NAMAA at SemEval-2026 Task 9: Comparing Generative, Retrieval-Augmented, and Discriminative Methods for Arabic Online Polarization Detection and Type Classification

Abdelbasset Djamai[1,6] **Sahara Hussam Al-Madi**[2,6] **Norah Naji Al-Zaid**[3,6]
Khloud Al Jallad[4,6] **Mona A. Azim**[5,6]

[1]Datategy, [2]Linguistic Security Institute, [3]King Khalid University,
[4]Arab International University, [5]Ain Shams University, [6]NAMAA Community
contact@adjamai.com

Abstract

Detecting polarization in online discourse is important for understanding social fragmentation, yet it remains difficult for Arabic due to dialect variation, informal writing, and implicit framing. In this paper, we study Arabic polarization modeling in the SemEval-2026 Task 9 (POLAR) setting, focusing on polarization detection (ST1) and polarization type classification (ST2). We compare three approaches: encoder fine-tuning, zero-shot prompting, and retrieval-augmented in-context learning (RAG-ICL), across six Arabic encoders and different LLMs. For ST1, RAG-ICL with Gemma-3-27b-it achieves the best result (test macro F1 = 0.83), while remaining competitive with the best fine-tuned encoder (0.82), and substantially outperforming zero-shot prompting. For ST2, a pipeline that first applies the best ST1 encoder as a hard filter and then performs RAG-ICL achieves an $F1_{macro}$ = 0.62. Prompt-language effects are model- and task-dependent, with some settings doing better with English prompts and others with Arabic prompts. Chain-of-thought, self-refinement, and contrastive prompting do not outperform standard RAG-ICL.

1 Introduction

Polarizing online content has become a critical concern on social media, with effects on public debate and social cohesion across many settings (Cinelli et al., 2021; Naseem et al., 2026b). This concern is also clear in Arabic-speaking spaces, where political, religious, and social tensions are often discussed in highly charged language. Despite this, building reliable detectors for Arabic is still difficult. The language is morphologically intricate, highly inflected, and diglossic, and posts often mix Modern Standard Arabic (MSA) with regional dialects (Levantine, Maghrebi, Egyptian, Gulf). In practice, texts also include informal spelling and code-switching (Hamed et al., 2025), as well as indirect signals such as sarcasm and euphemistic wording (Abu Farha et al., 2021, 2022).

While significant progress has been made in hate speech and offensive language detection for Arabic (Mubarak et al., 2020; Zaghouani et al., 2024), polarization detection remains a distinct and less-explored problem. Unlike hate speech, which targets individuals or groups with explicit hostility, polarization captures a broader phenomenon of "us vs. them" framing that may manifest subtly through narrative construction, selective emphasis, or implicit group vilification.

SemEval-2026 Task 9 (POLAR) (Naseem et al., 2026b) is a multilingual shared task on online polarization with three subtasks (ST): polarization detection (ST1), polarization type classification (ST2), and polarization manifestation identification (ST3). In this work, we tackle Arabic polarization through the POLAR framework, focusing on the Arabic track of ST1 and ST2. To study this setting from complementary angles, we evaluate three modeling families: zero-shot prompting with large language models (LLMs), retrieval-augmented in-context learning (RAG-ICL), and encoder fine-tuning. This approach provides a direct comparison between generative, retrieval-augmented, and discriminative approaches, and helps clarify how retrieval quality, prompt design, and model architecture affect performance across the two subtasks.

Our results show that domain-adapted encoder fine-tuning achieves competitive ST1 performance with RAG-ICL, making encoders a practical alternative when the higher inference cost of retrieval-augmented generation is a concern. We also find that LLM-based data augmentation helps ST1

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3562–3590

encoder performance, but its impact is model-dependent rather than uniformly positive. For ST2, RAG-ICL with pipeline filtering (using the best ST1 encoder to filter non-polarized texts) yields the best results. Prompt-language effects are not uniform across models and tasks, and pipeline filtering substantially reduces hallucination on non-polarized samples. Other evaluated methods–Chain-of-thought, self-refinement, and contrastive prompting–do not surpass standard RAG-ICL.

2 Related Work

Computational approaches to polarization have primarily focused on network-level phenomena such as echo chambers and ideological sorting, with text-level detection remaining less developed (Cinelli et al., 2021; Barberá, 2015). Closest to our work are stance detection and political ideology classification (Mohammad et al., 2016; Iyyer et al., 2014). POLAR is the first shared task to frame polarization as a multilingual, multicultural, and multievent text classification problem spanning 22 languages (Naseem et al., 2026b,a). Our work focuses on the Arabic partition of this dataset.

Arabic NLP has advanced considerably with pretrained encoder-only models such as AraBERT, MARBERT, and CAMeLBERT (Antoun et al., 2020; Abdul-Mageed et al., 2021; Inoue et al., 2021), applied to tasks such as sentiment analysis, offensive language, hate speech, and stance detection (Mubarak et al., 2020; Zaghouani et al., 2024; Alturayeif et al., 2024). Polarization detection in Arabic, however, remains underexplored.

LLMs have been applied to text classification through a range of prompting strategies, from zero-shot and few-shot in-context learning to chain-of-thought reasoning and self-refinement (Wang et al., 2025; Singh et al., 2025). Retrieval-augmented approaches extend ICL by dynamically selecting relevant demonstrations (Pradhan et al., 2026). Data augmentation via LLMs has also been explored to expand training data for numerous tasks (Arık et al., 2026; Cegin et al., 2025). We empirically compare all of these approaches for Arabic online polarization detection and type classification.

3 Task Description

POLAR (Naseem et al., 2026b) is a shared task on multilingual, multicultural, and multi-event online polarization encompassing 22 languages across three subtasks. ST1 is binary polarization detection: given a text, predict polarized (1) or not (0). ST2 is multi-label polarization type classification, assigning one or more labels from *Political*, *Racial/Ethnic*, *Religious*, *Gender/Sexual*, or *Other*. ST3 is multi-label polarization manifestation identification, classifying how polarization is expressed via labels such as *Stereotype*, *Vilification*, *Dehumanization*, *Extreme Language*, *Lack of Empathy*, or *Invalidation*. This work focuses exclusively on the Arabic track of ST1 and ST2.

4 Methods

Figure 1 provides an overview of all methods evaluated for both subtasks.

4.1 Subtask 1

4.1.1 Encoder Fine-Tuning

We fine-tune six Arabic BERT-based encoders—AraBERT v02, AraBERT v02-Twitter (Antoun et al., 2020), MARBERT v2 (Abdul-Mageed et al., 2021), QARiB (Abdelali et al., 2021), CAMeLBERT-Mix (Inoue et al., 2021), and AraModernBERT (Elshehy et al., 2026)—for ST1 binary classification. Each model is trained under two data conditions: the original training set and an augmented version. The augmented data is generated with DeepSeek V3.2 (DeepSeek-AI et al., 2025) through a two-phase pipeline. First, a random sample of training texts is analyzed to produce a reusable style guide capturing dialect variation, tonal patterns, and writing conventions. This guide then conditions the generation of two sample types per polarized training instance: *hard negatives*—offensive or critical texts rewritten to remove group-directed hostility while preserving surface form—and *paraphrases*—rewritings that preserve the polarization signal under dialect or stylistic variation. The prompts used for augmentation, alongside all prompts used in later experiments, are provided in Appendix F. The best-performing encoder is then used as the pipeline filter in all ST2 variants that apply filtering.

The augmented data is generated with DeepSeek V3.2

4.1.2 Zero-Shot Prompting

We evaluate five open-weight LLMs—DeepSeek V3.2 (DeepSeek-AI et al., 2025), Qwen3-235B-A22B (Yang et al., 2025), Gemma-3-27b-it (Team et al., 2025b), Llama-3.3-70B-Instruct (Dubey et al., 2024), and Fanar, an Arabic-centric model

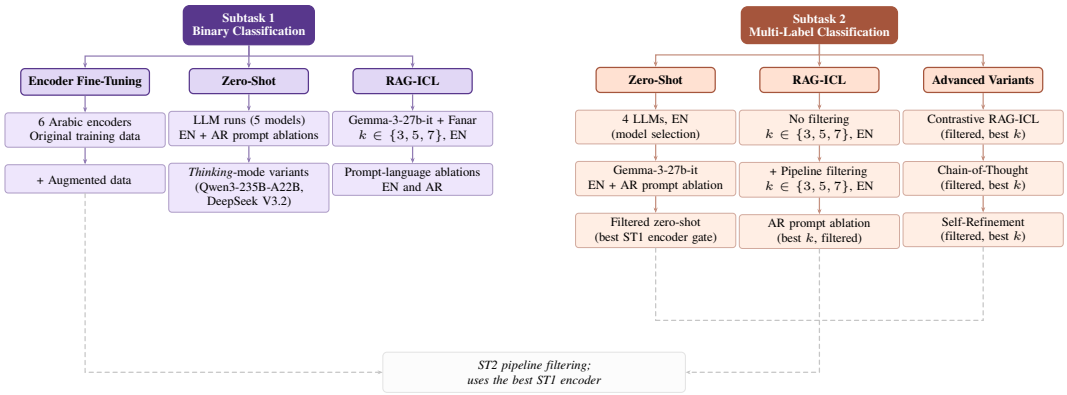

Figure 1: Overview of the ST1/ST2 experiment space, including method families, configuration settings, and ablations.

(Team et al., 2025a). Each model receives a task description, a definition of polarization, and the input text, with no labeled examples, and must produce a binary label. Primary runs use English prompts, and we add Arabic-prompt ablations for both the best-performing model and Fanar. We additionally test a *thinking*-mode variant for models that support it, in which internal step-by-step reasoning precedes the final answer, to assess whether extended reasoning benefits polarization detection.

