UMUSP at SemEval-2026 Task 9: Mitigating Cross-Lingual Interference via Selective Multilingual and Multitask Specialization

Julio Cesar Alves Araujo Fuganti[1], Tulio Ferreira Leite da Silva[1,2], Adelino Gala[2], Francisco S. Marcondes[2*], José Machado[2], Paulo Novais[2]
[1]University of São Paulo, Brazil
[2]ALGORITMI Center/LASI, University of Minho, Portugal

Abstract

This paper proposes a selective multilingual and multitask fine-tuning strategy for online polarization detection that improves cross-lingual stability over fully joint training. Covering all three subtasks — polarization detection (POLARDETECT), polarization type classification (POLARTYPE), and rhetorical manifestation identification (POLARMANIFEST) — across all 22 languages of the shared task, the approach introduces controlled specialization, where languages and subtasks are grouped empirically and separate specialist models are fine-tuned for each subset. Restricting parameter sharing substantially improves performance even without ensemble averaging, whereas ensembling jointly trained models fails to mitigate instability. The final specialist ensemble improves Task 3 macro-F1 from 0.3330 to 0.4920 and reduces cross-lingual dispersion (CV: 0.613 → 0.321). Under the official ranking framework, the system ranks 7th among 16 submissions with complete multilingual and multitask coverage and remains within 5% of the best system in 37.70% of evaluation conditions.

1 Introduction

Online polarization threatens democratic institutions and social cohesion (Waller and Anderson, 2021; Piazza, 2023), amplified by echo chambers and biased content that escalate ideological divides into hate speech and offline harm (Garimella, 2018; Piazza, 2023), motivating automatic detection systems for content moderation, policy development, and discourse analysis across diverse linguistic and cultural contexts.

SemEval-2026 Task 9 (Naseem et al., 2026a) addresses this need through the POLAR benchmark (Naseem et al., 2026b), comprising over 110,000 instances across 22 languages, with three subtasks: binary polarization detection (POLARDETECT), type classification (POLARTYPE), and rhetorical manifestation identification (POLARMANIFEST). Baseline results reveal a consistent performance hierarchy—best on binary detection, worst on manifestation—with disparities especially pronounced for lower-resource languages (Naseem et al., 2026b), a pattern consistent with cross-lingual interference in joint multilingual training (Wang et al., 2020; Zhang et al., 2023), though it remains unclear under which conditions fully joint training consistently benefits all language–task pairs.

This instability is interpreted as negative transfer (Wang et al., 2019; Zhang et al., 2023) and addressed through a controlled specialization strategy: language–task combinations are grouped based on empirical compatibility, separate specialist models are trained for each subset, and predictions are combined through subtask-specific ensembling.

The contributions of this paper are threefold:

- **Empirical evidence of performance instability** under fully joint multilingual multitask training on POLAR;
- **A controlled specialization strategy** that restricts parameter sharing to compatible language–task groups, improving cross-lingual stability;
- **A disentanglement of the effects of specialization and ensembling**, showing that restricting parameter sharing yields consistent gains independently of ensemble averaging, while ensembling jointly trained models fails to mitigate instability.

2 Related Work

2.1 Multilingual Transfer

Multilingual models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) enable

*Corresponding author: fm@di.uminho.pt

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3199–3208

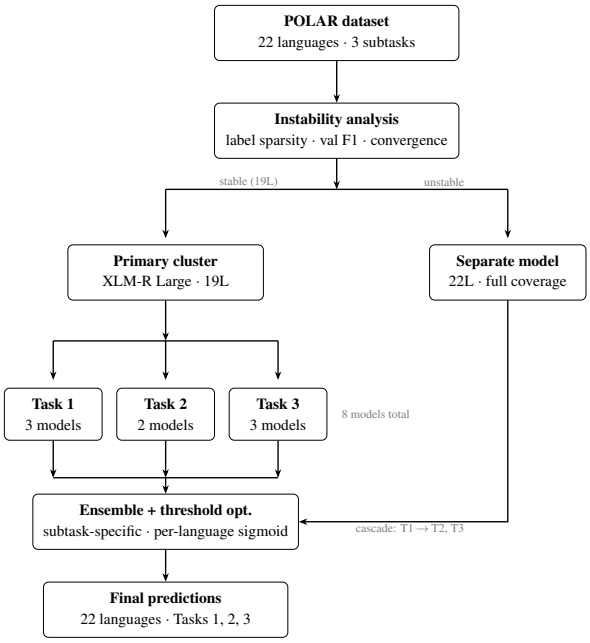

Figure 1: Overview of the selective specialisation pipeline. Languages identified as unstable are handled by a separate joint model; the remaining 19 languages are distributed across task-specific specialist models combined through subtask-specific ensembling with per-language threshold optimisation.

cross-lingual transfer from high- to lower-resource languages, but transfer effectiveness varies substantially across language pairs due to typological distance, domain mismatch, and data imbalance (Arivazhagan et al., 2019), resulting in uneven performance when languages differ significantly in distribution or supervision.

