IITPatna_ADE at #SMM4H-HeaRD 2026: Multilingual Adverse Drug Event Detection with LoRA-XLM-RoBERTa, Cross-Fold Ensembles, and Post-hoc Calibration

Sofia Jamil[1] **Manish Singh**[2] **Harshal Dharpure**[3] **Sriparna Saha**[4] **Rajiv Misra**[5]

[5]Department of Computer Science and Engineering
Indian Institute of Technology Patna, India

{sofia_2321cs16, manish_2511ai52, harshal_2511ai30, sriparna , rajivm}@iitp.ac.in

Abstract

We describe our submission to Task 1 of #SMM4H-HeaRD 2026: multilingual binary classification of adverse drug event (ADE) mentions in social media. Our system fine-tunes `xlm-roberta-large` with LoRA adapters and learned language embeddings, using two-stage training (CADEC translated domain adaptation, then five-fold cross-validation on the official training set). We ensemble the five fold checkpoints by mean logits, apply temperature scaling on the development set, and tune decision thresholds to maximize the official metric. On development, the final ensemble reaches macro-F_1 0.788 with a global threshold and 0.796 with per-language thresholds; our best official test submission achieves macro-F_1 0.616 (ID 678990).

1 Introduction

Task definition. Task 1 of #SMM4H-HeaRD 2026 frames **adverse drug event (ADE) mention detection** as a **binary classification** problem: for each social-media post, predict whether an adverse drug event is mentioned. The official leaderboard metric is an **unweighted macro-average** of per-language F_1 scores:

$$F_1^{\text{macro}} = \frac{1}{L} \sum_{\ell=1}^{L} F_\ell, \tag{1}$$

where L is the number of distinct languages in the evaluation set and F_ℓ is the binary F_1 on posts belonging to language ℓ. Every language contributes equally to the aggregate, regardless of post volume, making optimisation of accuracy or micro-F_1 a misaligned objective.

Data characteristics. Six languages appear in the labeled train/dev splits (de, en, fr, ja, ru, zh), while the blind test additionally evaluates Farsi (fa) and CADEC-tagged variants (Lopez-Garcia et al., 2026).

Language	F_1 (global th.)	F_1 (per-lang th.)	Δ
de	0.8000	0.8333	+0.033
en	0.7656	0.7656	0.000
fr	0.9831	0.9836	+0.001
ja	0.5963	0.6012	+0.005
ru	0.6983	0.6983	0.000
zh	0.8861	0.8916	+0.006
Macro	**0.7882**	**0.7956**	+0.007

Table 1: Development F_1 by language with a global threshold vs per-language thresholds.

Submission ID	Macro-F_1 (test)
678990	**0.616**
672993	0.3985

Table 2: Official blind test macro-F_1 reported by organizers.

Approach. We focus on a reproducible, metric-aligned pipeline: **low-rank adaptation (LoRA)** fine-tuning of **XLM-RoBERTa (XLM-R)** (`xlm-roberta-large`) with language embeddings, five-fold checkpoint ensembling, and post-hoc calibration (temperature and thresholds) tuned directly for macro-F_1.

2 Methodology

2.1 Evaluation Metric

For L languages, per-language binary precision P_ℓ, recall R_ℓ, and F_ℓ are computed over all development or test posts carrying language tag ℓ. The official metric is the unweighted macro-average in Eq. (1). We tune all temperatures and thresholds on the development set to maximise this quantity directly.

2.2 Architecture and LoRA Adaptation

The backbone `xlm-roberta-large` processes each input sequence and returns the hidden state $\mathbf{h} \in \mathbb{R}^{1024}$ at the first subword position (i.e., the

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 177–181

Language (test)	F_1
English (en)	0.7278
German (de)	0.6556
French (fr)	0.9239
Japanese (ja)	0.4967
Russian (ru)	0.5571
Chinese (zh)	0.8034
Farsi (fa)	0.4116
German CADEC (de_cadec)	0.8515
French CADEC (fr_cadec)	0.8269
Official macro-F_1 (Codalab leaderboard)	**0.6160**
Arithmetic mean of listed F_1 (9 languages)	0.695

Table 3: Official test F_1 by language for submission 678990 (organizer email). The **leaderboard macro-F_1** (0.616) is authoritative for ranking. The **arithmetic mean** of the nine per-language F_1 values is ≈0.695 under Eq. (1) as stated in this paper; these two quantities therefore *do not* match, and we report both explicitly pending organizer confirmation of the exact test evaluation script (language inventory, rounding, or aggregation).

