PSK at SemEval-2026 Task 9: Multilingual Polarization Detection Using Ensemble Gemma Models with Synthetic Data Augmentation

Srikar Kashyap Pulipaka
Independent Researcher
{srikar.kashyap@gmail.com}

Abstract

We present our system for SemEval-2026 Task 9: Multilingual Polarization Detection, a binary classification task spanning 22 languages. Our approach fine-tunes separate Gemma 3 models (12B and 27B parameters) per language using Low-Rank Adaptation (LoRA), augmented with synthetic data generated by a large language model (LLM). We employ three synthetic data strategies (direct generation, paraphrasing, and contrastive pair creation) using GPT-4o-mini, with a multi-stage quality filtering pipeline including embedding-based deduplication. We find that per-language threshold tuning on the development set yields 2 to 4% F1 improvements without retraining. We also use weighted ensembles of 12B and 27B model predictions with per-language strategy selection. Our final system achieves a mean macro-F1 of 0.811 across all 22 languages, ranking 2nd overall of the participating teams, with 1st place finishes in 3 languages and top-3 in 8 languages. We also find that alternative architectures (XLM-RoBERTa, Qwen3) that showed strong development set performance suffered 30 to 50% F1 drops on the test set, highlighting the importance of generalization.

1 Introduction

Team PSK participated in SemEval-2026 Task 9 (Naseem et al., 2026a) on detecting attitude polarization in social media text across 22 typologically diverse languages. The task requires binary classification of social media text as polarized or non-polarized, where polarization encompasses stereotyping, vilification, dehumanization, and intolerance.

The task presents several challenges: the dataset sizes vary across languages, polarization manifests differently across cultures, and some languages exhibit class imbalance. In our work, we explore per-language fine-tuned Gemma models with synthetic data augmentation, ensemble methods, and post-hoc threshold tuning. We find that robust generalization, rather than maximizing development set performance, is the critical factor for success.

The remainder of this paper is structured as follows: Section 2 gives a brief overview of related work, Section 3 describes the data, Section 4 describes our methodology, Section 5 discusses our results, and we conclude in Section 6.

2 Related Work

Recent work has explored LLM-generated data for text classification augmentation. Cegin et al. (2025) found that LLM augmentation primarily benefits low-data scenarios. Yong et al. (2024) demonstrated 5.6 to 8.9 point improvements for extremely low-resource languages using lexicon-conditioned generation. Our augmentation pipeline draws on backtranslation (Sennrich et al., 2016). Edunov et al. (2018) demonstrated that noisy back-translations provide stronger training signals than clean outputs. Our contrastive pair generation is inspired by counterfactual data augmentation for hate speech detection (Mostafazadeh Davani et al., 2021), where minimal-edit counterfactuals improve classifier robustness.

We use LoRA (Hu et al., 2022) for parameter-efficient fine-tuning, and our per-language threshold tuning follows work by Lipton et al. (2014) on F1 maximization and Pillai et al. (2013) on per-class threshold optimization for macro-averaged metrics.

3 Data

3.1 Task Data

SemEval-2026 Task 9 Subtask 1 uses the POLAR dataset (Naseem et al., 2026b), a binary classification task: given a social media text in one of 22 languages, the task is to classify it as polarized (1) or non-polarized (0). The primary evaluation

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1539–1546

metric is macro-averaged F1-score. The 22 languages are: Amharic (amh), Arabic (arb), Bengali (ben), Burmese (mya), Chinese (zho), English (eng), German (deu), Hausa (hau), Hindi (hin), Italian (ita), Khmer (khm), Nepali (nep), Odia (ori), Persian (fas), Polish (pol), Punjabi (pan), Russian (rus), Spanish (spa), Swahili (swa), Telugu (tel), Turkish (tur), and Urdu (urd). Training set sizes range from approximately 1,700 (Punjabi) to 7,000 (Swahili) samples. Several languages exhibit class imbalance, most notably Khmer (10:1 polarized), Hausa (8:1 non-polarized), Hindi (6:1 polarized), and Amharic (3:1 polarized).

We split the original data into 80% train and 20% validation before adding any synthetic data, ensuring the validation set contains only real data.

3.2 Data Augmentation

To provide more data to our models, we generated synthetic training data using GPT-4o-mini with three complementary strategies:

Direct Generation (50%). We generated new samples natively in the target language, covering five culturally relevant topic categories: political, ethnic/racial, religious, social class, and international relations.

