023) in an instruction-style prompting setup.

3.1 Efficiency, interpretability, and analysis metrics

For efficiency tracking, we wrap both training and inference with CodeCarbon (Lacoste et al., 2019) to estimate runtime emissions. Because XLM-R is our best-performing system, we focus the current interpretability analysis on this model using SHAP (Lundberg and Lee, 2017). We sample examples from true positives, true negatives, false positives, and false negatives, and save per-example SHAP text explanations as HTML files. A comparable logit-based interpretability analysis for LLaMA-2 is feasible, but we leave it for future work due to time constraints.

The official metric for Subtask 1 is macro-F1, which we report as the primary measure alongside accuracy. For additional analysis, we also report class-1 precision, recall, and F1 derived from the confusion matrix, confidence summaries from the per-example prediction logs, and Brier score (Brier, 1950) to assess probability calibration.

4 Experimental Setup

4.1 Data and splits

We participated in Subtask 1 and report results for English (ENG) and Spanish (SPA) only. For both languages, we used the official training split to build our systems and the official test split for final evaluation. For XLM-R, we created an internal validation split by randomly splitting the training data into 80% train and 20% validation (seed 42) to select the best checkpoint. For LLaMA-2, few-shot examples were sampled from the training split, and we selected the final few-shot setting ($k = 4$) based on development-set performance before running on the test split. We experimented with few-shot prompting using $k \in \{3, 4, 5, 6\}$. Development results are included in our Appendix Table 2.

4.2 Preprocessing

We apply no manual text cleaning (no lowercasing, URL removal, or normalization) and use the raw text as provided. For XLM-R, inputs are tokenized with truncation to a maximum sequence length of 256 tokens. For LLaMA-2, prompts are tokenized with truncation to a maximum sequence length of 1024 tokens. At the raw-text level, no ENG or SPA test post exceeds 256 whitespace-separated words, and therefore none exceeds 1024. The longest ENG post contains 56 words, while the longest SPA post contains 25 words.

4.3 XLM-R fine-tuning

Training uses a batch size of 16, learning rate 1×10^{-5}, weight decay 0.01, and 5 epochs. We evaluate once per epoch and select the best checkpoint using validation macro-F1, which is the task's main metric. The final model for each language is the best checkpoint saved at the end of training.

For inference on the test split, we run the fine-tuned classifier over each example and output predictions in the required submission format (`id`, `polarization`). We also output a per-example

behavior log that records the predicted label, softmax probabilities for classes 0 and 1, the maximum probability as a confidence value, and the logit gap between class 1 and class 0.

4.4 LLaMA-2 prompting

We prompt the model using an instruction-style template that defines polarization and constrains the output to a single character label (0 or 1). We use separate instruction prompts for English and Spanish. The full prompt templates are provided in Appendix Sections A.5 and A.6.

Our submitted configuration uses few-shot prompting with $k = 4$ examples. Few-shot examples are sampled from the training split with an approximately balanced mix of class 0 and class 1 using a fixed seed (42). Decoding uses greedy generation (`do_sample=False`) with `max_new_tokens=2`. Outputs are parsed by first checking the first generated character, and if needed falling back to searching the output text for a 0 or 1; if no digit is found, we default to label 0 to ensure a valid submission file. We additionally log a per-example behavior file including the predicted label and a confidence score based on the model's first-step probabilities for generating "0" versus "1" (and the logit gap).

The $k = 4$ few-shot examples were static, meaning the same set of examples was included in every prompt during test set evaluation to ensure consistent behavior across inputs.

4.5 Compute environment

XLM-R training and inference were run locally on a machine with an NVIDIA GTX 1080 Ti GPU. LLaMA-2 prompting runs were executed in Google Colab (free tier) using an NVIDIA T4 GPU. Where applicable, we fixed random seeds (42) to improve reproducibility. We used LLaMA-2 7B to establish a baseline for locally-deployable, open-source generative models, despite the availability of larger models like LLaMA-3, as it allows for a more direct efficiency comparison with our encoder-based approach. Because XLM-R and LLaMA-2 were run on different machines (local GTX 1080 Ti vs Colab T4), direct runtime and emissions comparisons should be interpreted as indicative rather than strictly controlled. This hardware split was due to limited computational resources, and we plan to repeat these measurements on a single machine in future work.

We tracked training and inference emissions using CodeCarbon and logged XLM-R training runs using Weights & Biases.

5 Results

The official evaluation metric for Subtask 1 is macro-F1, so we report macro-F1 as the primary measure alongside accuracy. Table 1 summarizes our results for English (ENG) and Spanish (SPA). We evaluate on the official test split using the released gold labels.