4.1.3 Retrieval-Augmented In-Context Learning

RAG-ICL grounds each prediction in semantically similar labeled examples retrieved from the training set. All training instances are encoded offline with a multilingual sentence encoder and stored in a vector index. At inference time, the k nearest neighbors of a test input are retrieved by cosine similarity and inserted before the input as labeled demonstrations. Retrieved examples carry binary labels. We evaluate the best-performing LLM together with Fanar using English prompts, and we additionally test Arabic-prompt ablations across the evaluated models and k values. We set $k \in \{3, 5, 7\}$ to study the effect of the number of retrieved demonstrations.

4.2 Subtask 2

4.2.1 Zero-Shot Prompting

We use the same prompting structure as in ST1, adapted for multi-label prediction: the prompt defines all five polarization categories and asks the model to output the applicable label set. We first compare four LLMs—Gemma-3-27b-it, Qwen3-235B-A22B, Llama-3.3-70B-Instruct, and Fanar—under zero-shot conditions with English prompts and no filtering, in order to select the primary best-performing model for the remaining experiments. Based on this comparison, we perform targeted zero-shot ablations: (1) English vs. Arabic prompting for both Fanar and the top-performing LLM, and (2) pipeline filtering for the best-scoring model under English prompting.

Pipeline Filtering. LLMs may assign non-zero type labels to non-polarized inputs, thereby inflating false positives. To mitigate this, we apply a two-stage pipeline. First, the best fine-tuned ST1 encoder predicts whether an input is polarized. Inputs predicted as non-polarized are assigned an all-zero type vector without invoking the LLM, while inputs predicted as polarized proceed to LLM-based type classification.

4.2.2 Retrieval-Augmented In-Context Learning

We use the same indexing and retrieval procedure as in ST1 (Section 4.1.3), except that retrieved examples now carry full multi-label annotations. The LLM produces a single prediction over all five categories in one call—rather than one call per category—to avoid inter-category inconsistencies. We evaluate RAG-ICL both with and without pipeline filtering across all k values. For the best configuration, we additionally run an Arabic-prompt ablation.

We further evaluate three prompting variants under the best-performing setting. *Contrastive RAG-ICL* retrieves both positive and negative examples for each category to expose the model to the decision boundary. *Chain-of-thought* (CoT) prompting asks the model to reason through each category before predicting. *Self-refinement* is a two-pass approach in which the model first generates a predic-

Model	ST1			ST2	
	FT	ZS	RAG	ZS	RAG+
Encoders (×6)	✓	—	—	*gate only*	
Large Language Models					
DeepSeek V3.2	—	✓	✗	✗	✗
Qwen3-235B-A22B	—	✓	✗	✓	✗
Fanar	—	✓	✓	✓	✗
Gemma-3-27b-it	—	✓	✓	✓	✓
Llama-3.3-70B-Instruct	—	✓	✗	✓	✗

Table 1: Model coverage per setting and subtask. **FT** = fine-tuning (encoder models only); **ZS** = zero-shot; **RAG+** = RAG-ICL and advanced variants. ✗ = not evaluated in that setting.

Model	Data	$F1_{mac}$	Prec.	Rec.
AraBERT v02	Orig.	0.7684	0.7143	0.8000
AraBERT v02	Aug.	0.7986	0.7303	0.8667
AraBERT v02-Twitter	Orig.	0.8260	0.8108	0.8000
AraBERT v02-Twitter	Aug.	**0.8516**	0.7976	**0.8933**
MARBERT v2	Orig.	0.8098	0.7590	0.8400
MARBERT v2	Aug.	0.7905	0.7632	0.7733
QARiB	Orig.	0.7806	0.7209	0.8267
QARiB	Aug.	0.8014	0.7917	0.7600
CAMeLBERT-Mix	Orig.	0.7887	**0.8596**	0.6533
CAMeLBERT-Mix	Aug.	0.7822	0.7826	0.7200
AraModernBERT	Orig.	0.7557	0.7125	0.7600
AraModernBERT	Aug.	0.7495	0.7089	0.7467

Table 2: ST1 encoder fine-tuning results on dev. Best model (AraBERT v02-Twitter + Aug.) is used as the ST2 pipeline filtering gate.

tion with reasoning and then reviews and revises it.

5 Experimental Setup

Dataset. We use the Arabic partition of the PO-LAR dataset (Naseem et al., 2026b). The split contains 3,380 training, 169 development (dev), and 1,521 test samples. In the development set, 44.4% (75/169) are polarized. For ST2, the label counts are Political (42), Racial/Ethnic (29), Other (28), Gender/Sexual (18), and Religious (14). We use the training split for both encoder fine-tuning and the RAG-ICL retrieval index. The development split is used for all evaluations and ablations. See Appendix A for a detailed dataset analysis.

Models. Table 1 summarizes the models used in each setting. For ST1, six Arabic BERT-based encoders (Section 4.1.1) are fine-tuned, and five LLMs are evaluated in the generative settings. For ST2, a subset of four LLMs (excluding DeepSeek V3.2) is evaluated. Full model identifiers and inference providers are given in Table 10 (Appendix B).

Experimental Settings. Embeddings are computed with `multilingual-e5-large` (Wang et al., 2024) using the prefix `"query: {text}"`. All LLM inference uses temperature $=$ 0.3; standard runs use `max_tokens` $=$ 256 and *thinking*-mode runs use 1024. Encoder fine-tuning hyperparameters along with other model/LLM details are given in Appendix B.

Evaluation. Our primary metric is macro F1. We also report precision and recall. For ST2, we additionally report hallucination rate (Hall.): the fraction of non-polarized inputs assigned at least one type label (lower is better). Additional metrics are provided in the full results tables in Appendix C.

6 Results

6.1 Subtask 1

6.1.1 Encoder Fine-Tuning

Table 2 reports fine-tuning results across all six encoders and both data conditions. AraBERT v02-Twitter trained on augmented data achieves the best dev performance ($F1_{mac}$=0.8516), and is used as the ST2 pipeline gate. We find that augmentation is beneficial but model-dependent: AraBERT variants and QARiB improve consistently, while MARBERT v2 and AraModernBERT show marginal degradation. CAMeLBERT-Mix achieves the highest precision with a much lower a recall.

6.1.2 Zero-Shot Prompting and RAG-ICL

Table 3 shows the main ST1 generative results. In the zero-shot setting, most models are conservative, achieving high precision (above 0.81) but low recall (below 0.20), which means they rarely predict the polarized class. Fanar is the exception, reaching an $F1_{mac}$ of 0.7905 with Arabic prompts. It maintains a much better balance between precision and recall, suggesting stronger zero-shot performance for Arabic text.

For models evaluated in both languages, Arabic prompts consistently perform better in the zero-shot setting. Fanar improves from 0.7524 to 0.7905 $F1_{mac}$, and Gemma-3-27b-it rises from 0.5113 to 0.5327, with both gains driven mostly by higher recall. Interestingly, *thinking* mode hurts performance for Qwen3-235B-A22B and DeepSeek V3.2, likely due to "overthinking". This suggests that explicit reasoning does not benefit this specific detection task.

RAG-ICL considerably improves Gemma-3-27b-it, boosting its recall from 0.17 to 0.84 and its F_1 by 0.33, with the best results at k=7. Across

Model	Method	Pr.	k	$F1_{mac}$	Prec.	Rec.
Fanar	Zero-Shot	EN	–	0.7524	0.7429	0.6933
Fanar	Zero-Shot	AR	–	**0.7905**	0.7632	0.7733
Gemma-3-27b-it	Zero-Shot	EN	–	0.5113	0.8125	0.1733
Gemma-3-27b-it	Zero-Shot	AR	–	0.5327	0.8333	0.2000
Llama-3.3-70B-Instruct	Zero-Shot	EN	–	0.4815	0.8333	0.1333
Qwen3-235B-A22B	Zero-Shot	EN	–	0.4735	0.9000	0.1200
Qwen3-235B-A22B	Zero-Shot†	EN	–	0.4331	0.7500	0.0800
DeepSeek V3.2	Zero-Shot	EN	–	0.4490	0.8750	0.0933
DeepSeek V3.2	Zero-Shot†	EN	–	0.4000	1.0000	0.0400
Gemma-3-27b-it	RAG-ICL	EN	3	0.8202	0.8000	0.8000
Gemma-3-27b-it	RAG-ICL	EN	5	0.8149	0.7821	0.8133
Gemma-3-27b-it	RAG-ICL	EN	7	**0.8446**	0.8182	0.8400
Fanar	RAG-ICL	AR	3	0.7546	0.7237	0.7333
Fanar	RAG-ICL	AR	5	0.7419	0.7162	0.7067
Fanar	RAG-ICL	AR	7	0.6230	0.6667	0.4267

Table 3: Key ST1 generative results on dev. †*Thinking* mode. Full results are given in Appendix C.

Model	Pr.	Filt.	$F1_{mac}$	Prec.	Rec.	Hall.↓
Gemma-3-27b-it	EN	✗	0.5425	0.5080	0.6070	0.6064
Gemma-3-27b-it	AR	✗	0.5315	0.4840	0.6037	0.8085
Gemma-3-27b-it	EN	✓	**0.5643**	**0.6711**	0.5366	**0.1702**
Fanar	EN	✗	0.4290	0.4507	0.4204	0.5957
Fanar	AR	✗	0.3817	0.4124	0.3959	0.7128
Qwen3-235B-A22B	EN	✗	0.5273	0.6151	0.4677	0.2447
Llama-3.3-70B-Instruct	EN	✗	0.4976	0.5482	0.4794	0.3511

Table 4: ST2 zero-shot results comparison on dev. Gemma-3-27b-it EN is selected as the primary ST2 model. Full results in Appendix C.

RAG-ICL runs, the effect of prompt language is mixed. Arabic prompts are stronger at lower k values, but English yields the best overall results for both Gemma-3-27b-it and Fanar at k=7. However, Fanar's performance generally worsens as k increases. Its zero-shot setup remains stronger than all RAG-based variants for each prompt language evaluated, suggesting that retrieved examples may often introduce noise rather than helpful guidance.