2.2 Multi-Task Learning and Negative Transfer

MTL improves performance through shared representations (Ruder, 2017), but insufficient task alignment can introduce negative transfer, degrading performance on one or more tasks (Wang et al., 2019; Zhang et al., 2023). Multilingual models are especially susceptible when linguistic or distributional differences are substantial (Arivazhagan et al., 2019; Wang et al., 2020).

2.3 Mitigating Negative Transfer

Strategies such as domain adaptation, selective transfer, and specialist models for language or domain subsets have been shown to improve robustness by isolating incompatible training signals while preserving beneficial transfer (Pan and Yang, 2010; Guo et al., 2018; Wang et al., 2019, 2020; Zhang et al., 2023). Our work adopts this perspective through controlled specialization restricted to empirically compatible language–task groups.

2.4 Cross-Lingual Stability and Evaluation

Aggregate metrics may obscure substantial variance across languages (Arivazhagan et al., 2019; Conneau et al., 2020), and recent work emphasizes assessing transfer reliability beyond global averages (Wang et al., 2019; Zhang et al., 2023). We analyze mean macro-F1 alongside standard deviation and coefficient of variation to characterize stability. Few prior studies explicitly disentangle unrestricted parameter sharing from ensembling under fully joint multilingual multitask training; our work addresses this gap.

3 Methodology

3.1 Base Architecture

XLM-RoBERTa Large (Conneau et al., 2020) is adopted as the backbone encoder for all tasks. With 24 transformer layers and 1024 hidden dimensions, XLM-R Large provides greater representational capacity than its base counterpart, with robust multilingual coverage across all 22 POLAR languages, including low-resource and non-Latin scripts. Its suitability is further supported by AfroXLMR-Large (Alabi et al., 2022)—the strongest POLAR baseline on African languages—being built upon it.

3.2 Empirical Observation of Negative Transfer

A fully joint configuration training all languages and subtasks simultaneously was first evaluated. The joint model exhibited instability across specific language–task combinations: certain languages experienced disproportionate performance drops, and Task 3 collapsed entirely for at least one language (minimum macro-F1 = 0.000). To understand this behavior, four configurations were compared: (1) fully joint training across all 22 languages, (2) joint training with a filtered language subset, (3) task-specialized training, and (4) selective grouping with controlled specialization and ensembling.

3.3 Selective Specialization Strategy

A deterministic grouping procedure based on label sparsity and empirical training stability is adopted. For each language ℓ, two statistics are computed: (i) the proportion of positive examples per label across subtasks and (ii) validation macro-F1 under fully

joint training. A language is marked *unstable* if at least one of the following holds for a language–task pair with available labeled data: (a) two or more labels have zero positive instances; (b) the overall positive rate for a subtask is below 1%; or (c) validation macro-F1 falls below 0.20 for any subtask after convergence. Languages that do not include a given subtask in the official dataset are not evaluated for sparsity under that subtask. Languages satisfying any of these conditions are excluded from the primary joint cluster. All grouping thresholds were defined during preliminary development and remained fixed for all subsequent experiments and final model selection.

Applying this procedure identifies Italian (zero positives for two Task 2 labels), Khmer (extreme polarization prevalence with highly skewed Task 3 distributions), and German (unstable convergence with oscillatory validation behavior and repeated performance collapse in Task 3) as unstable under fully joint optimization. Filtering these languages from the 22-language configuration increases Task 3 macro-F1 from 0.3330 to 0.4135, and further to 0.4241 with positional weighting.

The remaining languages form the primary specialist cluster (19 languages). Specialist models are fine-tuned on this cluster with identical training settings to the joint baseline. The excluded languages are handled by a separate model trained on the full 22-language data to preserve complete multilingual coverage. Per-language label statistics are reported in Appendix A.