Label-Preserving Paraphrasing (30%). We created paraphrases of real training samples with temperature 0.7, filtered to ensure cosine similarity below 0.90 with the original.

Contrastive Pairs (20%). We generated minimal pairs on the same topic, one polarized and one non-polarized, to sharpen class boundaries.

In preliminary experiments for lower-resource languages, we also explored backtranslation through pivot languages and cross-lingual transfer from related languages (e.g., Hindi to Nepali, Bengali to Odia).

Quality Filtering. We applied a multi-stage pipeline to ensure data quality: (1) basic cleaning and length filtering, (2) label leakage detection via regex patterns, (3) embedding-based deduplication using `paraphrase-multilingual-MiniLM-L12-v2` (Reimers and Gurevych, 2019) with a 0.90 cosine similarity threshold (both intra-synthetic and against original training data), and (4) round-trip translation consistency checks (threshold 0.70) for translated samples.

Synthetic samples were added only to the training portion at a configurable ratio. We generated approximately 1,000 synthetic samples per language.

4 Methodology

Figure 1 provides an overview of our end-to-end system pipeline.

4.1 Model Architecture

We chose Google's Gemma 3 family (Team et al., 2024) as our primary model because of its broad multilingual pretraining mix, which provides meaningful coverage across all 22 POLAR languages rather than concentrating on a narrow set of high resource ones. Our model comparison in Section 5 confirms this choice: Gemma transferred from development to test substantially better than XLM-RoBERTa or Qwen3, both of which have more skewed pretraining distributions. We fine-tuned the Gemma 3 family using LoRA with a sequence classification head. We used two model sizes:

- **Gemma 3 12B**: Our primary model. We used 8-bit quantization on a single A100 GPU, with LoRA rank 16, a learning rate of 5×10^{-5}, and trained for 3 epochs.
- **Gemma 3 27B**: Our ensemble model. We used 4-bit NF4 quantization (Dettmers et al., 2023) with gradient checkpointing across 2 A100 GPUs via `device_map="auto"`.

Both models use 2 output labels, eager attention, and `bfloat16` precision. We set the pad token to the end-of-sequence (EOS) token. All models were trained with batch size 1 to 2, gradient accumulation of 2 to 4 steps (effective batch size 4), warmup ratio 0.1, max sequence length 128 to 256 tokens, and random seed 42. We used the AdamW optimizer with cosine learning rate decay.

4.2 Ensemble and Threshold Tuning

Threshold Tuning. For each language, we searched over thresholds $t \in \{0.3, 0.35, 0.4, \ldots, 0.7\}$ to maximize development set macro-F1, rather than using the default $t = 0.5$.

Ensemble Strategies. For each language, we evaluated four strategies on the development set:

1. 12B predictions with tuned threshold
2. 27B predictions with tuned threshold

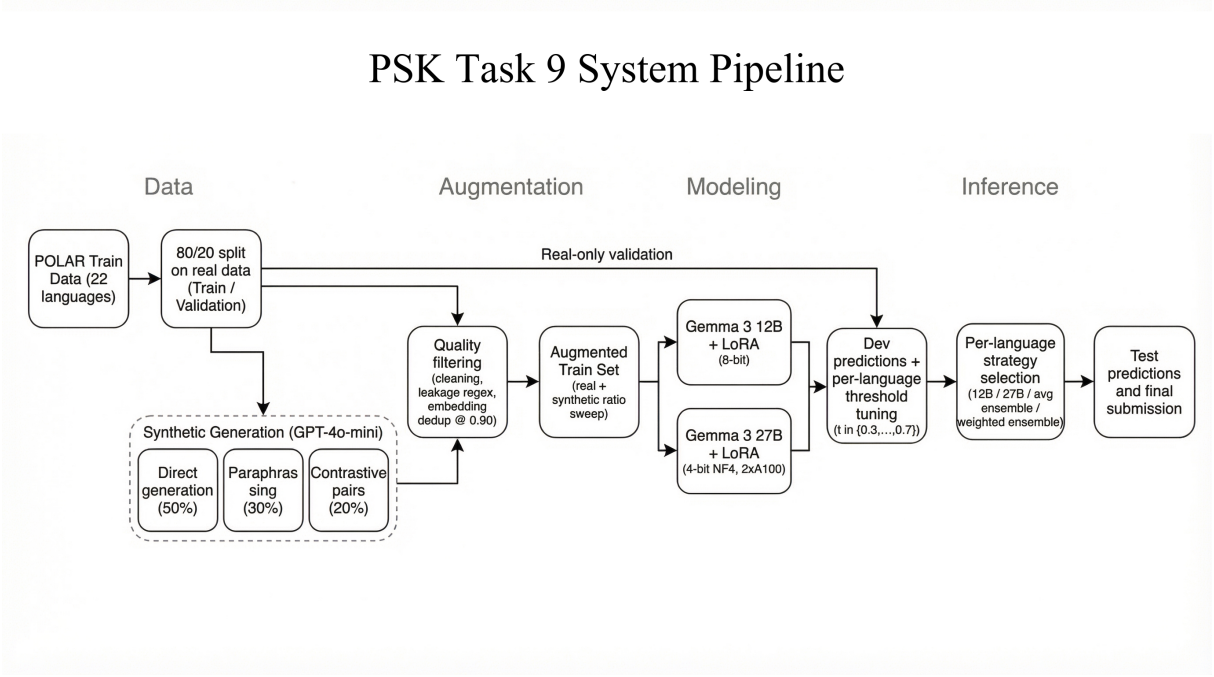

Figure 1: Overview of our multilingual polarization detection pipeline, from real-data splitting and synthetic augmentation to Gemma model training, threshold tuning, and per-language strategy selection.

3. Average ensemble: $p = \frac{p_{12B}+p_{27B}}{2}$

4. Weighted ensemble: $p = w \cdot p_{12B} + (1-w) \cdot p_{27B}$, with $w \in \{0.3, 0.4, 0.6, 0.7\}$

We selected the strategy achieving the highest development F1 per language.

4.3 Infrastructure

Experiments were run on a computing cluster with NVIDIA A100 40GB GPUs. The 12B model requires 1 GPU while the 27B model requires 2 GPUs.

5 Results

We first provide an overview of our official results, then discuss the results of our internal experiments on the development data.

5.1 Official Results

We were allowed to submit 5 systems. Table 1 summarizes our test submissions.

Our best official submission achieved a mean macro-F1 of 0.811 across all 22 languages. A post-hoc combined run that selected the best per-language result between Submission 1 (12B only) and Submission 4 (ensemble with threshold tuning) reached 0.812. Submission 2, which used XLM-RoBERTa and Qwen3 for selected languages, performed significantly worse with a mean F1 of 0.665.

Sub	Strategy	Mean F1
1	Best Gemma-12B per lang	0.797
2	Hybrid (XLM-R, Qwen3)	0.665
4	Ensemble + threshold	0.811
5	**Best of Sub 1 + Sub 4**	**0.812**

Table 1: Test submission results. Submission 2 used alternative architectures that failed to generalize. The final submission selects the best per-language results across submissions.

Synth. Ratio	Dev F1
0% (baseline)	0.812
10%	0.818
20%	0.820
30%	**0.822**
50%	0.819

Table 2: Mean development F1 for Gemma 3 12B by synthetic data ratio. A ratio of 30% is optimal.

5.2 Synthetic Data Ablation

Table 2 shows the impact of synthetic data ratio on mean development F1 across all languages.

We find that lower-resource languages benefit most from synthetic augmentation (+2.2%), while higher-resource languages see smaller gains (+1.9%). Some languages are still sensitive to syn-

Model	Dev F1	Test F1	Δ
Gemma 3 12B	0.827	0.797	−0.030
Gemma 3 27B	0.838	–	–
XLM-R Large	0.800	0.665*	−0.135
Qwen3-14B	0.811†	0.665*	−0.146

Table 3: Model comparison. *Hybrid submission. †Best dev F1 for selected languages only. Only Gemma models generalize reliably from dev to test.

Strategy	Languages Won
Ensemble weighted	9
Ensemble average	5
12B tuned only	4
27B tuned only	4

Table 4: Best strategy per language on the development set. Ensemble methods win for 14 out of 22 languages.

thetic data choices: Swahili improves at low ratios but does not benefit from higher ratios, and Khmer degrades by 6.4% at 50% synthetic ratio. This is consistent with the finding by Cegin et al. (2025) that augmentation primarily benefits lower-data scenarios.