6.2 Subtask 2

6.2.1 Zero-Shot Prompting

Table 4 compares four LLMs in the zero-shot ST2 setting. Among the unfiltered zero-shot runs, Gemma-3-27b-it with an English prompt gives the best macro F1. Qwen3-235B-A22B is more precise and produces fewer hallucinations, but its lower recall reduces overall F1. Fanar performs worst overall in this comparison. For the two models tested in both languages, English prompts outperform Arabic prompts, with lower hallucination and higher macro F1 in each case. Applying pipeline filtering to Gemma-3-27b-it (EN) further reduces hallucination from 0.61 to 0.17 and improves macro F1 from 0.54 to 0.56, despite a small drop in recall. We therefore use this filtered configuration in the following experiments.

Method	Pr.	Filt.	k	$F1_{mac}$	Prec.	Rec.	Hall.↓
ZS (ref.)	*EN*	✓	–	*0.5643*	*0.6711*	*0.5366*	*0.1702*
RAG-ICL	EN	✗	3	0.5590	0.5442	0.5826	0.5213
RAG-ICL	EN	✗	5	0.5935	0.5447	0.6607	0.5745
RAG-ICL	EN	✗	7	0.5457	0.5190	0.5856	0.5745
RAG-ICL	EN	✓	3	0.6044	0.6620	0.5689	0.1702
RAG-ICL	EN	✓	5	0.6146	0.6539	0.5909	0.1702
RAG-ICL	EN	✓	7	**0.6304**	**0.6903**	0.5927	**0.1702**
RAG-ICL	AR	✓	7	0.6161	0.6562	0.5895	0.1702
CoT	EN	✓	7	0.6196	0.6790	0.5818	0.1702
Contrastive	EN	✓	7	0.6062	0.6521	0.5771	0.1702
Self-Refinement	EN	✓	7	0.5692	0.6129	0.5678	0.1702

Table 5: ST2 RAG-ICL and advanced prompting results on dev (Gemma-3-27b-it). *ZS (ref.)* is the best zero-shot result from Table 4.

6.2.2 RAG-ICL and Advanced Prompting

All RAG-ICL and advanced prompting results are summarized in Table 5. Without filtering, k=5 is optimal, while performance drops at k=7. A larger retrieved set likely brings in more non-polarized examples, which increases false positives. With pipeline filtering, this pattern becomes more stable and performance rises to $F1_{mac}$=0.6304 at k=7. The ST1 gate noticeably improves the hallucination rate, bringing the it down to 0.1702 across all filtered runs. At the best k, the prompt-language ablation shows that English prompts outperform Arabic prompts, mainly because Arabic prompts reduce precision.

None of the advanced variants beats standard RAG-ICL. CoT comes closest, with $F1_{mac}$=0.6196 versus 0.6304 for the standard setup. Contrastive prompting also stays below the baseline (0.6062), while self-refinement performs worst among the advanced variants (0.5692), only slightly above the filtered zero-shot baseline (0.5643). Overall, the best ST2 results come from standard filtered RAG-ICL, not from adding extra reasoning or revision steps. This suggests that such extensions may struggle to correct sufficiently complex errors, such as those arising from implicit framing or dialectal nuance.

6.3 Test Set Results

Table 6 shows the results for selected test-set runs, which largely mirror our findings on the development set. For ST1, the official encoder submission is competitive with the (non-submitted) RAG-ICL system, which scored slightly higher at 0.8314. For ST2, our official submission used the non-filtered zero-shot Gemma-3-27b-it system (0.5292). The non-submitted filtered versions performed better, with filtered zero-shot at 0.5686 and filtered RAG-ICL at 0.6196.

ST	Configuration	$F1_{mac}$	Off.
1	AraBERT-v02-Tw + Aug.	0.8249	✓
1	Zero-Shot (Gemma-3-27b-it, EN)	0.5221	✗
1	RAG-ICL (Gemma-3-27b-it, EN, k=7)	**0.8314**	✗
2	Zero-Shot (Gemma-3-27b-it, EN)	0.5292	✓
2	Zero-Shot (Gemma-3-27b-it, EN, filtered)	0.5686	✗
2	RAG-ICL (Gemma-3-27b-it, EN, k=7, filtered)	**0.6196**	✗

Table 6: Selected test-set runs. ✓ = official submission; ✗ = non-submitted.

7 Error and Linguistic Analysis

We manually audited 40 predictions across both subtasks (20 each) from the best system per subtask, reviewed by an Arabic linguist and a computational linguist. More details and annotated examples are in Appendix E.

ST1. The most common failure is conflating personal insults with group-level hostility. Subtle "us vs. them" framing and rhetorical provocations are sometimes missed, while false positives cluster around sports banter and analytically charged vocabulary. Dialect variation compounds this further, with Gulf, Egyptian, and Iraqi posts mostly affected by unresolved region-specific idioms.

ST2. Models rely too heavily on surface tokens over pragmatic intent. The *Religious* label overactivates on routine expressions like "اللّٰه" *(Allah/God)* in non-polarized posts, while texts without obvious keywords are under-detected. More critically, models lack deep cultural context, making them miss the sectarian weight of regional terms and producing incorrect label combinations even when polarization is correctly detected.

8 Conclusion

We compared encoder fine-tuning, zero-shot prompting, and retrieval-augmented in-context learning for Arabic polarization detection and type classification. For ST1, RAG-ICL with Gemma-3-27b-it achieved the best test performance, but fine-tuned encoders were very nearly as effective and offer a more efficient alternative. For ST2, the best setup used the highest-scoring ST1 encoder as a gate for pipeline filtering combined with RAG-ICL. This shows that filtering is crucial for reducing spurious type predictions on non-polarized text.

More broadly, our findings show that prompt-language effects are inconsistent across models and tasks, and that advanced prompting strategies do not always improve performance. This highlights the need for careful empirical validation. We hope this study provides a useful benchmark for future work on Arabic polarization, especially in improving cultural grounding and type-level reasoning for fine-grained classification.

Limitations

Our study has several limitations. First, the small development set (169 samples) may limit the reliability of our ablation results. Second, RAG-ICL is much slower than zero-shot inference (∼2.2 s/sample vs. 0.2 s) due to embedding and retrieval overhead, which could make real-time deployment difficult. Additionally, due to time constraints and the scope of the study, we did not investigate how different embedding models or retrieval quality affect performance, nor did we explore efficiency trade-offs. Finally, the observed prompt-language effects are model-specific and may change with stronger Arabic-aligned or multilingual models.

Ethics Statement

Polarization detection systems carry inherent risks. False positives may lead to unjustified content moderation, while false negatives may allow harmful content to persist. Moreover, our system operates at the text level and does not perform user profiling. The definition of polarization used in this work may not capture all social and cultural perspectives on what constitutes polarizing discourse. We encourage careful human oversight in any deployment context.

Acknowledgments

We thank the POLAR task organizers for their efforts and resources, and the reviewers for their constructive comments and valuable feedback.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *Preprint*, arXiv:2102.10684.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th*

International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 7088–7105.

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. Semeval-2022 task 6: isarcasmeval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814, Seattle, United States. Association for Computational Linguistics.

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nora Alturayeif, Hamzah Luqman, Zaid Alyafeai, and Asma Yamani. 2024. StanceEval 2024: The First Arabic Stance Detection Shared Task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 774–782, Bangkok, Thailand. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15.

Ahmet Okan Arık, Gizem Parlayandemir, and Serra Çelik. 2026. Llm-based data augmentation for text classification on imbalanced datasets: A case study on fake news detection. *Egyptian Informatics Journal*, 33:100886.

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2025. Llms vs established text augmentation techniques for classification: When do the benefits outweight the costs? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10476–10496.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Omar Elshehy, Omer Nacar, Abdelbasset Djamai, Muhammed Ragab, Khloud Al Jallad, and Mona Abdelazim. 2026. AraModernBERT: Transtokenized initialization and long-context encoder modeling for Arabic. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*, pages 313–321, Rabat, Morocco. Association for Computational Linguistics.

Injy Hamed, Caroline Sabty, Slim Abdennadher, Ngoc Thang Vu, Thamar Solorio, and Nizar Habash. 2025. A survey of code-switched Arabic NLP: Progress, challenges, and future directions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4561–4585, Abu Dhabi, UAE. Association for Computational Linguistics.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection

using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2020. Overview of arabic offensive language and hate speech detection. *arXiv preprint arXiv:2005.01845*.

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

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acart"urk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026b. 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.

Manaranjan Pradhan, Naga Vemprala, and Naveen Gudigantala. 2026. Beyond zero-shot: Enhancing llm financial complaint classification with relevancy-driven rag-based few-shot prompting.

Juhi Singh, Ziqiao Ao, and Sebastian Antinome. 2025. Optimizing prompt refinement: Algorithmic strategies for llm-driven text classification tasks. *Available at SSRN 5221492*.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025a. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025b. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Zhiqiang Wang, Yanbin Lin, Jiajun Shen, and Xingquan Zhu. 2025. A survey of large language models for text classification: What, why, when, where, and how. *Authorea Preprints*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Wajdi Zaghouani, Hamdy Mubarak, and Md. Rafiul Biswas. 2024. So hateful! building a multi-label hate speech annotated arabic dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15044–15055, Torino, Italia. ELRA and ICCL.

A Dataset Analysis

Figure 2 summarizes the statistics of the Arabic training dataset (3,380 instances), annotated for polarization detection (ST1, binary), type classification (ST2, 5-class multi-label), and manifestation identification (ST3, 6-class multi-label).

Polarization Detection – Class balance (A). The training set is near-balanced at the ST1 level (55.3 % non-polarized, 44.7 % polarized), which is favorable for fine-tuning. The development set shows a similar ratio (44.4 % polarized), so dev-set ablations are not distorted by class skew.

Table 7 shows six representative instances (three per class) to illustrate the range of content and the annotation difficulty. Non-polarized texts may still mention sensitive topics (religion, geopolitics) without containing any polarization markers like group-directed hostility.