3.4 Ensemble Mechanism

Final predictions are obtained via subtask-specific ensembling of specialist models. For Subtask 1, logits are combined using a fixed weighted average; for Subtasks 2 and 3, logits are averaged uniformly. A per-label sigmoid produces independent probabilities, and thresholds are optimized per subtask and per language on the validation set. When multiple thresholds yield macro-F1 within 0.005 of the maximum for a given language–subtask pair, the largest threshold is selected to favor precision under label sparsity.

A lightweight task-specific gating layer is optionally applied to the pooled encoder representation before classification. The gate consists of a learned linear projection followed by element-wise scaling, enabling modest task-conditioned modulation without altering the backbone parameters.

Predictions follow a cascade: only instances classified as polarized in Subtask 1 are passed to Subtasks 2 and 3. This cascade is applied uniformly across all model configurations (joint and specialist) to ensure controlled comparison and enforce label consistency.

The final system deploys eight models. For the primary 19-language cluster, Subtask 1 ensembles three XLM-RoBERTa-Large models (positional weighting + oversampling; positional weighting + temperature balancing; gated + positional weighting). Subtask 2 combines two of these models with a gated model trained jointly on Subtasks 1–2. Subtask 3 ensembles a gated XLM-RoBERTa-Large, a gated mDeBERTa-base, and a jointly trained XLM-RoBERTa-Large model. The three excluded languages are handled by a single joint model to preserve full coverage.

4 Experimental Setup

4.1 Data and Splits

We used the official shared task datasets. Before the official validation set was released, training data was split 80/20; afterwards, the full training set was used for training and the official validation set for evaluation.

4.2 Training Configuration

XLM-RoBERTa-Large was fine-tuned for up to 6 epochs using AdamW (peak learning rate 2e-5, LLRD 0.95), with early stopping (patience 3, minimum delta 0.001, macro-F1). Effective batch size was 64 (32 per device, 2 gradient accumulation steps), with maximum sequence length 128. Training used NVIDIA A100 GPUs with BF16 and TF32. Positional weighting on BCEWithLogitsLoss mitigated class imbalance in Tasks 2 and 3, scaling positive labels inversely proportional to their frequency per language–subtask combination.

4.3 Evaluation Metrics

The primary metric is macro-F1, averaged across languages per subtask. Cross-lingual stability is assessed via standard deviation and coefficient of variation (CV) over per-language scores. All model selection and grouping decisions were made using the validation set; final evaluation results are reported on the official test set to mitigate overfitting concerns.

5 Results

5.1 Comparison with the POLAR Baseline

Per-language macro-F1 on the test set comparing the POLAR baseline (Naseem et al., 2026b) (M1, as reported on the official leaderboard) against the specialist ensemble (M2) is reported in Appendix B (Table 6). The specialist ensemble outperforms the POLAR baseline in the large majority of language–task combinations across all three subtasks. Notable exceptions occur for languages excluded from the primary specialist cluster (Italian, Khmer) and for Task 3 in a small number of high-resource languages where the joint baseline benefits from denser supervision signal.

5.2 Absolute Performance

Table 1 presents the progressive impact of controlled specialization. Fully joint training across all 22 languages yields the lowest performance, particularly for Tasks 2 and 3. Removing empirically unstable languages (Italian, Khmer, and German) substantially improves Task 3, and adding positional weighting further stabilizes minority labels. Task-specialized configurations produce additional gains. The final specialist ensemble achieves the strongest overall performance across all subtasks. Unless otherwise specified, all ablation and dispersion results are reported on the official validation set.

5.3 Disentangling Specialization and Ensembling

All comparisons isolate parameter-sharing structure while keeping language subsets, loss weighting, and cascade design fixed within each comparison. Threshold optimization is applied consistently inside each configuration family.

As shown in Table 8 (Appendix C), ensembling jointly trained multilingual models degrades macro-F1 across all subtasks relative to the single joint model. In contrast, specialist models evaluated individually already outperform the joint configuration, demonstrating that restricting parameter sharing yields consistent gains independently of ensemble averaging. Ensemble aggregation over specialists provides additional but smaller improvements, confirming that the primary performance gains arise from controlled specialization rather than ensembling alone.

