Gladiators at #SMM4H–HeaRD 2026: Multi-Seed XLM-RoBERTa Ensemble with Focal Loss and Per-Language Threshold Optimization for Multilingual Adverse Drug Event Detection

Ankit Singh
singhankit16@gmail.com

Abstract

This paper describes the Gladiators system for Task 1 of the SMM4H 2026 shared task on binary classification of adverse drug event (ADE) mentions in multilingual social media posts. Our system fine-tunes three XLM-RoBERTa-large models with different random seeds using focal loss (α=0.75, γ=2.0) and 3$\times$ positive oversampling, then averages their predicted probabilities and applies per-language threshold optimization. On the development set, our ensemble achieves a pooled binary F1 of 0.7505. On the official test set—which introduced surprise Farsi comprising 35.5% of samples—our system achieves F1 = 0.6039, above the competition mean (0.5465) and median (0.5798). We evaluated eleven approaches, document key negative results including catastrophic translation augmentation failures, and analyze precision–recall trade-offs under severe class imbalance. Post-evaluation, a six-model cross-regime ensemble improved dev F1 to 0.7585.

1 Introduction

Adverse drug events (ADEs) represent a significant public health concern, and social media has become an important source for pharmacovigilance signals (Huynh et al., 2016). The SMM4H 2026 Task 1 challenges participants to detect ADE mentions in multilingual social media posts across six training languages—English (en), German (de), French (fr), Japanese (ja), Russian (ru), and Chinese (zh)—with a surprise seventh language (Farsi) at test time.

Key challenges include: (1) severe class imbalance (2.4–10.1% positive), (2) diverse domains from tweets to forum posts, (3) limited training data for de/fr (70–83 positives), and (4) zero-shot transfer to an unseen script. We address these with XLM-RoBERTa-large (Conneau et al., 2020), focal loss (Lin et al., 2017), positive oversampling, and per-language threshold optimization.

Lang	Train	Dev	Pos%	Source
en	17,128	888	7.0	Tweets
ja	14,208	3,045	2.4	Tweets
ru	10,695	2,669	10.1	Reviews/Tweets
zh	2,248	379	10.0	Health Q&A
de	1,482	634	5.6	Forum posts
fr	976	418	7.2	Forum posts
Total	**46,737**	**8,033**	**6.4**	—

Table 1: Training and development data. Pos% is training positive rate.

2 Background and Datasets

Task 1 requires binary classification of social media posts for ADE mentions, evaluated via pooled binary F1-score across all non-CADEC samples. A CADEC subset (Karimi et al., 2015) of translated drug reviews (de/fr) is evaluated separately.

Training data. 46,737 samples across six languages (Table 1) with 6.4% overall positive rate. Sources include tweets (en, ja, ru), patient forums (de from KEEPHA, fr), drug reviews (Tutubalina et al., 2020), and health Q&A (zh). Class imbalance is severe, especially in Japanese (2.36%). German and French have only 83 and 70 positive samples with text 10$\times$ longer than tweets (625 vs 55 chars), creating significant domain heterogeneity.

Test data. 42,736 samples with surprise Farsi (15,184 samples, 35.5%)—the largest single language—followed by en (27.4%), ru (21.7%), ja (7.1%), zh (2.7%), de (2.6%), fr (2.6%).

3 Experimental Setup

3.1 Model and Training

We fine-tune XLM-RoBERTa-large (Conneau et al., 2020) (560M params) with a two-class classification head, max sequence length 256. Training uses AdamW (Loshchilov and Hutter, 2019) (lr=1×10^{-5}, cosine schedule, 10% warmup,

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 7–11

weight decay 0.01), effective batch size 32 (micro-batch 8, gradient accumulation 4), fp16, gradient clipping at 1.0 for 5 epochs on a single NVIDIA RTX 5070 Ti (12.8 GB VRAM, peak 11.8 GB used) with HuggingFace Transformers (Wolf et al., 2020).

Convergence and model selection. We train for a fixed 5 epochs with cosine learning rate decay to zero, evaluating dev F1 after each epoch and saving the best checkpoint. Training loss decreased monotonically across all seeds (seed 42: 0.033→0.014→0.007→0.003→0.002), indicating stable convergence without oscillation. We did not employ early stopping; instead, the cosine schedule naturally reduces the learning rate to zero by epoch 5, providing implicit regularization. The best epoch varied across seeds (seed 42: epoch 5, seed 123: epoch 3, seed 456: epoch 2), which motivates our ensemble approach—the variance in optimal stopping points suggests different seeds find complementary local minima, and averaging their predictions at inference is more robust than selecting any single checkpoint.