5.3 Model Comparison

We evaluated multiple architectures. Table 3 shows the results.

Only the Gemma models generalize reliably from development to test. XLM-RoBERTa and Qwen3 showed competitive development performance but suffered large drops on the test set. For example, XLM-R achieved 0.856 development F1 for Bengali but only 0.297 on test, a 56% absolute drop. This trend is similar to what we observed in our previous work (Mhalgi et al., 2024), where ensembles that performed best on the development set did not fully generalize to the test set.

5.4 Ensemble and Threshold Results

Table 4 shows the distribution of winning strategies across languages.

Ensemble methods won for 14 out of 22 languages. Optimal thresholds range from 0.3 (Nepali) to 0.7 (Khmer, Amharic, Bengali), showing that the models are not well calibrated at the default threshold of 0.5.

Table 5 shows per-language test F1 for our key submissions. Overall, 18 out of 22 languages improved from Submission 1 to Submission 4, with the largest improvement for Khmer (+8.7%).

Lang	Sub 1	Sub 4	Δ
khm	0.656	0.743	+8.7%
swa	0.774	0.811	+3.7%
nep	0.876	0.908	+3.2%
fas	0.804	0.828	+2.4%
hin	0.800	0.824	+2.4%
pol	0.814	0.835	+2.1%
ita	0.543	0.563	+2.0%
ori	0.793	0.811	+1.8%
eng	0.805	0.818	+1.3%
ben	0.828	0.837	+0.9%
Mean (all 22)	**0.811**		**+1.4%**

Table 5: Per-language test F1 improvements (top 10 shown). 18 out of 22 languages improved from Submission 1 to Submission 4.

5.5 Analysis

Threshold Sensitivity. We find that the models exhibit probability miscalibration. For example, Russian predictions have a mean probability of 0.246 (under-confident), while Khmer averages 0.919 (over-confident). This miscalibration makes the default 0.5 threshold suboptimal, which explains why threshold tuning alone yields significant gains.

Leaderboard Performance. On the official leaderboard, our system ranked 2nd overall out of 60 teams, with a mean F1 of 0.811 compared to 0.818 for the top-ranked system. We achieved 1st place in 3 languages (Amharic, Hindi, Swahili), top-3 in 8 languages, and top-10 in 17 out of 22 languages. Our weakest rankings are in Italian (25th) and Spanish (14th), which are also our lowest-performing languages in absolute terms. Full per-language rankings are provided in Appendix D.

Remaining Challenges. Italian (0.563) is our worst performing language by a wide margin. On closer inspection, the POLAR Italian split reveals a topic coverage gap rather than a simple class ratio shift. The training and development sets contain zero examples labeled with the `political` or `other` topic categories, yet these two categories account for 41% of the Italian test set (631 of 1538 samples) and roughly 87% of its positive class, as shown in Table 6. Since POLAR topic labels only co-occur with polarized rows, the model was effectively evaluated on a positive sub population it never observed during training.

Topic	Train	Dev	Test
political	0 (0.0%)	0 (0.0%)	412 (26.8%)
racial/ethnic	746 (22.4%)	37 (22.3%)	143 (9.3%)
religious	285 (8.5%)	14 (8.4%)	69 (4.5%)
gender/sexual	381 (11.4%)	19 (11.4%)	61 (4.0%)
other	0 (0.0%)	0 (0.0%)	219 (14.2%)
positive rate	41.0%	41.6%	47.3%

Table 6: Italian POLAR topic label and positive class distribution across splits. The `political` and `other` categories are entirely absent from train and development but account for 41% of test samples and about 87% of test positives, which explains the Italian test set performance drop.

6 Conclusion

We conducted an extensive analysis of Gemma-based models for multilingual polarization detection, investigating the effectiveness of synthetic data augmentation, ensemble methods, and per-language threshold optimization. We find that model generalization matters more than development set performance: Gemma was the only architecture that reliably transferred from development to test, while XLM-RoBERTa and Qwen3 suffered large drops. Threshold tuning provides 2 to 4% improvement without retraining. Ensemble methods combining 12B and 27B predictions are effective for 14 out of 22 languages. Synthetic data provides modest gains of 2 to 3%, primarily for low-resource languages, with 30% being the optimal ratio. Our final system achieves 0.811 mean macro-F1 across 22 languages, placing 2nd overall on the official leaderboard.