5.4 Cross-Lingual Stability

Table 2 reports dispersion statistics for the fully joint baseline and the final specialist ensemble (with threshold optimization). The specialist ensemble improves both mean performance and cross-lingual consistency across all subtasks. The largest effect occurs in Task 3, where the baseline collapses to a minimum per-language F1 of 0.000 (CV = 0.613), while the specialist ensemble raises this floor to 0.213 and reduces CV to 0.321.

5.5 Shared-Task Robustness

Table 7 (Appendix C) evaluates robustness under the official shared-task ranking framework. Our system ranks 7th among 16 systems with complete multilingual and multitask coverage. It remains within 5% of the best system in 37.70% of evaluation conditions, and exceeds a 20% performance drop in 18.03% of cases. These results reflect competitive but conservative behavior, prioritizing cross-lingual stability over aggressive optimization for specific language–task pairs.

6 Analysis

6.1 Conditions for Negative Transfer

Negative transfer is most evident in language–task combinations characterized by extreme label imbalance or sparse supervision. Task 3 provides the clearest illustration: under fully joint multilingual multitask training, at least one language collapses entirely (minimum macro-F1 = 0.000; CV = 0.613). This collapse indicates unstable cross-lingual parameter sharing rather than insufficient model capacity.

Languages with highly skewed distributions exacerbate this instability. For example, Hausa exhibits near-zero positive rates in Task 2, while Khmer presents extremely high polarization prevalence and strongly imbalanced manifestation labels. When trained jointly with more balanced languages, such distributions introduce gradient signals that disproportionately affect shared representations, harming vulnerable language–task pairs. Per-language breakdowns in Appendix B confirm this pattern on the test set: Khmer and Italian, which are handled by the fallback joint model, show degraded Task 3 performance relative to the POLAR baseline, while the 19-language specialist cluster consistently improves. Languages handled by the fallback joint model consistently underperform the specialist cluster, suggesting that isolation

Model Configuration	Tasks	Lang.	Task 1	Task 2	Task 3
XLM-R Large (joint multilingual multitask)	1, 2, 3	22	0.7771	0.4808	0.3330
XLM-R Large (joint, no ita, khm, deu)	1, 2, 3	19	0.8067	0.5129	0.4135
XLM-R Large + Positional Weighting	1, 2, 3	19	0.8028	0.5246	0.4241
XLM-R Large + Positional Weighting (task-specialized)	1, 2	19	**0.8225**	**0.5511**	–
XLM-R Large + Positional Weighting (task-specialized)	1, 3	19/16	0.8178	–	**0.4906**
XLM-R Large + Positional Weighting (task-specialized)	2, 3	19/16	–	0.4800	0.4125
Selective Specialist Ensemble + Threshold Opt. (final)	1, 2, 3	19	**0.8431**	**0.5861**	**0.4920**

Table 1: Ablation study showing the effect of controlled specialization on multilingual multitask polarization detection (average macro-F1 per subtask). Results reported on the official validation set.

Subtask	Configuration	Mean ↑	Std. Dev. ↓	CV ↓	Min ↑	Max ↑
Task 1	Baseline (22L Multi-task)	0.777	0.081	0.104	0.585	0.910
	Specialist Ensemble	**0.848**	**0.047**	**0.055**	**0.744**	**0.920**
Task 2	Baseline (22L Multi-task)	0.481	0.161	0.335	0.184	0.772
	Specialist Ensemble	**0.586**	**0.135**	**0.231**	**0.359**	**0.790**
Task 3	Baseline (22L Multi-task)	0.333	0.204	0.613	0.000	0.789
	Specialist Ensemble	**0.492**	**0.158**	**0.321**	**0.213**	**0.801**

Table 2: Cross-lingual performance dispersion on the validation set. Mean is the unweighted macro-F1 across languages. Std. Dev. and CV measure cross-lingual consistency (lower is better). Min and Max represent per-language extremes.

mitigates interference but does not fully compensate for extreme label imbalance.

Empirically, filtering unstable languages (Italian, Khmer, German) increases Task 3 macro-F1 from 0.3330 to 0.4135. Adding positional weighting further improves performance to 0.4241 without architectural modification. These gains confirm that the observed degradation stems from negative transfer induced by incompatible language–task distributions, rather than representational limitations of the encoder.

6.2 Specialists vs. Generalists

Ensembling jointly trained multilingual models degrades macro-F1 across all subtasks relative to the single joint baseline (e.g., Task 3: 0.3330 → 0.3101). In contrast, specialist models evaluated individually already outperform the joint configuration (Task 3 mean = 0.4305), indicating that restricting parameter sharing yields consistent improvements independently of ensemble averaging.