On the official leaderboard, our submission ranked 38/44 for English and 32/37 for Spanish.

We selected $k = 4$ based on development-set performance and report test results here. Averaged across English and Spanish, XLM-R achieves 0.760 accuracy and 0.753 macro-F1, while LLaMA-2 achieves 0.7018 accuracy and 0.6845 macro-F1.

The best development scores were in the same overall range as the test results for both models: reported as accuracy / macro-F1, XLM-R changed from 0.7688 / 0.7489 to 0.7810 / 0.7681 in ENG and from 0.6727 / 0.6727 to 0.7386 / 0.7384 in SPA, while LLaMA-2 changed from 0.7313 / 0.7124 to 0.7376 / 0.7193 in ENG and from 0.6364 / 0.6099 to 0.6660 / 0.6497 in SPA.

Overall, the fine-tuned XLM-R classifier performs best across both languages. The prompted LLaMA-2 system is competitive on English but drops on Spanish, which reduces its average performance across the two languages.

5.1 Additional metrics and error patterns

To better understand the models' behavior, we compute class-1 precision, class-1 recall, and class-1 F1 (binary), derived from the confusion matrices produced by our evaluation script.

The Spanish LLaMA-2 results indicate a strong tendency to predict the positive class: the model produces many false positives (FP=418) compared to true negatives (TN=335), which increases recall for class 1 but lowers precision and reduces macro-F1.

XLM-R shows a more balanced precision-recall tradeoff across both languages, with higher macro-F1 overall (Table 1).

Class 0 metrics are reported in Appendix Table 4.

Metric	Base. (ENG)	XLM-R (ENG)	LLaMA-2 (ENG)	Base. (SPA)	XLM-R (SPA)	LLaMA-2 (SPA)
Accuracy	—	0.7810	0.7376	—	0.7386	0.6660
Macro-F1	0.7802	0.7681	0.7193	0.7266	0.7384	0.6497
Precision_1	—	0.6863	0.6387	—	0.7430	0.6108
Recall_1	—	0.7430	0.6567	—	0.7197	0.8925
F1_1	—	0.7135	0.6475	—	0.7312	0.7253
Mean conf $\pm$ SD	—	0.8439 ± 0.1377	0.8178 ± 0.1515	—	0.8147 ± 0.1249	0.9229 ± 0.1067
Conf (correct) $\pm$ SD	—	0.8666 ± 0.1298	0.8304 ± 0.1533	—	0.8367 ± 0.1159	0.9138 ± 0.1193
Conf (wrong) $\pm$ SD	—	0.7581 ± 0.1328	0.7826 ± 0.1407	—	0.7525 ± 0.1286	0.9410 ± 0.0724
Mean Brier $\pm$ SD	—	0.1514 ± 0.2495	0.2177 ± 0.2871	—	0.1820 ± 0.2617	0.3960 ± 0.4229

Table 1: Summary of test-set performance, class-1 metrics, and calibration statistics compared to the official POLAR baselines. Confidence and Brier rows are reported as mean $\pm$ standard deviation across examples.

6 Analysis

This section analyses the main performance differences between the two approaches and highlights typical model behavior beyond the headline macro-F1 scores (Table 1).

6.1 Cross-lingual performance and class bias

Across both languages, the fine-tuned XLM-R system outperforms the prompted LLaMA-2 system (Table 1). The gap is modest in English (macro-F1 0.7681 vs 0.7193), but widens in Spanish (0.7384 vs 0.6497). This suggests that few-shot prompting transfers less reliably to Spanish.

The class-1 metrics in Table 1 show that the Spanish drop is largely explained by a strong positive-class bias in LLaMA-2. In Spanish, LLaMA-2 reaches very high recall for class 1 (Rec_1 0.8925) but at the cost of low precision (Prec_1 0.6108), indicating many false positives. The confusion matrix confirms this: FP=418 compared to TN=335, meaning the model predicts class 1 for the majority of Spanish inputs. Full confusion matrices are provided in Appendix Table 3. In contrast, the Spanish gold distribution is close to balanced (753 class-0 vs 735 class-1), so this behavior reflects a genuine skew in the model's decisions rather than class imbalance in the data.

In English, LLaMA-2 is less skewed: its predicted positive rate is close to the true positive rate, and the main limitation is lower recall for class 1 compared to XLM-R (Rec_1 0.6567 vs 0.7430). Overall, these patterns suggest that prompting yields a workable baseline in English, but in Spanish it tends to treat many non-polarized posts as polarized.

This Spanish bias could be mitigated through logit-based threshold calibration. Since our pipeline already records the logit gap between the "0" and "1" tokens, requiring a specific margin to predict the polarized class - rather than relying on a default greedy selection - might reduce the high false positive rate. Furthermore, the performance drop suggests that LLaMA-2 may misinterpret Spanish emotional intensity or evaluative markers as identity-based hostility, potentially due to the English-centric nature of its underlying training and the translated instructions.