Polarization Type Frequency (B). Among polarized instances, *Political* is the most frequent type (780), followed by *Racial/Ethnic* (583) and *Other* (565). *Religious* is the rarest (283), nearly three times less frequent than *Political*. This imbalance directly predicts the per-category results in Table 13: the best-performing categories (Political, Racial/Ethnic) are also the most represented.

Polarization Type Co-occurrence (C). Political and Racial/Ethnic co-occur most frequently (348 instances), forming the dominant label pair. The Political + Racial/Ethnic + Religious triplet is the most common three-label combination (88 instances). Gender/Sexual has relatively low co-occurrence with Political (68) but high co-occurrence with Other (175), suggesting it captures a qualitatively distinct phenomenon from overtly ideological polarization. These patterns motivate a joint multi-label prediction strategy over per-category binary calls.

Label Cardinality (D). ST2 labels are sparse: most polarized instances carry 1 or 2 labels (684 and 620, respectively), with very few carrying 4 or 5. ST3 manifestation labels are denser, peaking at 3 labels per instance and extending to 6, reflecting that a single polarizing text typically employs multiple rhetorical strategies simultaneously.

Text Length (E). Texts are short overall (median 14 words), consistent with a social-media origin. Polarized texts are modestly longer (median 15 vs. 13 words for non-polarized), suggesting that polarizing content requires more elaboration. The heavy right tail for non-polarized texts indicates some lengthy neutral posts, but the bulk of classification decisions involve short, dense texts where implicit cues dominate.

Manifestation Type Frequency (F). Although ST3 is outside the scope of this work, we include it for completeness. Vilification (1,256) and Stereotype (1,127) are the dominant manifestation types, together covering the vast majority of polarized instances. Extreme Language (1,027) is also prevalent. Dehumanization (370) and Invalidation (274) are rarer and likely more severe forms. The high frequency of Vilification and Stereotype aligns with the group-directed hostility patterns observed in our ST1 error analysis.

A.1 Lexical Fingerprints (ST1 & ST2)

Table 8 lists the most frequent content words (after Arabic stopword removal) for the ST1 binary split and for each ST2 type. Several patterns are noteworthy.

- **The dual role of "الله" (God).** The word "الله" is the most frequent unigram across almost all categories, including both polarized and non-polarized texts. While often used in everyday Arabic expressions, its high prevalence in polarized text reflects its use in curses, religious vilification, and sectarian othering (as seen in Examples 3 and 6).

- **Geopolitical salience.** Country and region names dominate polarized text: الخليج (the Gulf), سوريا (Syria), مصر (Egypt), تونس (Tunisia). These map directly onto the *Political* and *Racial/Ethnic* categories.

- **Sectarian markers in *Religious*.** الشيعة (Shia), الكنيسة (the church), المسلمين (Muslims), and السنة (Sunni) are the top terms for this label—all four directly name religious groups.

- **Gender discourse.** احلام (Ahlam, a pop-culture figure), المرأة (the woman), التحرش (harassment), and المثلية (homosexuality) surface under *Gender/Sexual*, spanning celebrity gossip to social issues.

- **Entertainment in *Other*.** احلام (Ahlam) and كاظم (Kazem, i.e. the singer Kazem Al-Saher) indicate that celebrity-related discourse is a notable source of polarization in Arabic social media.

#	Arabic (original)	English (translation)	Label
1	في الرياض.. لقاء ودي يجمع قادة دول الخليج والرئيس المصري وملك الأردن	"In Riyadh…a friendly meeting brings together Gulf leaders, the Egyptian president, and the King of Jordan."	0
2	افضل نادي بالكون برشلونة. اسطورة كرة القدم هو ليونيل اندريس ميسي. وخلص ختمنة كرة القدم	"The best club in the universe is Barcelona. The legend of football is Lionel Andrés Messi. End of discussion."	0
3	إن الله وزّع الأرزاق بالعدل، ولم يوزعها بالتساوي؛ لإن العدل مبدأ أسمى من المساواة	"God distributed sustenance with justice, not equally; for justice is a higher principle than equality."	0
4	أغنى دولة عربية تحارب أفقر دولة عربية اللهم العن أمراء البترول	"The richest Arab country is fighting the poorest Arab country—may God curse the oil princes."	1
5	علي المصريين مقاطعة الحديث في التليفون لانه اختراع الكفار.. ونعود للحمام الزاجل	"Egyptians should boycott phone calls since it's an infidel invention…let's go back to carrier pigeons."	1
6	أنت ذبابة تريد بتغريدتك هذه أن تفرّق بين المذاهب لا أكثر!! سبحان الله تعالى	"You are a fly trying to divide the sects with this tweet, nothing more!! Glory be to God."	1

Table 7: Examples of training instances (3 non-polarized, 3 polarized).

Category	Top unigrams (Arabic / English)
Polarized	الله (God), الخليج (Gulf), احلام (Ahlam), سوريا (Syria), العرب (Arabs)
Non-Pol.	الله (God), كأس (cup), العالم (world), احلام (Ahlam), نادي (club)
Political	الله (God), سوريا (Syria), الدولة (state), العرب (Arabs), النظام (regime)
Racial/Eth.	الخليج (Gulf), الله (God), العرب (Arabs), مصر (Egypt), سوريا (Syria)
Religious	الله (God), الشيعة (Shia), الكنيسة (church), المسلمين (Muslims), السنة (Sunni)
Gender/Sex.	الله (God), احلام (Ahlam), المرأة (woman), التحرش (harassment), المثلية (homosexuality)
Other	الله (God), احلام (Ahlam), الخليج (Gulf), كاظم (Kazem), علي (Ali)

Table 8: Top-5 content words per category (stopwords removed).

A.2 Qualitative Examples

Table 9 presents five representative polarized texts selected to span different topics, events, and label configurations.

The following observations can be made:

- Examples 1 and 2 show how political and regional identity polarization can surface even in seemingly mundane contexts (sports, diplomacy).
- Example 3 illustrates religious vilification with direct cursing.
- Example 4 highlights how gender-based harassment is a systemic, pragmatic phenomenon.
- Example 5 criticizes religious *education* rather than a religious group itself, carrying both *Political* and *Religious* labels and illustrating the blurred boundaries between categories.

B Experimental Setup Details

Encoder fine-tuning settings. We fine-tune all models for 12 epochs with a batch size of 128, using AdamW with a learning rate of 5×10^{-5}, weight decay of 0.1, and a warmup ratio of 0.1. Label smoothing is set to 0.1, and early stopping is applied with a patience of 3 epochs.

Software and hardware resources. Encoder fine-tuning is conducted on an NVIDIA P100 GPU using the Hugging Face (Transformers, Sentence-Transformers) and PyTorch frameworks. All LLM evaluations are performed through API (see below).

LLMs. Table 10 lists all LLMs evaluated with their full model identifiers and providers.

Name	Model ID	Provider
Qwen3-235B-A22B	`qwen/qwen3-235b-a22b`	NVIDIA NIM
DeepSeek V3.2	`deepseek-ai/deepseek-v3.2`	NVIDIA NIM
Gemma-3-27B-it	`google/gemma-3-27b-it`	NVIDIA NIM
Llama-3.3-70B-Instruct	`meta/llama-3.3-70b-instruct`	NVIDIA NIM
Fanar	`Fanar`	Fanar API

Table 10: Evaluated LLMs with full identifiers.

LLM inference configuration. All models use a temperature of 0.3 (0.7 for data augmentation). Standard and *thinking*-mode runs use `max_tokens` = 256 and 1024 respectively.

C Full Results Tables

Tables 11 and 12 provide the complete ranking of all generative / RAG-based experiments conducted for both Subtask 1 and Subtask 2, respec-

#	Topic	Arabic (original)	English (translation)	ST2
1	Sports / regime	فالنسيا قدم مباراة كبيرة فعل المستحيل ولم يتمكن من تحقيق الفوز، بينما نادي النظام قدم كرة مقرفة ومقززة وبقي ينتظر مساعدات الحكم	"Valencia played a great match and did the impossible but couldn't win, while the regime's club played disgusting football, waiting for referee help."	OTH
2	Gulf identity	ارجل واشجع من جميع دول الخليج..مواقف مخزية لدول الخليج	"Braver than all Gulf states…disgraceful stances by Gulf countries."	POL, RAC, OTH
3	Sectarianism	هذا جعلها شريكة لله عليك من الله ما تستحق. منافق مبتدع الاسلوب جديد من النفاق والله لن ينالك منه سوى الخزي في الدنيا والعذاب في الآخرة	"He made her a partner of God…you hypocrite innovator…you will earn nothing but disgrace in this life and torment in the hereafter."	REL
4	Gender	يقال أن سبب التحرش في الشارع ملابسها، يا ترى ما سبب التحرش بالمرأة في مواقع التواصل الاجتماعي..؟ جمال خطها مثلا	"They say street harassment is caused by her clothes—then what causes online harassment of women…her handwriting maybe?"	GEN
5	Religion / education	لا اوافق لان معلم الدين للاسف مش بيعلموا ولادنا سلوكيات حسنه دول هايعلموهم العنف والتطرف. قاتلوهم. اذبحوهم	"I disagree because religion teachers unfortunately don't teach our kids good behavior—they teach them violence and extremism: 'fight them', 'slaughter them'"	POL, REL

Table 9: Diverse polarized examples with original Arabic and English translation (by the authors. Label abbreviations: **POL** = Political, **RAC** = Racial/Ethnic, **REL** = Religious, **GEN** = Gender/Sexual, **OTH** = Other.

tively, on the development set, comparing all models, prompt variants, and configuration choices.

D Per-Category Breakdown (ST2)

Category	Prec.	Rec.	F1	TP	FP	FN	TN
Political	0.7556	**0.8095**	**0.7816**	34	11	8	116
Racial/Ethnic	0.6667	0.6897	0.6780	20	10	9	130
Religious	0.7500	0.6429	0.6923	9	3	5	152
Gender/Sexual	0.7500	0.5000	0.6000	9	3	9	148
Other	0.5294	0.3214	0.4000	9	8	19	133
Macro	*0.6903*	*0.5927*	*0.6304*	–	–	–	–

Table 13: Per-category results for best ST2 system (RAG-ICL, k=7, filtered, Gemma-3-27b-it, EN). Green : well-detected categories. Orange : difficult categories. Low recall on *Other* (0.32) might suggest label ambiguity and class imbalance.