Ensemble aggregation over specialists provides additional but smaller gains (Task 3: 0.4305 → 0.4425 without threshold optimization; 0.4920 in the final system with threshold optimization). The primary performance improvements therefore arise from controlled specialization rather than ensembling alone. Controlled parameter sharing achieves a more favorable balance between transfer and interference mitigation than either unrestricted joint training or full task isolation.

7 Conclusion

Experiments on POLAR demonstrate that unrestricted parameter sharing under fully joint multilingual multitask training can induce negative transfer, resulting in instability and performance collapse for specific language–task combinations. Restricting parameter sharing to empirically compatible language clusters and combining specialist models through subtask-specific ensembling yields consistent improvements in macro-F1 and cross-lingual stability.

The final specialist ensemble improves Task 3 macro-F1 from 0.3330 to 0.4920 and reduces cross-lingual dispersion (CV: 0.613 → 0.321), substantially raising the performance floor across languages. These findings indicate that robustness and stability should be treated as primary evaluation dimensions in multilingual benchmarks where supervision quality and label distributions vary significantly across languages.

Limitations

The grouping strategy relies on validation data to identify unstable language–task combinations and may not generalize in settings with limited development supervision. Controlled specialization increases computational cost due to multiple fine-tuning runs and ensemble construction. Additionally, grouping decisions are heuristic rather than formally optimized; future work could explore principled compatibility estimation, adaptive parameter-sharing mechanisms, or dynamic routing strategies to mitigate negative transfer without manual cluster definition. Code is available at https://github.com/Fugant1/UMUSP-SemEval2026-Task9.

Acknowledgements

This work was supported by AMALIA (Automatic Multimodal Language Assistant with Artificial Intelligence), the Portuguese Large Language Model Project included in measure RE-C05-i08 of the national "Programa de Recuperação e Resiliência." Tulio Ferreira Leite da Silva also acknowledges the support of the São Paulo Research Foundation (FAPESP), grant no. 20/15160-7.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 959–986.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *Preprint*, arXiv:1907.05019.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Kiran Garimella. 2018. *Polarization on Social Media*. Ph.D. thesis, Aalto University, Finland.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703, Brussels, Belgium.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acarürk, 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 Acarürk, 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.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

James A. Piazza. 2023. Political polarization and political violence. *Security Studies*, 32(3):476–504.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Isaac Waller and Ashton Anderson. 2021. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888):264–268.

Zirui Wang, Zihang Dai, Barnabas Poczos, and Jaime Carbonell. 2019. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11285–11294.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of EMNLP*, pages 4438–4450.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. 2023. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329.

A Per-Language Label Distributions

This appendix reports per-language positive label proportions for Tasks 1–3. These distributions illustrate the heterogeneity, sparsity, and skewness

conditions used to define unstable language–task combinations in Section 3. In particular, extreme imbalance (near-zero positives) and highly skewed label prevalence directly motivated the filtering and specialization strategy discussed in Sections 3 and 6.

Language	Polarization	Political	Racial/Ethnic	Religious	Gender/Sexual	Other
arb	0.447	0.232	0.172	0.084	0.109	0.167
ben	0.427	0.340	0.007	0.019	0.005	0.101
mya	0.581	0.253	0.053	0.031	0.106	0.451
eng	0.365	0.357	0.087	0.035	0.022	0.039
hin	0.853	0.735	0.121	0.587	0.115	0.131
nep	0.503	0.172	0.140	0.079	0.052	0.118
urd	0.695	0.673	0.544	0.553	0.512	0.507
zho	0.496	0.059	0.227	0.020	0.169	0.086
amh	0.755	0.667	0.259	0.020	0.006	0.248
hau	0.107	0.049	0.032	0.026	0.008	0.004
ori	0.289	0.210	0.050	0.063	0.033	0.037
fas	0.739	0.439	0.024	0.096	0.060	0.242
pol	0.420	0.366	0.090	0.036	0.046	0.065
pan	0.493	0.308	0.059	0.079	0.112	0.089
rus	0.306	0.139	0.098	0.041	0.056	0.024
spa	0.503	0.273	0.189	0.159	0.134	0.134
swa	0.501	0.027	0.355	0.035	0.022	0.079
tel	0.537	0.216	0.170	0.090	0.133	0.238
tur	0.489	0.447	0.169	0.152	0.048	0.048
ita	0.411	0.000	0.224	0.085	0.114	0.000
khm	0.908	0.183	0.015	0.034	0.017	0.659
deu	0.476	0.409	0.185	0.111	0.059	0.138

Table 3: Proportion of positive examples by language (Task 1 and Task 2).