6.2 Behavioral analysis and confidence calibration

Both systems output per-example confidence signals as described in Section 3. For XLM-R, confidence is derived from the classifier softmax probability. For LLaMA-2, confidence is derived from the first-step logits for the digit tokens "0" and "1". [1]

A clear behavioral difference is that LLaMA-2 is more overconfident in Spanish than in English. On Spanish test data, the mean predicted confidence is high (0.923), but errors often have even higher confidence than correct predictions (mean confidence wrong 0.941 vs correct 0.914). This matches the large number of Spanish false positives and suggests poor calibration: the model frequently assigns strong confidence to polarized predictions that are actually non-polarized.

This is also reflected by Brier score (Brier, 1950) (lower is better), which measures how well predicted probabilities match the true labels. Unlike macro-F1 and accuracy, which only evaluate the final predicted label, Brier score also captures the quality of the predicted probability itself, making it useful for assessing calibration as well as correctness. LLaMA-2 has a much higher mean Brier score in Spanish (0.396) than in English (0.218). In comparison, XLM-R is lower on Spanish (0.182), supporting the observation that the fine-tuned classifier is better calibrated in this setting. Calibration summary statistics are provided in Table 1.

[1] In our test runs, this confidence signal was available for all examples.

6.3 Interpretability with SHAP

As XLM-R is our best-performing model, we focus our interpretability analysis on this system. Overall, the explanations align with the task definition: the model assigns strong positive contributions to identity-group references and inflammatory terms, particularly when appearing in combination.

For an English polarized true positive (e.g., *"they've lost control of illegal immigration..."*), XLM-R predicted class 1 with predicted probability $p_1 = 0.755$. SHAP analysis showed that 'illegal' made the strongest positive contribution ($+0.339$), followed by 'they' ($+0.115$) and 'lost' ($+0.064$). Interestingly, 'immigration' made a small negative contribution (-0.031), slightly pulling the prediction toward the non-polarized class despite the final correct label.

In a Spanish false negative containing identity terms (e.g., *"él es de origen judío, jesús nació en palestina,"*), the model incorrectly predicted class 0 ($p_1 = 0.198$). SHAP indicates that this occurs because negative contributions from terms like 'origen' (-0.094) and 'es' (-0.026) outweigh the comparatively weak positive signals from identity-related subwords such as 'je' ($+0.057$), 'palestin' ($+0.046$), and 'judi' ($+0.014$). These patterns are consistent with the error trends in Table 1, indicating that Spanish posts containing identity terms without explicitly inflammatory or derogatory descriptors can be misclassified as non-polarized. A comparable logit-based interpretability analysis for LLaMA-2 is feasible, but we leave it for future work due to time constraints. Across examples, identity-related words and negative descriptors contributed most to correct polarized predictions, while neutral terms often pushed predictions toward non-polarized.

6.4 Efficiency and environmental impact

We tracked XLM-R training and inference, and LLaMA-2 inference, using CodeCarbon. Reported inference time covers the prediction loop after model loading and setup. Because these measurements were collected on different hardware, the timing and emissions differences reported here should be interpreted as setup-specific rather than as a fully controlled benchmark.

For XLM-R, fine-tuning took approximately 4.7 minutes for English (283 s, 0.0097 $kgCO_2e$) and 3.8 minutes for Spanish (229 s, 0.0080 $kgCO_2e$) on an NVIDIA GTX 1080 Ti. XLM-R inference over the full test set took 14.6 seconds for English (0.00050 $kgCO_2e$) and 13.4 seconds for Spanish (0.00046 $kgCO_2e$).

For LLaMA-2 with Unsloth 4-bit quantization on a Colab T4 GPU, inference took 10.3 minutes for English (0.0069 $kgCO_2e$) and 11.8 minutes for Spanish (0.0082 $kgCO_2e$).

Overall, at inference time the prompted LLaMA-2 pipeline is roughly 48x slower and about 16x higher in estimated emissions across ENG+SPA compared to XLM-R inference. However, XLM-R incurs an upfront training cost (0.0177 $kgCO_2e$ across both languages) that is comparable to a single LLaMA-2 inference pass across both languages (0.0151 $kgCO_2e$). In practice, if the fine-tuned XLM-R model is reused for multiple inference runs, its training cost is amortized and the encoder-based approach becomes substantially more efficient. This highlights that the XLM-R approach is not only more accurate in this multilingual setting but is also significantly more sustainable for large-scale production deployment, providing a more meaningful comparison in terms of real-world utility. A summary of CodeCarbon measurements is provided in Appendix Table 5.