`[CLS]`-equivalent token). A learned **language embedding** $\mathbf{e}_\ell \in \mathbb{R}^{64}$ (projected to 1024 dimensions) is added to $\mathbf{h}$ to make language identity explicitly available to the classifier. An unknown-language `unk` row handles test-time language codes absent from training. Classifier dropout (0.1) is applied before a single linear layer that produces logits $\mathbf{z} \in \mathbb{R}^2$. LoRA (Hu et al., 2022) attaches trainable low-rank adapters to the `query` and `value` projection matrices of every transformer attention layer. With frozen pre-trained weights $W \in \mathbb{R}^{d\times k}$ and low-rank factors $A \in \mathbb{R}^{r\times k}$, $B \in \mathbb{R}^{d\times r}$, the effective weight becomes:

$$W_{\text{eff}} = W + \frac{\alpha}{r} BA, \tag{2}$$

where rank r=16, scaling α=32, and LoRA dropout is 0.05.

2.3 Input Formatting

We prepend metadata tokens encoding language and source information before the input text.

2.4 Training Objectives

The primary classification loss is **focal loss** (Lin et al., 2017):

$$\mathcal{L}_{\text{focal}} = -\big(1 - p_y\big)^{\gamma} \log p_y, \tag{3}$$

where p_y is the model's predicted probability for the correct class and γ=2.0. Two auxiliary terms are added:

Supervised contrastive loss (Khosla et al., 2020):

$$\mathcal{L}_{\text{ctr}} = -\sum_i \frac{1}{|\mathcal{P}(i)|} \sum_{j\in\mathcal{P}(i)} \log \frac{e^{\mathbf{h}_i\cdot\mathbf{h}_j/\tau}}{\sum_{k\neq i} e^{\mathbf{h}_i\cdot\mathbf{h}_k/\tau}}, \tag{4}$$

where $\mathcal{P}(i)$ is the set of in-batch positives sharing the same label as example i and τ is a temperature hyperparameter. This loss pushes same-class representations together and separates different-class ones.

R-Drop consistency loss (Liang et al., 2021):

$$\mathcal{L}_{\text{kl}} = \frac{1}{2}\Big[\mathrm{KL}\big(p^{(1)} \,\|\, p^{(2)}\big) + \mathrm{KL}\big(p^{(2)} \,\|\, p^{(1)}\big)\Big], \tag{5}$$

where $p^{(1)}$ and $p^{(2)}$ are the output distributions from two independent stochastic forward passes (different dropout masks) of the same batch. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda_{\text{ctr}}\,\mathcal{L}_{\text{ctr}} + \lambda_{\text{kl}}\,\mathcal{L}_{\text{kl}}, \tag{6}$$

with small positive weights λ_{ctr} and λ_{kl} set to implementation defaults in our codebase.

2.5 Two-Stage Training Procedure

Stage 1 – Domain adaptation. We pre-train the model on a translated version of the CADEC corpus (Karimi et al., 2015), a dataset of annotated patient forum posts about adverse drug reactions. Translation to each target language allows the model to adapt to informal medical vocabulary before seeing the task-specific labels, reducing domain shift between the clinical register of CADEC and the informal register of social media. Training runs for at most 3 epochs with early stopping.

Stage 2 – Task fine-tuning with cross-validation. We run K=5-fold stratified cross-validation on the official training CSV, stratifying by (label, language) pair whenever each stratum contains at least two samples. For each fold, the model is initialised from the Stage 1 checkpoint and fine-tuned on the training partition. After each epoch, macro-F_1^{macro} is evaluated on the *full* development set (not the held-out fold), and the checkpoint with the best development macro-F_1^{macro} is saved. This gives five independent best checkpoints $\theta^{(0)}, \ldots, \theta^{(4)}$.

Optimiser: AdamW with learning rate 2×10^{-5}, weight decay 0.01, cosine schedule with warmup ratio 0.06, sequence length 256, batch size 8, gradient accumulation steps 2, mixed-precision (fp16), and early stopping with patience 3 on the development macro-F_1^{macro}.