Language	Stereotype	Vilification	Dehumanization	Extreme Language	Lack of Empathy	Invalidation
arb	0.334	0.372	0.109	0.304	0.170	0.081
ben	0.060	0.241	0.107	0.047	0.019	0.018
mya	0.000	0.000	0.000	0.000	0.000	0.000
eng	0.151	0.265	0.121	0.240	0.111	0.182
hin	0.497	0.654	0.182	0.506	0.568	0.658
nep	0.268	0.314	0.066	0.271	0.106	0.150
urd	0.623	0.647	0.556	0.621	0.562	0.572
zho	0.301	0.185	0.050	0.081	0.079	0.048
amh	0.547	0.484	0.132	0.306	0.176	0.160
hau	0.043	0.012	0.035	0.030	0.009	0.002
ori	0.100	0.116	0.007	0.134	0.016	0.034
fas	0.131	0.573	0.043	0.169	0.099	0.080
pol	0.000	0.000	0.000	0.000	0.000	0.000
pan	0.164	0.403	0.220	0.239	0.124	0.244
rus	0.000	0.000	0.000	0.000	0.000	0.000
spa	0.278	0.306	0.089	0.242	0.239	0.106
swa	0.396	0.412	0.128	0.239	0.298	0.234
tel	0.112	0.225	0.025	0.134	0.263	0.228
tur	0.408	0.324	0.109	0.432	0.096	0.040
ita	0.000	0.000	0.000	0.000	0.000	0.000
khm	0.683	0.015	0.012	0.023	0.110	0.065
deu	0.359	0.301	0.149	0.218	0.267	0.163

Table 4: Proportion of positive examples by language (Task 3).

B Per-Language Results on Validation and Test Sets

Tables 5 and 6 report per-language macro-F1 scores for all 22 languages across Tasks 1–3, comparing the joint baseline (M1) against the specialist ensemble (M2). Table 5 covers the official validation set; Table 6 covers the official test set, with M1 corresponding to the POLAR baseline (Naseem et al., 2026b) as reported on the official leaderboard. Dashes (—) indicate languages for which Task 3 annotations are not provided in the official dataset. Bold values indicate the better-performing system for each language–task pair.

	Task 1		Task 2		Task 3	
Lang.	**M1**	**M2**	**M1**	**M2**	**M1**	**M2**
arb	**0.844**	0.826	0.537	**0.638**	0.412	**0.596**
ben	**0.859**	0.852	0.184	**0.400**	0.082	**0.249**
mya	0.866	**0.901**	0.365	**0.523**	—	—
eng	0.813	**0.822**	**0.367**	0.359	0.375	**0.500**
hin	0.755	**0.829**	0.636	**0.789**	0.744	**0.749**
nep	0.909	**0.909**	0.762	**0.785**	0.457	**0.545**
urd	0.730	**0.789**	0.771	**0.789**	0.789	**0.801**
zho	0.882	**0.910**	0.618	**0.661**	0.294	**0.443**
amh	0.709	**0.779**	0.368	**0.472**	0.278	**0.494**
hau	0.783	**0.803**	0.249	**0.452**	0.000	**0.213**
ori	0.767	**0.837**	0.547	**0.677**	0.131	**0.276**
fas	0.770	**0.859**	0.558	**0.582**	0.189	**0.379**
pol	**0.833**	0.827	0.500	**0.629**	—	—
pan	0.745	**0.879**	0.329	**0.435**	0.494	**0.593**
rus	0.804	**0.829**	0.532	**0.674**	—	—
spa	0.736	**0.744**	0.617	**0.654**	0.386	**0.457**
swa	0.802	**0.821**	0.313	**0.417**	0.374	**0.566**
tel	0.648	**0.898**	0.272	**0.491**	0.126	**0.435**
tur	0.860	**0.895**	0.522	**0.699**	0.395	**0.570**
ita	**0.638**	0.623	**0.367**	0.361	—	—
khm	**0.584**	0.520	**0.681**	0.633	0.159	**0.204**
deu	**0.748**	0.741	**0.471**	0.464	0.302	**0.367**

Table 5: Per-language macro-F1 on the validation set. M1 = joint baseline; M2 = specialist ensemble.