Table 13 shows a clear performance gap across labels. *Political* is the strongest category (F1 = 0.78), while *Racial/Ethnic* and *Religious* are relatively stable (F1 = 0.68 and 0.69). *Gender/Sexual* decreases mainly because of lower recall (0.50), suggesting the model misses many valid cases. The *Other* label is the hardest by a wide margin (F1 = 0.40, recall = 0.32, 19 FN).

Rank	Method	Model	Pr. Lang.	k	$F1_{mac}$	Prec.	Rec.	Acc.	Spec.
1	RAG-ICL	Gemma-3-27b-it	EN	7	**0.8446**	0.8182	0.8400	0.8462	0.8511
2	RAG-ICL	Gemma-3-27b-it	AR	3	0.8442	0.8267	0.8267	0.8462	0.8617
3	RAG-ICL	Gemma-3-27b-it	AR	5	0.8388	0.8077	0.8400	0.8402	0.8404
4	RAG-ICL	Gemma-3-27b-it	AR	7	0.8211	0.7848	0.8267	0.8225	0.8191
5	RAG-ICL	Gemma-3-27b-it	EN	3	0.8202	0.8000	0.8000	0.8225	0.8404
6	RAG-ICL	Gemma-3-27b-it	EN	5	0.8149	0.7821	0.8133	0.8166	0.8191
7	Zero-Shot	Fanar	AR	–	0.7905	0.7632	0.7733	0.7929	0.8085
8	RAG-ICL	Fanar	AR	3	0.7546	0.7237	0.7333	0.7574	0.7766
9	Zero-Shot	Fanar	EN	–	0.7524	0.7429	0.6933	0.7574	0.8085
10	RAG-ICL	Fanar	AR	5	0.7419	0.7162	0.7067	0.7456	0.7766
11	RAG-ICL	Fanar	EN	3	0.7403	0.7286	0.6800	0.7456	0.7979
12	RAG-ICL	Fanar	EN	5	0.7299	0.7027	0.6933	0.7337	0.7660
13	RAG-ICL	Fanar	EN	7	0.6520	0.6667	0.5067	0.6686	0.7979
14	RAG-ICL	Fanar	AR	7	0.6230	0.6667	0.4267	0.6509	0.8298
15	Zero-Shot	Gemma-3-27b-it	AR	–	0.5327	0.8333	0.2000	0.6272	0.9681
16	Zero-Shot	Gemma-3-27b-it	EN	–	0.5113	0.8125	0.1733	0.6154	0.9681
17	Zero-Shot	Llama-3.3-70B-Instruct	EN	–	0.4815	0.8333	0.1333	0.6036	0.9787
18	Zero-Shot	Qwen3-235B-A22B	EN	–	0.4735	0.9000	0.1200	0.6036	0.9894
19	Zero-Shot	DeepSeek V3.2	EN	–	0.4490	0.8750	0.0933	0.5917	0.9894
20	ZS (Think)	Qwen3-235B-A22B	EN	–	0.4331	0.7500	0.0800	0.5858	0.9787
21	ZS (Think)	DeepSeek V3.2	EN	–	0.4000	1.0000	0.0400	0.5740	1.0000

Table 11: Complete ranking of all LLM-based (generative/RAG-ICL-based) 21 experiments conducted for Subtask 1 on the development set.

Rank	Method	Model	Pr. Lang.	k	Filt.	$F1_{mac}$	$F1_{mic}$	P_{mac}	P_{mic}	R_{mac}	R_{mic}	Acc	Ham.	Hall.
1	RAG-ICL	Gemma-3-27b-it	EN	7	✓	**0.6304**	0.6559	**0.6903**	0.6983	0.5927	0.6183	**0.6272**	**0.1006**	0.1702
2	CoT	Gemma-3-27b-it	EN	7	✓	0.6196	0.6475	0.6790	0.6991	0.5818	0.6031	**0.6272**	0.1018	0.1702
3	RAG-ICL	Gemma-3-27b-it	AR	7	✓	0.6161	0.6429	0.6562	0.6694	0.5895	0.6183	0.6095	0.1065	0.1702
4	RAG-ICL	Gemma-3-27b-it	EN	5	✓	0.6146	0.6454	0.6539	0.6750	0.5909	0.6183	0.6331	0.1053	0.1702
5	Contrastive	Gemma-3-27b-it	EN	7	✓	0.6062	0.6320	0.6521	0.6639	0.5771	0.6031	0.6154	0.1089	0.1702
6	RAG-ICL	Gemma-3-27b-it	EN	3	✓	0.6044	0.6311	0.6620	0.6814	0.5689	0.5878	0.6331	0.1065	0.1702
7	RAG-ICL	Gemma-3-27b-it	EN	5	✗	0.5935	0.5993	0.5447	0.5361	**0.6607**	**0.6794**	0.4497	0.1408	0.5745
8	Self-Refine	Gemma-3-27b-it	EN	7	✓	0.5692	0.6215	0.6129	0.6500	0.5678	0.5954	**0.6272**	0.1124	0.1702
9	Filtered ZS	Gemma-3-27b-it	EN	–	✓	0.5643	0.6192	0.6711	0.6852	0.5366	0.5649	0.6036	0.1077	**0.1702**
10	RAG-ICL	Gemma-3-27b-it	EN	3	✗	0.5590	0.5591	0.5442	0.5270	0.5826	0.5954	0.4320	0.1456	0.5213
11	RAG-ICL	Gemma-3-27b-it	EN	7	✗	0.5457	0.5655	0.5190	0.5157	0.5856	0.6260	0.4083	0.1491	0.5745
12	Zero-Shot	Gemma-3-27b-it	EN	–	✗	0.5425	0.5586	0.5080	0.5094	0.6070	0.6183	0.3964	0.1515	0.6064
13	Zero-Shot	Gemma-3-27b-it	AR	–	✗	0.5315	0.5281	0.4840	0.4651	0.6037	0.6107	0.3846	0.1692	0.8085
14	Zero-Shot	Qwen3-235B-A22B	EN	–	✗	0.5273	0.5487	0.6151	0.6526	0.4677	0.4733	0.5503	0.1207	0.2447
15	Zero-Shot	Llama-3.3-70B-Instruct	EN	–	✗	0.4976	0.5388	0.5482	0.5789	0.4794	0.5038	0.5207	0.1337	0.3511
16	Zero-Shot	Fanar	EN	–	✗	0.4290	0.4656	0.4507	0.4656	0.4204	0.4656	0.3728	0.1657	0.5957
17	Zero-Shot	Fanar	AR	–	✗	0.3817	0.4179	0.4124	0.4088	0.3959	0.4275	0.3018	0.1846	0.7128

Table 12: Complete ranking of all 17 experiments conducted for Subtask 2 on the development set. **Acc** (Exact Match): fraction of samples where all 5 labels are exactly correct (strictest metric). **Ham.** (Hamming Loss): fraction of individual label slots predicted incorrectly (lower is better; e.g., 0.10 means 10% wrong label slots). **Hall.** (Hallucination Rate): among truly non-polarized samples (all 5 labels are 0), fraction where the model predicts at least one category.

Figure 2: Dataset analysis of the Arabic training partition. **(A)** ST1 class distribution. **(B)** ST2 type frequencies among polarized instances. **(C)** ST2 co-occurrence matrix (diagonal = marginal counts). **(D)** Label cardinality distributions for ST2 and ST3 among polarized instances. **(E)** Text length density by class. **(F)** ST3 manifestation frequencies (included for completeness; not covered in this work).

E Detailed Linguistic Analysis

To gain deeper insights into our system's limitations and reasoning capabilities, we conducted a manual qualitative audit of the model's outputs. Two experts, an Arabic linguist and a computational linguist, reviewed a stratified sample of 40 instances (20 for each subtask, from the best system of each of them). The annotators evaluated the linguistic nuances of the Arabic texts, the validity of the LLM-generated rationales (for ST2), and the accuracy of the ground-truth labels. Following independent review, the linguists reconciled their findings to identify recurring model blind spots, linguistic patterns (dialectal correlations, keyword bias, ...), and annotation inconsistencies. Examples are grouped by error pattern, with annotations on dialect, framing, and failure mode.

E.1 ST1: Binary Detection

✓ Correct — Political Polarization (Levantine)

Text: والله اذا بدك تحكي على البيع والشراء ما رح تلاقي اكثر من ال الاسد باعوا سورية لكل دول العالم لكن معلمك بشار الأسد لص وغبي
Gold: Polarized **Model:** Polarized ✓
Analysis: Explicit political attack on Assad (and the Assad family) with insulting language (لص وغبي — "thief and stupid"). The second-person pronoun معلمك ("your teacher/master") creates clear "us vs. them" framing. Levantine dialect markers: بدك, رح.

✓ Correct — Group Hostility (Gulf)

Text: المشكله وهو طبال ضاغطكم وكاتم بلوفكم ومخليكم تنابحون ليل نهار ولا طق لكم خبر، شلون اجل لو مو طبال [2 ROFL emojis]
Gold: Polarized **Model:** Polarized ✓
Analysis: Group-directed hostility via plural pronoun كم. Dehumanizing verb تنابحون ("you bark at each other"). Gulf dialect: شلون, اجل. Sarcasm reinforced by emojis.

✓ Correct — Reporting on Division (MSA)

Text: الرئيس السوري أحمد الشرع يقول إن النظام السابق اعتمد على الانقسام الطائفي خلال 54 سنة وكان يخوّف الطوائف من الأغلبية في سوريا
Gold: Not polarized **Model:** Not polarized ✓
Analysis: Reports a political statement *about* sectarian division without adopting a hostile stance. Model correctly separates meta-discussion of polarization from polarizing content itself.