Text preprocessing. Light normalization: URLs → `[URL]`, mentions → `@USER`, PII → `[PLACE]`, repeated chars collapsed. Casing and function words are preserved.

3.2 Handling Class Imbalance

Focal loss (Lin et al., 2017) with γ=2.0, α=0.75: $\mathcal{L}_{\text{FL}} = -\alpha_t(1-p_t)^\gamma \log(p_t)$, down-weighting easy negatives. The combination of focal loss and oversampling primarily improves *recall* for minority-class positives: without these, the model achieves high precision but near-zero recall on languages like Japanese (2.36% positive), predicting almost everything as negative. With our imbalance handling, the ensemble achieves balanced precision and recall (e.g., ja: P=0.72, R=0.61 at τ=0.70), trading a small precision decrease for substantial recall gains.

3× oversampling of positives, raising representation from 6.4% to ∼17% (54,503 samples).

Per-language thresholds. Grid search over $[0.20, 0.80]$ maximizing F1 per language on dev. Optimal thresholds ranged from 0.32 (en) to 0.70 (ja), reflecting that languages with more training positives produce better-calibrated, more confident predictions requiring higher thresholds.

3.3 Multi-Seed Ensemble

Three models with seeds 42, 123, 456; averaged $P(\text{ADE})$ with re-optimized thresholds. Seeds showed complementary strengths: seed 42 peaked at epoch 5 (F1=0.731, best de/zh), seed 123 at epoch 3 (F1=0.738, best fr/ja), seed 456 at epoch 2 (F1=0.728, best en). No single seed dominated, and best epochs varied from 2 to 5.

Why 3 seeds? We chose 3 seeds as a practical trade-off: the ensemble gain from 1→3 models was +1.95% F1 (0.731→0.7505), while adding a 4th model of comparable quality (the best individual Farsi-augmented seed at 0.740) to form a 4-model ensemble yielded diminishing returns. Computational cost scales linearly (∼105 min/seed), and 3 models kept total training under 6 hours on a single GPU.

4 Alternative Approaches

Table 2 summarizes all eleven approaches.

Zero-shot baselines. GLiNER (Zaratiana et al., 2023) (NER, F1=0.244) and mDeBERTa-XNLI (Laurer et al., 2024) (NLI, F1=0.155) failed due to low precision: GLiNER predicted 622 positives (>500 FP) and NLI predicted 5,177 (∼4,700 FP) from only 508 true positives.

XLM-R-base (278M). Achieved F1=0.704. Scaling to large improved all languages, with the largest gains for data-scarce languages: de +15.7%, zh +9.5%.

Translation augmentation for de/fr. Translating EN/RU positives to DE/FR via NLLB-200 (NLLB Team et al., 2022) (4,560 samples) *hurt* performance (0.704→0.687; fr −8.3%). The domain mismatch between source tweets (short, informal) and target forum posts (long, formal medical narratives) meant translated instances introduced distributional noise rather than useful signal.

GLiNER feature ensemble. Logistic regression combining XLM-R probabilities with 9 GLiNER features yielded F1=0.704—identical to XLM-R alone. Feature importance: XLM-R coefficient +7.01 vs <+0.23 for all GLiNER features.

Japanese BERT. `bert-large-japanese-v2` (Devlin et al., 2019) with MeCab tokenization achieved ja F1=0.662—identical to the multilingual ensemble.

The bottleneck is data scarcity (335 positives), not tokenization.

Architecture-diverse ensemble. Adding mDeBERTa-v3-base (F1=0.686) to the 3-seed XLM-R ensemble yielded F1=0.7502—worse than XLM-R-only (0.7505). The quality gap causes dilution. *Practical note*: mDeBERTa-v3's fp16 HuggingFace weights cause NaN; requires fp32 loading with gradient checkpointing (2.3× slower).

4.1 Post-Evaluation: Farsi Augmentation

v1: Positives-only (failed). Translating 2,280 EN/RU positives into Farsi caused the model to learn "Farsi script = positive" (99.9% predicted positive)—a *Clever Hans* effect invisible on dev (-0.34% F1 only) since dev has no Farsi.

v2: Balanced (successful). Adding 6,000 negatives (8,280 total, 27.5% positive) fixed the issue (dev F1=0.7512). However, augmentation disrupted probability calibration across *all* languages: notably, the Japanese threshold shifted from 0.70 to 0.27—a dramatic change indicating that introducing 8,280 Farsi samples (15% of the augmented dataset) altered the model's learned probability distributions globally, not just for Farsi. This is a significant risk of multilingual fine-tuning: augmentation for one language can silently degrade calibration for others, requiring full re-optimization of all thresholds.