	Task 1		Task 2		Task 3	
Lang.	**M1**	**M2**	**M1**	**M2**	**M1**	**M2**
arb	0.795	**0.831**	0.485	**0.625**	0.390	**0.584**
ben	**0.852**	0.819	0.288	**0.307**	0.086	**0.236**
mya	0.821	**0.872**	0.477	**0.686**	—	—
eng	0.780	**0.787**	0.333	**0.427**	0.410	**0.485**
hin	0.737	**0.805**	**0.791**	0.713	0.234	**0.710**
nep	0.879	**0.900**	0.721	**0.751**	0.131	**0.535**
urd	0.789	**0.793**	0.712	**0.772**	0.531	**0.802**
zho	0.869	**0.896**	0.669	**0.751**	0.000	**0.388**
amh	0.715	**0.770**	0.371	**0.561**	0.443	**0.490**
hau	0.775	**0.797**	0.203	**0.363**	**0.745**	0.143
ori	0.776	**0.813**	0.560	**0.561**	**0.384**	0.232
fas	**0.842**	0.800	0.462	**0.551**	0.200	**0.347**
pol	0.724	**0.806**	0.449	**0.528**	—	—
pan	**0.789**	0.761	0.365	**0.392**	0.456	**0.481**
rus	0.745	**0.775**	**0.590**	0.572	—	—
spa	0.726	**0.789**	0.593	**0.650**	**0.508**	0.479
swa	0.757	**0.774**	**0.441**	0.426	0.220	**0.543**
tel	0.644	**0.881**	0.314	**0.440**	**0.673**	0.423
tur	0.695	**0.779**	0.470	**0.563**	**0.769**	0.488
ita	**0.677**	0.501	**0.375**	0.165	—	—
khm	0.659	**0.730**	**0.626**	0.608	**0.609**	0.343
deu	0.671	**0.723**	0.407	**0.512**	0.348	**0.495**

Table 6: Per-language macro-F1 on the test set. M1 = POLAR baseline (Naseem et al., 2026b); M2 = UMUSP specialist ensemble.

C Additional Results

Team	Mean Drop ↓	Variance ↓	Worst Drop ↓	% within 5% ↑	% >20% ↓
SMASH	**0.0320**	**0.0021**	0.2029	**80.65**	**3.23**
NYCU-NLP	0.0560	0.0066	0.3328	67.74	9.68
PolaFusion	0.0787	0.0069	0.5351	45.16	6.45
YEZE	0.0959	0.0082	0.4213	37.10	12.90
Sagarmatha	0.0985	0.0092	**0.5817**	37.10	12.90
AIvengers	0.1011	0.0088	0.6120	33.87	9.68
Lingo Research Group	0.1268	0.0170	0.5264	37.10	19.35
UMUSP	0.1285	0.0152	0.6995	37.10	22.58
MSqrd	0.1288	0.0096	0.5035	17.74	19.35
DigiS-FBK	0.1886	0.0296	0.7809	14.52	37.10
maggam	0.2835	0.0308	0.8140	6.45	67.74

Table 7: Cross-lingual robustness comparison among all 11 systems with complete multilingual and multitask coverage (62 evaluation conditions: 22 languages × Subtasks 1–2 + 18 languages × Subtask 3). mdok-style is excluded as an outlier: its score of 0.000 in ST3/hau yields a worst drop of 1.000, rendering cross-lingual robustness metrics non-comparable. Mean Drop and Worst Drop measure relative degradation compared to the best system per condition. Variance captures dispersion across languages and tasks. Bold values indicate best performance per column.

Configuration	Ensemble	Languages	Task 1	Task 2	Task 3
Joint multilingual multitask (baseline)	No	22	0.7771	0.4808	0.3330
Joint multilingual multitask	Yes	22	0.7651	0.4541	0.3101
Specialist models (mean)	No	18–19	0.8110	0.5403	0.4305
Specialist ensemble (without threshold optimization)	Yes	18–19	0.8138	0.5525	0.4425

Table 8: Effect of specialization and ensembling on multilingual multitask performance (macro-F1). "Specialists (mean)" corresponds to the average performance of individually evaluated specialist models without ensembling.