For future work, we plan to explore more sophisticated calibration methods such as temperature scaling and investigate topic coverage gaps of the kind we observed in Italian, where an entire class of polarization topics was absent from the training data.

Limitations

Our system trains separate models per language, requiring significant compute resources. The synthetic data quality for low-resource languages (Khmer, Swahili) may be limited, as GPT-4o-mini may not generate authentic text in these languages. Our threshold tuning is optimized on a small development set and may not fully transfer to the test distribution. Finally, the synthetic data mix of 50% direct generation, 30% paraphrasing, and 20% contrastive pairs was chosen heuristically, based on the intuition that direct generation should contribute the bulk of new topical content, paraphrasing should preserve label boundaries on real samples, and contrastive pairs should sharpen class separation. We did not run a systematic ablation over these strategy ratios and leave that study to future work.

Ethics Statement

Our system processes social media text that may contain offensive or polarizing content. The synthetic data generation process involved prompting LLMs to produce polarizing content for training purposes only. These synthetic samples are not representative of any real individuals or groups and are used solely for classifier training.

Acknowledgments

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2025. LLMs vs established text augmentation techniques for classification: When do the benefits outweigh the costs? In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized language models. In *Advances in Neural Information Processing Systems*, volume 36.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. 2014. Thresholding classifiers to maximize F1 score. *Computing Research Repository*, arXiv:1402.1892.

Shrirang Mhalgi, Srikar Kashyap Pulipaka, and Sandra Kübler. 2024. IUCL at PAN 2024: Using data augmentation for conspiracy theory detection. In

Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, Grenoble, France.

Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. 2021. Improving counterfactual generation for fair hate speech detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 92–101, Online. 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, 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.

Ignazio Pillai, Giorgio Fumera, and Fabio Roli. 2013. Threshold optimisation for multi-label classifiers. *Pattern Recognition*, 46(7):2055–2065.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Zheng Xin Yong, Cristina Menghini, and Stephen H. Bach. 2024. LexC-Gen: Generating data for extremely low-resource languages with large language models and bilingual lexicons. *Computing Research Repository*, arXiv:2402.14086.

A Full Per-Language Results

Table 7 shows complete test F1 scores for all 22 languages across our key submissions.

B Synthetic Data Impact by Language

Table 8 shows the optimal synthetic ratio for each language with the 12B model.

C Prompt Strategies

We used GPT-4o-mini for all synthetic data generation, with different prompt templates per strategy. All prompts instruct the model to output only the generated text in the target language, without labels, explanations, or meta-commentary.

Direct Generation. We provided the model with a definition of polarization and asked it to generate a social media post in the target language. For polarized samples, the prompt specifies that the post should target a specific group using vilifying, stereotyping, or dehumanizing language and show an us-vs-them mentality. For non-polarized samples, the prompt asks for a post that discusses the same types of topics in a neutral or balanced way, without vilifying anyone. Both prompts include a randomly sampled topic hint from five categories (political, ethnic/racial, religious, social class, international relations), each containing six specific topics. We used a temperature of 0.9 and a maximum of 250 tokens.

Paraphrasing. We provided the model with a real training sample and asked it to rewrite it using different words and phrasing while preserving the exact same meaning, sentiment, and level of polarization. The prompt specifies that the output should be in natural target-language expressions and within 20% of the original word count. We used a lower temperature of 0.7 to keep paraphrases close to the originals.

Contrastive Pairs. We asked the model to generate two versions of a social media post about a given topic: one polarized (containing vilification, stereotyping, or an us-vs-them mentality) and one non-polarized (neutral, balanced viewpoints). Both versions are required to discuss the same topic, be