✗ Error (False Negative) — Missed Non-Political Target (Egyptian)

Text: هو الفن هبط وباظ من قليل عشان مبقاش في موهبه أصلا ولا حتى عمل جيد
Gold: Polarized **Model:** Not polarized ✗
Analysis: Attacks an entire professional group (artists) with sweeping generalizations: "no talent at all." The model fails to recognize group-directed hostility when the target is *non-political*. Egyptian dialect: عشان, مبقاش, باظ.
Failure mode: Political-bias in polarization concept.

✗ Error (False Negative) — Implicit Polarization via Questions (MSA)

Text: كيف نسوية ؟ وما علاقتها ب تحرر المرأة وكيف يمكن ان تعبر عن حقوق المرأة وهي متدينة كيف ؟
Gold: Polarized **Model:** Not polarized ✗
Analysis: Though phrased as questions, the text attacks feminists by implying feminism contradicts religion — delegitimization through rhetorical skepticism. The model struggles with polarization in question form rather than direct attack.
Failure mode: Implicit framing not detected.

✗ Error (False Negative) — Implicit "Us vs. Them" (MSA)

Text: الفجوة الهائلة بين ملايين المسحوقين الفقراء وزمرة من المنتفعين ستخلقُ فوران للشارع قد يصل الى العنف.
Gold: Polarized **Model:** Not polarized ✗
Analysis: Divides society into two opposing groups: ملايين المسحوقين الفقراء ("millions of crushed poor") vs. زمرة من المنتفعين ("a group of profiteers"). Creates sharp "us vs. them" framing through social analysis rather than explicit attack.
Failure mode: Structural polarization missed.

✗ Error (False Positive) — Individual Attack (Gulf)

Text: معزة وصارت مغنية... هي تستحق لقب ملكة عن جدارة..ملكة التفاهة والرقاعة
Gold: Not polarized **Model:** Polarized ✗
Analysis: Insults *one* specific singer (معزة — "goat", ملكة التفاهة — "queen of triviality"). No group generalization. Model cannot distinguish individual attacks from group-directed hostility.
Failure mode: Individual/group confusion.

✗ Error (False Positive) — Sports Humor (Iraqi)

Text: شوف كوريا صارت فريق شعبي كجال البرازيل واحنه من نلعب كجالهم ما انخلي دعاء ما نقراه هههههههه
Gold: Not polarized **Model:** Polarized ✗
Analysis: Sports banter comparing Korean and Brazilian football teams with Iraqi humor. Playful exaggeration ("we read every prayer"), not group-directed hostility. Model misreads passionate tone and dialect markers as polarization.
Failure mode: Tone/register misinterpretation.

✗ Error (False Positive) — Social Commentary (Levantine)

Text: المشكلة انه حالات الاغتصاب بتكون كمان بأماكن النقاب فيها سائد. بس مع هيك بينحط الحق عالضحية
Gold: Not polarized **Model:** Polarized ✗
Analysis: Discusses victim-blaming in sexual assault cases. Sensitive keywords (الاغتصاب, النقاب) trigger the model, but the text is analytical social commentary, not group-directed attack.
Failure mode: Keyword sensitivity without context.

E.2 ST2: Type Classification

✓ Exact Match — Political + Racial + Religious

Text: رئيس الدولة كافر والشعب ساكت خاطرو شعب طحان
Gold: political, racial/ethnic, religious
Model: political, racial/ethnic, religious ✓
LLM Rationale: The text explicitly attacks the president (political), uses the term "kafir" / infidel (religious), and insults the people as "tahhan" (racial/ethnic/social).
Analysis: All three labels correctly captured from explicit cues. The model's rationale aligns perfectly with the pragmatic intent of the text, successfully parsing the intersection of political critique and religious/social slurs.

✗ Keyword Bias — Appearance → Religious Hallucination

Text: الساقط شوف بطنه عبالك بطن الشمر وصخ ... ساقط مختل عقليا المفروض هذول إعدام وهاي البربوك لابسه عبايه وحجاب وهي وحده ساقطه ولكم طفله عمرها ست سنوات يا ساقطين شنو ذنبها
Gold: gender/sexual, other, political
Model: gender/sexual, other, religious ✗ (hallucinated *religious*; missed *political*)
LLM Rationale: The text mentions "abaya" and "hijab" (عبايه وحجاب), indicating that the attack is based on religious appearance, thus constituting religious polarization.
Analysis: The LLM explicitly cites clothing items as evidence for a *religious* label. However, these are visual appearance markers used here in a *moral and gendered* condemnation context (attacking the hypocrisy of the individuals). This is a clear case of **keyword bias**: surface-level vocabulary triggered a religious classification without any contextual understanding.

✗ Clever Hans & Intersectionality — Misread Intent

Text: لا عزة ولا شرف ولا نخوة ولا كرامة 100 راجل مكانش فيهم لي سلكها ولا عاونها بصح على شكون لوكان كان كاين جزائري واحد تما يوقفهم العزة و النيف عند مواليه
Gold: gender/sexual, racial/ethnic
Model: racial/ethnic ✗ (missed *gender/sexual*)
LLM Rationale: "The text expresses a lack of honor and courage, specifically contrasting it with what a جزائري (Algerian) would do. This targets Algerians as a group and implies a negative judgment about their character, thus constituting racial/ethnic polarization."
Analysis: The LLM completely hallucinates the intent. It sees the keyword "Algerian" alongside negative words ("no honor") and assumes it insults Algerians, when the text actually *praises* them while attacking the masculinity and chivalry (نخوة) of the "100 men" who failed to protect a woman. The model misses the intersectional gender dimension entirely and gets the ethnic sentiment backwards despite predicting the correct label.

✗ Cultural Gap — Sectarian Context Missed

Text: داعش اليوم ليس هي امتداد لداعش امس لقد ماتت داعش بعد تحرير نينوى ودير الزور وما تبقى هم رداء داعش متهري ينتهون بانتهاء الحشد
Gold: political, racial/ethnic
Model: political, religious ✗ (hallucinated *religious*; missed *racial/ethnic*)
LLM Rationale: The text discusses ISIS, an Islamic extremist group, which inherently constitutes religious and political polarization.
Analysis: The model interprets ISIS as a religious group (surface-level association), missing the ethnic/sectarian dimension of the regional conflict (Iraq, Nineveh, Deir ez-Zor). Crucially, it fails to recognize the cultural and sectarian weight of الحشد (Popular Mobilization Forces), treating the conflict as purely religious/political and missing the localized ethnic/sectarian nuances.

F Complete Prompt Templates

All prompts are shown verbatim as used in experiments. Placeholders are marked with `{braces}`.

F.1 Data Augmentation (ST1)

Used with DeepSeek V3.2 to generate augmented training data for encoder fine-tuning. The pipeline has two phases.

Phase 1: Style Analysis. A random sample of 20 training texts is analyzed to produce a reusable style guide.

System — Style Analysis

You are a computational linguist.

User — Style Analysis

Analyze the following Arabic Social Media samples.
DATASET SAMPLES:
`{examples_text}`
OUTPUT FORMAT:
Provide a concise "Style Guide" paragraph describing:

1. Dialects present.
2. Swearing/Insult patterns in Polarized text.
3. Tone differences.

This guide is for a generative model to mimic this exact distribution.

Phase 2: Augmentation. For each polarized training sample, two variants are generated.

System — Augmentation

You are an expert Arabic Data Augmentor.
STYLE GUIDE: `{style_guide}`

User — Augmentation

TASK: Generate 2 variations of the Target Text.

1. PARAPHRASE (Label 1): Rewrite in a different dialect/style from the dataset. MUST remain POLARIZED (hateful/divisive).
2. HARD NEGATIVE (Label 0): Rewrite to contain the same criticism but REMOVE hate/insults/generalization. Must be CRITICAL but NOT POLARIZED.

REFERENCE EXAMPLES (Mimic Style):
`{few_shot_text}`
TARGET TEXT:
"`{text}`"
RESPONSE FORMAT (JSON ONLY):

```
{``paraphrase'': ``Arabic text'', ``paraphrase_dialect'': ``dialect'', ``hard_negative'':
``Arabic text''}
```

Generation uses temperature = 0.7 and max_tokens = 2048.

F.2 Subtask 1 — Binary Polarization Detection

Zero-Shot (EN)

System — Zero-Shot (EN)

```
You are an expert in Arabic sociolinguistic analysis, specializing
in detecting social polarization.

=== TASK ===
Determine whether an Arabic text contains POLARIZING content
(binary: yes or no).

=== DEFINITION OF POLARIZATION ===
Polarization is content that creates "us vs. them" social division by:
1. TARGETING a specific SOCIAL GROUP (not just an individual person)
2. EXPRESSING negative sentiment through stereotyping, vilification,
   or dehumanization

=== WHAT IS NOT POLARIZATION ===
- Personal insults against individuals (without group generalization)
- Criticism or disagreement without hate/vilification
- Neutral discussion of controversial topics
- Positive or encouraging content

=== RESPONSE FORMAT ===
Output ONLY valid JSON:
{"polarized": 0 or 1, "reasoning": "brief explanation"}

0 = Text does NOT contain polarization
1 = Text DOES contain polarization
```

User — Zero-Shot (EN)

```
Classify this Arabic text for polarization.

TEXT:
"""{text}"""

Does this text target a specific SOCIAL GROUP with negative sentiment
(stereotyping, vilification, or dehumanization)?

If YES -> polarized: 1
If NO -> polarized: 0

Output ONLY JSON: {"polarized": 0 or 1, "reasoning": "brief explanation"}
```

Zero-Shot (AR)

System — Zero-Shot (AR)

أنت خبير في تحليل الخطاب العربي للكشف عن الاستقطاب الاجتماعي.

=== المهمة ===

حدد ما إذا كان النص العربي يحتوي على محتوى استقطابي (ثنائي: نعم أو لا).