Six-model weighted ensemble. Combining original and Farsi-augmented models: $P_{\text{final}} = (w_{\text{orig}} \cdot \bar{P}_{\text{orig}} + w_{\text{v2}} \cdot \bar{P}_{\text{v2}})/(w_{\text{orig}} + w_{\text{v2}})$ with w_{orig}=1.0, w_{v2}=0.7 achieved best dev F1=0.7585, improving *every language*. Not submitted due to timing.

5 Results

Per-language dev F1 and optimized thresholds for the submitted 3-seed ensemble: de .698/τ=.56, en .815/.32, fr .783/.36, ja .662/.70, ru .730/.49, zh .960/.66.

Official test results. Table 3 shows test performance vs. the competition. Our system (F1=0.6039) exceeded the mean (+5.7%) and median (+2.4%), with the strongest advantages on ja (+5.7% vs median), fa (+5.5%), and CADEC-de (+4.4%). German was our only below-median language (-0.5%).

Approach	Dev F1
GLiNER zero-shot NER	0.244
NLI zero-shot (mDeBERTa-XNLI)	0.155
XLM-R-base fine-tuned	0.704
XLM-R-base + translation aug.	0.687
XLM-R-base + GLiNER ensemble	0.704
XLM-R-large (best single seed)	0.738
3-seed XLM-R-large ensemble	**0.7505**
Japanese BERT (ja only)	0.662
4-model (3×XLM-R + mDeBERTa)	0.7502
Post-evaluation:	
3-seed Farsi-augmented (v2)	0.7512
6-model weighted ensemble	0.7585

Table 2: All approaches evaluated (pooled binary F1 on dev). Bold = official submission.

Lang	N	Ours	Mean	Med.	Dev
en	11.7k	.726	.685	.701	.815
de	1.1k	.651	.664	.656	.698
fr	1.1k	.705	.681	.696	.783
ja	3.0k	.606	.534	.549	.662
ru	9.3k	.553	.533	.550	.730
zh	1.1k	.826	.804	.821	.960
fa	15.2k	.435	.367	.380	—
de_c	87	.904	.833	.860	.914
fr_c	87	.883	.843	.883	.914
All	**42.6k**	**.604**	**.547**	**.580**	**.751**

Table 3: Test F1 vs. competition mean/median, with dev F1 for gap comparison. de_c/fr_c = CADEC (excluded from All).

Generalization gap. The dev→test gap (-14.7%) stems from (a) surprise Farsi (35.5% of test, F1=0.435) and (b) distribution shift, especially in Russian (0.730→0.553, gap -17.8%). Russian's gap was universal across teams (competition median ru F1=0.550), strongly suggesting test-set distribution shift—likely a change in source platform or time period for the Russian test data (e.g., different drug discussion forums or a shift toward newer drugs not represented in the RuDReC-era training data) rather than model-specific overfitting. Chinese also dropped notably (0.960→0.826).

6 Discussion

Why zero-shot Farsi worked relatively well. Our Farsi F1 (0.435) exceeded the competition mean (0.367) by +6.8% despite zero Farsi training data. We attribute this to XLM-R-large's pretraining on 2.5TB of data covering 100 languages including Farsi, which provides shared subword representations between Farsi and languages with similar medical vocabulary (borrowed English/French

pharmaceutical terms are common in Farsi medical text). Additionally, our per-language threshold fallback (0.515 = mean of all dev thresholds) happened to be reasonable; post-hoc analysis showed Farsi's probability distribution had 0.873 histogram overlap with Russian's, and Farsi's mean predicted probability (0.106) closely matched French (0.108), suggesting the model successfully projected Farsi into a similar representation space as trained languages.

Forum posts vs. tweets. The model handled long forum posts (de/fr, avg 625 chars) and short tweets (ja/en, avg 55–101 chars) differently. For tweets, the model relies on concise ADE signals ("made me sick," "terrible side effects"); for forum posts, it must filter through lengthy medication histories and identify ADE mentions buried in multi-paragraph narratives. German's lower F1 (0.698 dev, 0.651 test) despite higher positive rate than Japanese reflects this difficulty: forum posts frequently discuss medications *without* reporting ADEs, creating false positive pressure that tweets rarely exhibit. The 256-token truncation also affects $\sim$5% of German/French posts, potentially losing relevant context.