7 Conclusion

We presented our system for SemEval-2026 Task 9 (POLAR), Subtask 1, and compared fine-tuning XLM-RoBERTa with prompting LLaMA-2 7B for polarization detection in English and Spanish. Overall, XLM-R achieved higher macro-F1 and accuracy across both languages, while LLaMA-2 was competitive in English but dropped substantially in Spanish. Our analysis suggests that this drop is driven by a strong positive-class bias in Spanish, which increases recall but produces many false positives and poor calibration.

From an efficiency perspective, LLaMA-2 inference was substantially slower and higher in estimated emissions than XLM-R inference in our setup, while XLM-R incurs an upfront training cost that can be amortized when the model is reused. In future work, we would test additional languages, explore prompt variants and calibration methods to reduce the Spanish false positive rate, and extend analysis to Subtasks 2 and 3, where the POLAR benchmark indicates substantially higher difficulty (Naseem et al., 2026b). Polarization detection models may support moderation and research, but they can also be misused for censorship or to target spe-

cific groups. We recommend using such systems with human oversight and auditing for bias across languages, topics, and identity groups.

Acknowledgments

We thank the organizers of SemEval-2026 Task 9 (POLAR) for creating the dataset, running the shared task, and providing the evaluation infrastructure.

References

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Martin Bucher and Marco Martini. 2024. Fine-tuned 'small' LLMs (still) significantly outperform zero-shot generative AI models in text classification. *Preprint*, arXiv:2406.08660.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, 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. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *Preprint*, arXiv:1910.09700.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"urk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Renáta Németh. 2023. A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of Computational Social Science*, 6(1):289–313.

Branislav Pecher, Jan Cegin, Robert Belanec, Ivan Srba, Jakub Simko, and Maria Bielikova. 2026. Better as generators than classifiers: Leveraging LLMs and synthetic data for low-resource multilingual classification. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2840–2857.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

A Appendix

A.1 Development set few-shot sweep (LLaMA-2)

Lang	k	Accuracy	Macro-F1
ENG	3	0.7250	0.6599
ENG	4	0.7313	0.7124
ENG	5	0.6750	0.6737
ENG	6	0.6750	0.6732
SPA	3	0.5697	0.5666
SPA	4	0.6364	0.6099
SPA	5	0.5273	0.3772
SPA	6	0.5576	0.4384

Table 2: Development set few-shot sweep for LLaMA-2 (k=3 to k=6), evaluated on the development gold labels.

A.2 Confusion matrices (test split)

Lang	Model	TN	FP	FN	TP
ENG	XLM-R	738	181	137	396
ENG	LLaMA-2	721	198	183	350
SPA	XLM-R	570	183	206	529
SPA	LLaMA-2	335	418	79	656

Table 3: Counts derived from the confusion matrix on the test split.

A.3 Class 0 metrics (test split)

Lang	Model	Precision_0	Recall_0	F1_0
ENG	XLM-R	0.8434	0.8030	0.8227
ENG	LLaMA-2	0.7976	0.7845	0.7910
SPA	XLM-R	0.7345	0.7570	0.7456
SPA	LLaMA-2	0.8092	0.4449	0.5741

Table 4: Class 0 metrics on the test split.

A.4 Efficiency summary (CodeCarbon)

Model	Stage	Lang	Time	Emissions (kgCO_2e)
XLM-R	Train	ENG	4.7 min	0.0097
XLM-R	Train	SPA	3.8 min	0.0080
XLM-R	Inference	ENG	14.6 s	0.00050
XLM-R	Inference	SPA	13.4 s	0.00046
LLaMA-2	Inference	ENG	10.3 min	0.0069
LLaMA-2	Inference	SPA	11.8 min	0.0082

Table 5: CodeCarbon estimates in our setup (inference time excludes model loading and setup). XLM-R and LLaMA-2 were run on different machines, so these comparisons are indicative only.

A.5 LLaMA-2 Prompt (English)

Task: Attitude polarization detection. Output 1 if the text shows strong us-vs-them framing, contempt, blame, or generalised negative claims toward a group or side. Output 0 otherwise. Rule: Return ONLY a single character: 0 or 1.

A.6 LLaMA-2 Prompt (Spanish)

Instrucciones: Detecta la polarización de actitudes en el siguiente texto. Si el texto muestra una fuerte comparación de 'nosotros contra ellos', desprecio, culpa o declaraciones negativas generalizadas realizadas por un grupo o un bando, el resultado tiene que ser 1. En caso contrario el resultado tiene que ser 0. Regla: Devolver SOLO un carácter: 0 o 1