Lang	Sub 1	Sub 4	Δ	Strategy	Threshold
amh	0.800	0.797	−0.3%	27B tuned	0.7
arb	0.848	0.848	+0.0%	27B tuned	0.4
ben	0.828	0.837	+0.9%	27B tuned	0.7
deu	0.721	0.728	+0.7%	Ens. weighted (0.3)	0.45
eng	0.805	0.818	+1.3%	Ens. weighted (0.6)	0.4
fas	0.804	0.828	+2.4%	Ens. weighted (0.6)	0.6
hau	0.793	0.800	+0.7%	Ens. weighted (0.7)	0.5
hin	0.800	0.824	+2.4%	Ens. average	0.6
ita	0.543	0.563	+2.0%	Ens. weighted (0.6)	0.45
khm	0.656	0.743	+8.7%	12B tuned	0.7
mya	0.874	0.877	+0.3%	12B tuned	0.35
nep	0.876	0.908	+3.2%	Ens. weighted (0.3)	0.3
ori	0.793	0.811	+1.8%	Ens. weighted (0.7)	0.4
pan	0.805	0.812	+0.7%	Ens. average	0.45
pol	0.814	0.835	+2.1%	Ens. average	0.3
rus	0.807	0.806	−0.1%	Ens. average	0.55
spa	0.770	0.779	+0.9%	Ens. weighted (0.6)	0.55
swa	0.774	0.811	+3.7%	Ens. average	0.65
tel	0.893	0.882	−1.1%	27B tuned	0.5
tur	0.802	0.809	+0.7%	Ens. weighted (0.3)	0.5
urd	0.803	0.803	−0.0%	12B tuned	0.35
zho	0.917	0.919	+0.2%	12B tuned	0.6
Mean	**0.797**	**0.811**	**+1.4%**		

Table 7: Complete per-language test F1 results with strategy and threshold used.

similar in length (30 to 60 words), and feel authentic to the target language. The output uses a structured format (`POLARIZED:` and `NON_POLARIZED:`) that we parse programmatically. We used a temperature of 0.8 and a maximum of 500 tokens.

Backtranslation and Cross-Lingual Transfer. In exploratory experiments, we used Google Cloud Translate for backtranslation by translating source text to a pivot language (English by default, with multi-pivot using English, German, French, and Spanish) and then back to the source language. For cross-lingual transfer experiments, we translated training data from a related higher-resource language to the target language (e.g., Hindi to Nepali, Bengali to Odia).

D Leaderboard Rankings

Table 9 shows our per-language rankings on the official Subtask 1 leaderboard.

Lang	Base	Best	Ratio
amh	0.781	0.781	0%
arb	0.814	0.828	30%
ben	0.832	0.832	0%
deu	–	0.792	30%
eng	0.808	0.811	50%
fas	0.847	0.849	30%
hau	0.766	0.804	20%
hin	0.848	0.865	30%
ita	0.686	0.701	50%
khm	0.628	0.692	30%
mya	0.853	0.887	20%
nep	0.870	0.890	10%
ori	0.891	0.891	0%
pan	0.859	0.859	0%
pol	0.792	0.834	30%
rus	0.781	0.827	30%
spa	0.757	0.757	0%
swa	0.819	0.833	10%
tel	0.907	0.907	0%
tur	0.809	0.817	10%
urd	0.785	0.799	10%
zho	0.921	0.930	30%

Table 8: Optimal synthetic ratio per language for Gemma 3 12B. Base = 0% synthetic. 6 out of 22 languages perform best without synthetic data.

Lang	Rank	Ours	Best	Δ
amh†	**1**	**0.800**	**0.800**	**0.000**
arb†	2	0.848	0.849	−0.0004
ben	12	0.837	0.863	−0.025
deu‡	8	0.728	0.761	−0.033
eng†	3	0.818	0.825	−0.008
fas‡	4	0.828	0.835	−0.007
hau	12	0.800	0.834	−0.034
hin†	**1**	**0.824**	**0.824**	**0.000**
ita	25	0.563	0.730	−0.167
khm‡	6	0.743	0.774	−0.032
mya‡	9	0.877	0.891	−0.014
nep	12	0.908	0.924	−0.016
ori‡	5	0.811	0.826	−0.015
pan†	2	0.812	0.826	−0.014
pol†	3	0.835	0.843	−0.008
rus‡	6	0.807	0.830	−0.024
spa	14	0.779	0.803	−0.024
swa†	**1**	**0.811**	**0.811**	**0.000**
tel‡	4	0.893	0.905	−0.012
tur†	3	0.809	0.833	−0.024
urd‡	4	0.803	0.820	−0.017
zho‡	7	0.919	0.932	−0.013
Avg	**2**	**0.811**	**0.818**	**−0.007**

Table 9: Per-language rankings on the official Subtask 1 leaderboard. Δ = Ours − Best. †Top 3, ‡Top 4–9. Bold = 1st place. Average Best score is computed over participants who participated in at least 11 of 22 languages.