=== تعريف الاستقطاب ===
الاستقطاب هو محتوى يخلق انقسامًا اجتماعيًا ("نحن ضد هم") من خلال:
١. استهداف مجموعة اجتماعية محددة (وليس فردًا فقط)
٢. التعبير عن مشاعر سلبية من خلال التنميط أو التشويه أو التجريد من الإنسانية

=== ما ليس استقطابًا ===
- الإهانات الشخصية للأفراد (بدون تعميم على المجموعة)
- النقد أو الاختلاف بدون كراهية/تشويه
- المناقشة المحايدة للمواضيع المثيرة للجدل
- المحتوى الإيجابي أو المشجع

=== تنسيق الإخراج ===
أجب بـ JSON فقط:
{"polarized": 0 أو 1, "reasoning": "تفسير مختصر"}

0 = النص لا يحتوي على استقطاب
1 = النص يحتوي على استقطاب

User — Zero-Shot (AR)

صنّف هذا النص العربي من حيث الاستقطاب.

النص:
"""{text}"""

هل يستهدف هذا النص مجموعة اجتماعية محددة بمشاعر سلبية (تنميط، تشويه، أو تجريد من الإنسانية)؟

إذا نعم -> polarized: 1
إذا لا -> polarized: 0

أجب بـ JSON فقط: {"polarized": 0 أو 1, "reasoning": "تفسير مختصر"}

RAG-ICL (EN)

System Prompt: same as Zero-Shot (EN) above.

User — RAG-ICL (EN)

```
Below are similar examples from the training data with their ground
truth labels.

=== SIMILAR TRAINING EXAMPLES ===

--- Example {i}: {POLARIZED/NOT POLARIZED} (similarity: {score}) ---
"{example_text}"
```

```
[repeated for k examples]

=== TEXT TO CLASSIFY ===
"""{text}"""

Does this text contain social polarization? Compare with the similar
examples above.
Remember: Polarization requires targeting a GROUP (not individual)
with negative sentiment.

Output ONLY JSON:
{"polarized": 0/1, "reasoning": "Brief explanation"}
```

RAG-ICL (AR)

System Prompt: same as Zero-Shot (AR) above.

User — RAG-ICL (AR)

فيما يلي أمثلة مشابهة من بيانات التدريب مع تصنيفاتها.

=== أمثلة تدريب مشابهة ===

--- مثال {i}: {مستقطب/غير مستقطب} (تشابه: {score}) ---
"{example_text}"

[يتكرر لـ k أمثلة]

=== النص المطلوب تصنيفه ===
"""{text}"""

هل يحتوي هذا النص على استقطاب اجتماعي؟ قارن مع الأمثلة أعلاه.
تذكر: الاستقطاب يتطلب استهداف مجموعة (وليس فردًا) بمشاعر سلبية.

أجب بـ JSON فقط:
{"polarized": 0/1, "reasoning": "تفسير مختصر"}

F.3 Subtask 2 — Multi-Label Polarization Type Classification

ST2 system prompts are composed of two blocks that are concatenated: (1) a *Core Definition Block* shared by all ST2 methods, and (2) a *Task Block* that varies depending on whether pipeline filtering from ST1 is applied. Both blocks are shown first; the per-method user prompts follow.

System Prompt — Core Definition Block (EN)

System Prompt — Core Definition Block (EN)

```
You are an expert in Arabic sociolinguistic analysis, specializing
in detecting social polarization.

=== DEFINITION OF POLARIZATION ===
Polarization is content that creates "us vs. them" social division by:
1. TARGETING a specific SOCIAL GROUP (not just an individual person)
```

```
2. EXPRESSING negative sentiment through stereotyping, vilification,
   or dehumanization

=== WHAT IS NOT POLARIZATION ===
- Personal insults against individuals (without group generalization)
- Criticism or disagreement without hate/vilification
- Neutral discussion of controversial topics
- Factual reporting of events or conflicts

=== CRITICAL DISTINCTIONS ===
1. INDIVIDUAL vs GROUP: Insulting one person =/= polarization.
   Insulting their GROUP = polarization.
2. FACTUAL vs HOSTILE: Reporting conflict =/= polarization.
   Adding hostile generalizations = polarization.
3. OPINION vs ATTACK: Disagreeing with ideas =/= polarization.
   Attacking people for their identity = polarization.
```

System Prompt — Core Definition Block (AR)

System Prompt — Core Definition Block (AR)

أنت خبير في التحليل اللغوي الاجتماعي العربي، متخصص في اكتشاف الاستقطاب الاجتماعي.

=== تعريف الاستقطاب ===
الاستقطاب هو محتوى يخلق انقساماً اجتماعياً بمنطق "نحن ضد هم" من خلال:
١. استهداف مجموعة اجتماعية محددة (وليس مجرد فرد)
٢. التعبير عن مشاعر سلبية عبر التنميط أو التشويه أو نزع الإنسانية

=== ما ليس استقطاباً ===
- الإهانات الشخصية للأفراد (بدون تعميم على المجموعة)
- النقد أو الاختلاف بدون كراهية أو تشويه
- المناقشة المحايدة للمواضيع المثيرة للجدل
- التقارير الإخبارية الموضوعية

=== تمييزات جوهرية ===
١. فرد مقابل مجموعة: إهانة شخص واحد =/= استقطاب. إهانة مجموعته = استقطاب.
٢. حقائق مقابل عداء: نقل الأحداث =/= استقطاب. إضافة تعميمات عدائية = استقطاب.
٣. رأي مقابل هجوم: الاختلاف مع الأفكار =/= استقطاب. مهاجمة الناس بسبب هويتهم = استقطاب.

System Prompt — Task Block, Unfiltered (EN)

Appended to the Core Definition Block (EN) when no ST1 filter is applied.

System Prompt — Task Block, Unfiltered (EN)

```
=== TASK ===
Classify Arabic text into polarization types. A text can have
ZERO, ONE, or MULTIPLE types.
```

```
=== CATEGORIES ===
1. POLITICAL -- Targets political groups, governments, ideologies
2. RACIAL/ETHNIC -- Targets nationalities, ethnic groups, races,
   regional origins
3. RELIGIOUS -- Targets religious groups, sects, beliefs
4. GENDER/SEXUAL -- Targets based on gender, sexual orientation
5. OTHER -- Targets professional groups, social classes,
   or ambiguous collective targets

=== RESPONSE FORMAT ===
Output ONLY valid JSON:
{"political": 0/1, "racial/ethnic": 0/1, "religious": 0/1,
  "gender/sexual": 0/1, "other": 0/1,
  "reasoning": "brief explanation"}
```

System Prompt — Task Block, Unfiltered (AR)

Appended to the Core Definition Block (AR) when no ST1 filter is applied.

System Prompt — Task Block, Unfiltered (AR)

```
=== المهمة ===
صنّف النص العربي حسب أنواع الاستقطاب. يمكن أن يحتوي النص على صفر أو نوع واحد أو أنواع متعددة.

=== الفئات ===
١. سياسي -- يستهدف مجموعات سياسية أو حكومات أو أيديولوجيات
٢. عرقي/إثني -- يستهدف جنسيات أو مجموعات عرقية أو أصول إقليمية
٣. ديني -- يستهدف مجموعات دينية أو طوائف أو معتقدات
٤. جنسي/نوعي -- يستهدف على أساس الجنس أو التوجه الجنسي
٥. أخرى -- يستهدف مجموعات مهنية أو طبقات اجتماعية أو أهداف جماعية غامضة

=== صيغة الإجابة ===
أجب فقط بـ JSON صالح:
{"political": 0/1, "racial/ethnic": 0/1, "religious": 0/1,
  "gender/sexual": 0/1, "other": 0/1,
                                    "reasoning": "شرح مختصر"}
```

System Prompt — Task Block, Filtered (EN)

Appended to the Core Definition Block (EN) when ST1 pipeline filtering is active.

System Prompt — Task Block, Filtered (EN)

```
=== TASK ===
The input text has ALREADY been identified as containing polarized
content. Your task is to classify WHICH TYPE(S) of polarization it
contains. A text can have ONE or MULTIPLE types simultaneously.

=== CATEGORIES ===
```

```
1. POLITICAL -- Targets political groups, governments, ideologies
2. RACIAL/ETHNIC -- Targets nationalities, ethnic groups, races,
   regional origins
3. RELIGIOUS -- Targets religious groups, sects, beliefs
4. GENDER/SEXUAL -- Targets based on gender, sexual orientation
5. OTHER -- Targets professional groups, social classes,
   or ambiguous collective targets

=== RESPONSE FORMAT ===
Output ONLY valid JSON:
{"political": 0/1, "racial/ethnic": 0/1, "religious": 0/1,
  "gender/sexual": 0/1, "other": 0/1,
  "reasoning": "brief explanation"}

IMPORTANT: At least ONE category should be 1.
```

System Prompt — Task Block, Filtered (AR)

Appended to the Core Definition Block (AR) when ST1 pipeline filtering is active.

System Prompt — Task Block, Filtered (AR)

```
=== المهمة ===
تم تحديد النص التالي مسبقاً على أنه يحتوي على محتوى استقطابي.
مهمتك هي تصنيف أي نوع (أو أنواع) من الاستقطاب يحتوي عليها.
يمكن أن يحتوي النص على نوع واحد أو أنواع متعددة في آنٍ واحد.

=== الفئات ===
١. سياسي -- يستهدف مجموعات سياسية أو حكومات أو أيديولوجيات
٢. عرقي/إثني -- يستهدف جنسيات أو مجموعات عرقية أو أصول إقليمية
٣. ديني -- يستهدف مجموعات دينية أو طوائف أو معتقدات
٤. جنسي/نوعي -- يستهدف على أساس الجنس أو التوجه الجنسي
٥. أخرى -- يستهدف مجموعات مهنية أو طبقات اجتماعية أو أهداف جماعية غامضة

=== صيغة الإجابة ===
أجب فقط بـ JSON صالح:
{"political": 0/1, "racial/ethnic": 0/1, "religious": 0/1,
  "gender/sexual": 0/1, "other": 0/1,
  "reasoning": "شرح مختصر"}

مهم: يجب أن تكون فئة واحدة على الأقل = ١.
```

Zero-Shot (EN)

System Prompt: Core Definition Block (EN) + Task Block, Unfiltered (EN).

```
Classify the following Arabic text for polarization types.

TEXT:
"""{text}"""

ANALYSIS CHECKLIST:
1. Does this text target a specific SOCIAL GROUP (not just an individual)?
2. Does it express negative sentiment through stereotyping, vilification,
   or dehumanization?
3. Which categories apply? (Can be zero, one, or multiple)

Output ONLY JSON:
{"political": 0/1, "racial/ethnic": 0/1, "religious": 0/1,
  "gender/sexual": 0/1, "other": 0/1,
  "reasoning": "brief explanation"}
```

Zero-Shot (AR)

System Prompt: Core Definition Block (AR) + Task Block, Unfiltered (AR).