Calibration risks of multilingual augmentation. The Farsi v2 augmentation (8,280 samples, 15% of augmented data) caused the Japanese threshold to shift from 0.70$\rightarrow$0.27—meaning the model's probability outputs for Japanese became dramatically less confident. This occurs because the shared encoder's representations shift globally when new language data is introduced: the model partially "reallocates" representation capacity to accommodate Farsi, subtly affecting all languages. This finding cautions that *any* data augmentation in multilingual settings requires full re-evaluation of all per-language thresholds, not just the target language.

Key negative findings. (1) Translation augmentation hurt when domains mismatch (tweets$\rightarrow$forums: fr -8.3%). (2) Augmenting only positives into a new language creates catastrophic shortcuts (Farsi v1: 99.9% positive), invisible on dev sets lacking that language. (3) Architecture diversity requires quality parity: mDeBERTa-v3-base diluted the ensemble. (4) GLiNER entity features add zero value atop a fine-tuned classifier.

Key positive findings. (1) Model scale: XLM-R-large gained +15.7% (de), +9.5% (zh) over base for data-scarce languages. (2) Multi-seed ensembling: +1.95% F1 from complementary seed-induced errors. (3) Cross-regime ensembling (6-model, dev 0.7585) improved every language.

7 Conclusion

Our multi-seed XLM-RoBERTa-large ensemble with focal loss and per-language thresholds achieved F1=0.6039, above competition median, across seven languages including zero-shot Farsi. Key lessons: (1) larger shared encoders beat language-specific models, (2) translation augmentation requires domain alignment and class balance, (3) cross-regime ensembling is more effective than architectural diversity, and (4) multilingual augmentation can silently disrupt calibration across all languages.

Limitations

Our approach has several limitations. First, the 256-token truncation affects approximately 5% of German/French forum posts, potentially discarding ADE-relevant context that appears later in long narratives. Second, per-language threshold optimization requires a labeled dev set for each target language, making it inapplicable to truly zero-shot scenarios—our Farsi threshold was a heuristic fallback (mean of trained-language thresholds) rather than optimized. Third, the 3-seed ensemble triples inference cost linearly, which may be prohibitive for real-time pharmacovigilance monitoring. Fourth, our system was evaluated on a single shared task; generalization to other ADE corpora or clinical text remains untested. Finally, we did not explore parameter-efficient fine-tuning (e.g., LoRA) or distillation, which could provide comparable performance at lower computational cost.

References

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics, pages 4171–4186.

Trung Huynh, Yulan He, Alistair Willis, and Stefan Rueger. 2016. Adverse drug reactions discovery from open data: A perspective on social media mining. *ACM Computing Surveys*, 49(4):1–35.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. CADEC: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis*, 32(1):84–100.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Baber, Antonios Baez, Prangthip Bali, and Natascia Battaglia. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2020. The Russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics*, 36(17):4603–4610.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. GLiNER: Generalist model for named entity recognition using bidirectional transformer. *arXiv preprint arXiv:2311.08526*.

A Training Dynamics

Table 4 shows the epoch-by-epoch training progression for the XLM-R-large seed 42 model. Training loss decreases monotonically, while dev F1 improves steadily from 0.702 to 0.731. The cosine schedule reduces LR to zero by epoch 5, providing a natural convergence criterion without early stopping.

Ep.	Loss	LR	Dev F1	Best?
1	0.0331	9.7e-6	0.702	
2	0.0145	7.5e-6	0.726	
3	0.0069	4.1e-6	0.721	
4	0.0030	1.2e-6	0.729	
5	0.0016	0.0	0.731	✓

Table 4: Training progression for XLM-R-large seed 42. Loss is training focal loss; LR is the cosine-decayed learning rate.

For seed 123, dev F1 peaked at epoch 3 (0.738) then slightly decreased (0.735, 0.733 at epochs 4–5), while seed 456 peaked at epoch 2 (0.728) with minor fluctuations thereafter. This variation in optimal epochs motivated our approach of training all seeds for the full 5 epochs and selecting the best checkpoint per seed, rather than applying a uniform early stopping criterion.

B Seed Ablation

To justify the 3-seed choice, we examined performance of subsets:

Configuration	Dev F1
Best single seed (s123)	0.738
2-seed (s42 + s123)	0.745
2-seed (s123 + s456)	0.743
3-seed (s42 + s123 + s456)	**0.7505**

Table 5: Seed ablation: diminishing returns from 2→3 seeds (+0.5–0.8%), but still meaningful given the tight competition margins.