2.6 Ensemble Inference and Post-hoc Calibration

Logit averaging: For each test example, all five fold models produce logits $\mathbf{z}^{(k)} \in \mathbb{R}^2$, $k \in \{0, \ldots, 4\}$. The ensemble logit is:

$$\overline{\mathbf{z}} = \frac{1}{K} \sum_{k=0}^{K-1} \mathbf{z}^{(k)}. \tag{7}$$

Averaging in logit space (rather than probability space) has been shown to give slightly better calibrated uncertainty estimates when models have similar architectures (Guo et al., 2017).

Per-language temperature scaling: A per-language scalar temperature $T_\ell > 0$ is fitted on the development predictions to minimise negative log-likelihood:

$$\mathbf{p} = \mathrm{softmax}\left(\frac{\overline{\mathbf{z}}}{T_\ell}\right). \tag{8}$$

$T_\ell > 1$ softens over-confident distributions; $T_\ell < 1$ sharpens under-confident ones. This step is essential when some languages have far fewer positive examples than others, since the model's confidence may be systematically biased.

Threshold selection: After calibration, we sweep thresholds $\tau \in (0, 1)$ on the development set and pick the value maximising macro-F_1^{macro}. We compare a single *global* threshold (used in the Codalab submission) versus *per-language* thresholds (development analysis only, Table 1).

3 Dataset

Training and development: Table 4 summarizes the official development split by language; ADE positives are rare overall.

Blind test set: The blind test set includes the six training languages plus **Farsi (fa)**, **German CADEC (de_cadec)**, and **French CADEC (fr_cadec)**, none of which appear with labels during training. Unknown language codes are mapped to an `unk` language embedding row at inference time.

4 Experimental Setup

We use a standard PyTorch/`transformers` stack with LoRA support. Training runs on a single GPU (A100 40GB); a fold takes a few hours and inference takes minutes. Table 5 lists the main hyperparameters. We iteratively moved from single-fold cross-entropy baselines to the final pipeline by adding structured prompts, focal loss with imbalance handling, auxiliary regularizers (contrastive/R-Drop), and calibrated thresholding; submission 672993 is an earlier configuration, while 678990 is the final full-stack submission.

Lang.	Dev N	#ADE	ADE rate (%)
de	634	35	5.5
en	888	61	6.9
fr	418	30	7.2
ja	3045	72	2.4
ru	2669	272	10.2
zh	379	38	10.0
Total	**8033**	**508**	**6.3**

Table 4: Development set distribution by language in `dev_data_SMM4H_2026_Task_1.csv`.

Setting	Value
LoRA rank r / α / dropout	16 / 32 / 0.05
Language embedding dimension	64
Max sequence length	256
Train batch / grad. accum.	8 / 2
Learning rate / weight decay	2×10^{-5} / 0.01
Focal γ; λ_{ctr} / λ_{kl}	2.0; 0 / 0 (submission)
Random seed; CV folds K	42; 5
Stage 1 / Stage 2 epochs (cap)	3 / 10
Dev global threshold (re-run)	0.95
Dev per-language T_ℓ (re-run)	de 0.37, en 0.62, fr 0.26, ja 0.49, ru 0.70, zh 0.47

Table 5: Hyperparameters and artifact pointers for submission 678990 (`smm4h_ade/config.py` defaults unless noted). Exact threshold and temperature scalars are exported by `evaluate.py`, not hard-coded.

5 Results and Discussion

Cross-Validation Results (Diagnostic)

The variance across folds (0.662–0.748) reflects the sensitivity of individual checkpoints to training data partition, particularly for low-resource language strata. Fold 1 and Fold 3 achieve the highest individual development scores; Folds 2 and 4 are weaker in isolation but still contribute to ensemble diversity.

Development: Language-wise Metrics (Final Ensemble)

We report language-wise metrics and confusion counts for the final ensemble in Tables 7 and 8.

French achieves near-perfect precision (1.00) with only one false negative, consistent with the comparatively clean textual style of French social posts in this corpus. Japanese has the lowest F_1 (0.596) despite the largest support, reflecting its

Fold k	Best dev macro-F_1 (single checkpoint)
0	0.717
1	0.748
2	0.668
3	0.734
4	0.662
Mean	0.706

Table 6: Per-fold checkpoint selection scores on development (from `final_results.json`).