صنّف النص العربي التالي حسب أنواع الاستقطاب.

النص:
"""{text}"""

قائمة التحقق:
١. هل يستهدف هذا النص مجموعة اجتماعية محددة (وليس فرداً فقط)؟
٢. هل يعبّر عن مشاعر سلبية عبر التنميط أو التشويه أو نزع الإنسانية؟
٣. أي الفئات تنطبق؟ (يمكن أن تكون صفر أو واحدة أو أكثر)

أجب فقط بـ JSON:

```
{"political": 0/1, "racial/ethnic": 0/1, "religious": 0/1,
  "gender/sexual": 0/1, "other": 0/1,
  "reasoning": "شرح مختصر"}
```

RAG-ICL (EN)

System Prompt: Core Definition Block (EN) + Task Block, Filtered (EN).

```
Below are similar examples from the training data with their ground
truth labels.

=== SIMILAR TRAINING EXAMPLES ===

--- Example {i} (similarity: {score}) ---
Text: "{example_text}"
Categories: {active_categories}
Labels: political={0/1}, racial/ethnic={0/1}, religious={0/1},
        gender/sexual={0/1}, other={0/1}

[repeated for k examples]

=== TEXT TO CLASSIFY ===
"""{text}"""

Based on the similar examples above, classify this text for ALL five
polarization types.
```

```
ANALYSIS CHECKLIST:
1. Does this text target a specific SOCIAL GROUP?
2. Does it express negative sentiment through stereotyping, vilification,
   or dehumanization?
3. Which categories apply? Compare with the examples above.

Output ONLY JSON:
{"political": 0/1, "racial/ethnic": 0/1, "religious": 0/1,
  "gender/sexual": 0/1, "other": 0/1,
  "reasoning": "brief explanation"}
```

RAG-ICL (AR)

System Prompt: Core Definition Block (AR) + Task Block, Filtered (AR).

```
فيما يلي أمثلة مشابهة من بيانات التدريب مع تصنيفاتها الصحيحة.

=== أمثلة تدريبية مشابهة ===

--- مثال {i} (تشابه: {score}) ---
النص: "{example_text}"
الفئات: {active_categories}
التصنيفات: religious={0/1}, racial/ethnic={0/1}, political={0/1},
other={0/1} gender/sexual={0/1},

[يتكرر لـ k أمثلة]

=== النص المطلوب تصنيفه ===
"""{text}"""

بناءً على الأمثلة المشابهة أعلاه، صنّف هذا النص لجميع أنواع الاستقطاب الخمسة.

قائمة التحقق:
١. هل يستهدف هذا النص مجموعة اجتماعية محددة؟
٢. هل يعبّر عن مشاعر سلبية عبر التنميط أو التشويه أو نزع الإنسانية؟
٣. أي الفئات تنطبق؟ قارن مع الأمثلة أعلاه.

أجب فقط بـ JSON:
{"political": 0/1, "racial/ethnic": 0/1, "religious": 0/1,
  "gender/sexual": 0/1, "other": 0/1,
  "reasoning": "شرح مختصر"}
```

Contrastive (EN)

System Prompt: Core Definition Block (EN) + Task Block, Filtered (EN).

```
Below are POSITIVE examples (polarized with specific categories) and
NEGATIVE examples (not matching those categories).

=== POSITIVE EXAMPLES (polarized) ===
```

```
1. "{text}" -> {categories}
[...]

=== NEGATIVE EXAMPLES (not polarized or different type) ===
1. "{text}" -> not polarized / different categories
[...]

=== TEXT TO CLASSIFY ===
"""{text}"""

Compare with both positive and negative examples to make your decision.

Output ONLY JSON:
{"political": 0/1, "racial/ethnic": 0/1, "religious": 0/1,
  "gender/sexual": 0/1, "other": 0/1,
  "reasoning": "brief explanation"}
```

Contrastive (AR)

System Prompt: Core Definition Block (AR) + Task Block, Filtered (AR).

```
فيما يلي أمثلة إيجابية (مستقطبة مع فئات محددة) وأمثلة سلبية (لا تطابق تلك الفئات).

=== أمثلة إيجابية (مستقطبة) ===
{categories} -> "{text}" 1.
[...]

=== أمثلة سلبية (غير مستقطبة أو نوع مختلف) ===
غير مستقطب / فئات أخرى -> "{text}" 1.
[...]

=== النص المطلوب تصنيفه ===
"""{text}"""

قارن مع الأمثلة الإيجابية والسلبية لاتخاذ قرارك.

أجب فقط بـ JSON:
{"political": 0/1, "racial/ethnic": 0/1, "religious": 0/1,
  "gender/sexual": 0/1, "other": 0/1,
"reasoning": "شرح مختصر"}
```

Chain-of-Thought (EN)

System Prompt: Core Definition Block (EN) + Task Block, Filtered (EN). User Prompt: identical to RAG-ICL (EN) above, with the `ANALYSIS CHECKLIST` block replaced by the following.

```
Below are similar examples from the training data with their ground
truth labels.

=== SIMILAR TRAINING EXAMPLES ===

--- Example {i} (similarity: {score}) ---
Text: "{example_text}"
Categories: {active_categories}
```

```
Labels: political={0/1}, racial/ethnic={0/1}, religious={0/1},
        gender/sexual={0/1}, other={0/1}

[repeated for k examples]

=== TEXT TO CLASSIFY ===
"""{text}"""

Based on the similar examples above, classify this text for ALL five
polarization types.

Think step by step:
1. What social group (if any) is being targeted in this text?
2. What negative sentiment (if any) is expressed?
   (stereotyping, vilification, dehumanization)
3. How does this compare to the similar examples above?
4. Which specific categories apply?

Output ONLY JSON:
{"political": 0/1, "racial/ethnic": 0/1, "religious": 0/1,
  "gender/sexual": 0/1, "other": 0/1,
  "reasoning": "brief explanation"}
```

Chain-of-Thought (AR)

System Prompt: Core Definition Block (AR) + Task Block, Filtered (AR). User Prompt: identical to RAG-ICL (AR) above, with قائمة التحقق *replaced by the following.*

```
فيما يلي أمثلة مشابهة من بيانات التدريب مع تصنيفاتها الصحيحة.

=== أمثلة تدريبية مشابهة ===

--- مثال {i} (تشابه: {score}) ---
النص: "{example_text}"
الفئات: {active_categories}
التصنيفات: political={0/1}, racial/ethnic={0/1},
religious={0/1}, gender/sexual={0/1},
other={0/1}

[يتكرر لـ k أمثلة]

=== النص المطلوب تصنيفه ===
"""{text}"""

بناءً على الأمثلة المشابهة أعلاه، صنّف هذا النص لجميع أنواع الاستقطاب الخمسة.

فكّر خطوة بخطوة:
١. ما المجموعة الاجتماعية (إن وُجدت) المستهدفة في هذا النص؟
٢. ما المشاعر السلبية (إن وُجدت) المعبّر عنها؟ (تنميط، تشويه، نزع إنسانية)
٣. كيف يقارن هذا بالأمثلة المشابهة أعلاه؟
٤. أي الفئات المحددة تنطبق؟

أجب فقط بـ JSON:
```

```
{"political": 0/1, "racial/ethnic": 0/1, "religious": 0/1,
  "gender/sexual": 0/1, "other": 0/1,
                                        "reasoning": "شرح مختصر"}
```

Self-Refinement (EN)

System Prompt: Core Definition Block (EN) + Task Block, Filtered (EN). Pass 1 uses the RAG-ICL (EN) user prompt above. Pass 2 uses the following user prompt.

User Prompt — Pass 2

```
You previously classified this text as:
{previous_json_output}

TEXT:
"""{text}"""

Review your classification. Consider:
1. Did you correctly identify the targeted group?
2. Could any categories be missing or incorrectly assigned?
3. Is your reasoning consistent with the definitions?

If your classification is correct, output the same JSON.
If it needs correction, output the corrected JSON.

Output ONLY JSON:
{"political": 0/1, "racial/ethnic": 0/1, "religious": 0/1,
  "gender/sexual": 0/1, "other": 0/1,
  "reasoning": "brief explanation"}
```

Self-Refinement (AR)

System Prompt: Core Definition Block (AR) + Task Block, Filtered (AR). Pass 1 uses the RAG-ICL (AR) user prompt above. Pass 2 uses the following user prompt.

User Prompt — Pass 2

صنّفت هذا النص سابقاً كالتالي:
{previous_json_output}

النص:
"""{text}"""

راجع تصنيفك. فكّر في:
١. هل حددت المجموعة المستهدفة بشكل صحيح؟
٢. هل هناك فئات ناقصة أو مُعيّنة بشكل خاطئ؟
٣. هل تبريرك متسق مع التعريفات؟

إذا كان تصنيفك صحيحاً، أعد نفس الـ JSON.
إذا كان يحتاج تصحيحاً، أخرج الـ JSON المصحح.

أجب فقط بـ JSON:

```
{"political": 0/1, "racial/ethnic": 0/1, "religious": 0/1,
  "gender/sexual": 0/1, "other": 0/1,
```

```
{"reasoning": "شرح مختصر"}
```