Language	Precision	Recall	F_1	Support n
de	0.8667	0.7429	0.8000	634
en	0.7313	0.8033	0.7656	888
fr	1.0000	0.9667	0.9831	418
ja	0.5393	0.6667	0.5963	3045
ru	0.7216	0.6765	0.6983	2669
zh	0.8537	0.9211	0.8861	379
Macro (unweighted)	0.7854	0.7962	**0.7882**	8033

Table 7: Development metrics for the final 5-fold XLM-R ensemble with calibration and a global threshold (from `dev_eval_official_metric.json`).

extreme imbalance (2.4% ADE rate) and the challenges of tokenisation and domain transfer. Russian's false-negative count (88) is the highest in absolute terms, suggesting that the model under-recalls ADEs in Russian.

5.1 Thresholding Analysis

Table 1 compares global and per-language thresholds on the development set. Tuning a separate decision threshold per language improves macro-F_1^{macro} from 0.788 to **0.796**, with the largest gain for German (+0.033). English and Russian show no gain because the calibrated probabilities for those languages happen to be well-aligned with the global optimum. These per-language thresholds were computed on the development set only; applying them directly to the test set would constitute overfitting to development label frequencies, so the Codalab submission used the global threshold.

5.2 Discussion

Ensemble gain over single folds: Ensembling five fold checkpoints yields a clear gain over a typical single fold on development (0.706 → 0.788+), indicating reduced variance; calibration/threshold tuning adds a smaller improvement.

Language difficulty: On the blind test, the largest drop is for the unseen language group `fa` (0.412), while CADEC-tagged subsets remain strong (de_cadec 0.852; fr_cadec 0.827).

Language	TN	FP	FN	TP
de	595	4	9	26
en	809	18	12	49
fr	388	0	1	29
ja	2932	41	24	48
ru	2326	71	88	184
zh	335	6	3	35

Table 8: Development confusion matrix counts by language for the final ensemble (from `dev_eval_official_metric.json`).

Development-to-test generalisation gap: The official test macro-F_1^{macro} (0.616) is considerably lower than the development score (0.788). Multiple factors contribute: (i) domain and temporal shift between the development and test corpora; (ii) three previously unseen language-level groups (fa, de_cadec, fr_cadec) whose performance drags down the macro average; and (iii) the use of a global decision threshold on the test run, whereas per-language thresholds (+0.007) could not be safely applied without development-like labels on test.

6 Conclusion

We presented IITPatna_ADE's system for #SMM4H-HeaRD 2026 Task 1, based on LoRA-adapted `xlm-roberta-large` with learned language embeddings, two-stage training (CADEC pre-training followed by five-fold stratified cross-validation), mean-logit ensemble inference, and per-language temperature scaling with threshold tuning. We carefully distinguished *diagnostic* fold-level scores (mean dev 0.706) from the *final* ensemble development metrics (0.788 global / 0.796 per-language), and reported the official

Limitations

All calibration is performed on a single development split with multiple dependent tuning steps (§2.6); no statistical significance tests are available on the blind test due to the absence of per-example labels. The `Reason` field is empty for the majority of examples in the training and development CSVs, limiting the utility of the reasoning-style input format. Finally, our system is tuned exclusively for macro-F_1^{macro} and may not generalise to other evaluation metrics such as micro-F_1 or per-class AUC.

References

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

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 *International Conference on Learning Representations*.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Melissa Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. In *Proceedings of the 4th Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33.

Xiaobo Liang, Lijun Wu, Juntao Li, Qi Wang, Qin Zhang, Jun Zhou, and Ming Liu. 2021. R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34.

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 (ICCV)*.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Ganesh Chandrasekar, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z. Klein, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, Abeed Sarker, and 8 others. 2026. Overview of the 11th social media mining for health (#smm4h) and health real-world data (HeaRD) shared tasks at ACL 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

A Acknowledgement

The authors gratefully acknowledge the financial support provided by the Prime Minister's Research Fellowship (PMRF), Government of India. The authors additionally thank the Aryabhatta Supercomputing Centre (ASC) at the Indian Institute of Technology Patna for providing the computational resources utilized in this work.