Team TIET at #SMM4H-HeaRD 2026: Fine-tuned Biomedical Transformers with Language-Balanced Sampling for Patient Metadata and Multilingual ADE Detection

Divrose Kaur **Jatin Bedi** **Jasmeet Singh**
Thapar Institute of Engineering & Technology
divrose19@gmail.com jatin.bedi@thapar.edu jasmeet.singh@thapar.edu

Abstract

We present Team TIET's systems for two shared tasks at #SMM4H-HeaRD 2026: Task 5 (detection of patient metadata in SARS-CoV-2 sequencing papers) and Task 1 (multilingual adverse drug event detection across six languages plus an unseen Farsi subset). For Task 5 we explore iterative LLM prompting followed by fine-tuning `BiomedBERT-base` with weighted cross-entropy loss and probability threshold optimization, achieving F1 = 0.760 on the official test set (above the competition mean of 0.729). For Task 1 we fine-tune `XLM-RoBERTa-base` with a combined language- and class-balanced sampling strategy and per-language threshold tuning, achieving macro F1 = 0.497 overall (0.608 excluding the unseen Farsi subset). We report empirical findings on BERT+LLM ensemble failure with bimodal probability distributions, the superiority of base over large model variants under limited data, and the importance of language-balanced gradient contribution in multilingual classification.

1 Introduction

Social media and biomedical literature provide complementary windows onto public health. The #SMM4H-HeaRD 2026 shared tasks challenge systems to mine these sources for clinically relevant signals (Lopez-Garcia et al., 2026). We participated in two tasks: **Task 5**, binary classification of sentences from SARS-CoV-2 sequencing articles for patient metadata; and **Task 1**, binary classification of social media posts in six languages for adverse drug event (ADE) mentions evaluated by macro-averaged F1. Both tasks exhibit severe class imbalance (2.4–22.5% positive rate) and require domain-adapted representations. All experiments were conducted on commodity hardware (standard laptop, 16 GB RAM, no dedicated GPU) with free-tier cloud compute (Google Colab T4/H100, Groq API).

2 Task 5: Patient Metadata Detection

2.1 Task and Data

Given a sentence from a PubMed SARS-CoV-2 sequencing article, the system predicts whether it reports patient metadata (age, sex, symptoms, disease severity, lab results, treatments, outcomes, etc.) *directly associated with sequences generated in that specific study*. Mentions of prior-study metadata, mere collection descriptions, and epidemiological analyses are negative examples. The dataset contains 15,504 training / 2,214 validation / 4,429 test sentences at 13.3% positive rate. Organizer baselines: BiomedBERT-Large fine-tuned (F1 = 0.776) and Llama-3-70B zero-shot (F1 = 0.558).

2.2 Stage 1: LLM Prompting

We first evaluated `llama-3.3-70b-versatile` via the Groq API; however, at ≈900 tokens per request the 100k daily free-tier limit covered only 68 samples—insufficient for dataset-scale evaluation. This motivated switching to a locally-hosted 7B model (`qwen2.5:7b` via Ollama, ≈13 s/sentence on CPU), enabling full validation set evaluation.

We iterated four prompt versions, alternating between over-permissive and over-strict rules. Each iteration was evaluated on the full 2,214-sentence validation set; errors were manually annotated to identify systematic failure patterns. After analysis of 135 errors (92 FPs + 43 FNs) from the third iteration, we derived a 7-step decision tree encoding the critical distinction: *stated values* ("median age 30.5 years" → positive) vs. *analytical relationships* ("older patients had higher viral loads" → negative). Our final prompt iteration, incorporating this decision tree, reached F1 = 0.59 on the full validation set. As shown in Table 1, larger cloud models did not consistently exceed this, suggesting the ceiling may reflect single-sentence context limits rather than model scale alone, though we note

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 268–271

Model	F1	P	R
qwen2.5:7b local (decision-tree)	0.59	0.63	0.56
llama-4-scout-17b (Groq)	0.53	0.53	0.54
qwen3-32b thinking (Groq)	0.34	0.53	0.25

Table 1: Task 5 LLM prompting results. Local model (qwen2.5:7b) evaluated on full validation set; cloud models on a fixed 200-sample hard subset. Prompts used a 7-step decision tree distinguishing stated values from analytical relationships.

this inference is limited to the models and dataset tested here.

2.3 Stage 2: Fine-tuned BiomedBERT

We fine-tuned `BiomedBERT-base` (Gu et al., 2022) with weighted cross-entropy (positive class weight $= 6.5\times$). Configuration: MAX_LEN=256, batch=32, LR=2e-5, 4 epochs, linear warmup (10%), AdamW, weight decay=0.01, gradient clipping=1.0; training $\approx$25 min on a T4 GPU.

Threshold tuning. BERT's output probability distribution was highly bimodal: true positives clustered near 0.997 and true negatives near 0.0004, with very few examples in between. We swept $\tau \in [0.1, 0.9]$ in increments of 0.01 on the validation set; τ=0.88 was optimal (val F1 = 0.800 vs. 0.788 at default τ=0.5). The bimodal distribution means threshold choice has large impact: small deviations from 0.88 cause sharp drops in validation F1.

BiomedBERT-Large. Memory constraints on T4 required reducing MAX_LEN to 128 and batch to 8. The base model outperformed Large (val F1 = 0.800 vs. 0.797): truncated sequences lose information and the larger model suffers relative data starvation given the limited training set size.

BERT+LLM ensemble. We swept ensemble weights combining BERT probabilities with the high-recall prompt variant (F1 = 0.51, recall = 0.68). The optimum was `llm_weight=0.00`: BERT's bimodal probability distribution leaves no ambiguous region for the LLM to correct, meaning the LLM's uncertain predictions added noise rather than complementary signal. LLM prompting served as an exploratory stage to understand the task rather than as a component of the final system.

2.4 Task 5 Results

Table 2 compares our system against the organizer baselines.

System	F1	P	R
Org. zero-shot (Llama-3-70B)	0.558	—	—
Ours: qwen2.5:7b (decision-tree)	0.590	0.630	0.560
Org. fine-tuned (BERT-Large)	0.776	—	—
Ours: BERT-base (val, τ=0.88)	0.800	0.776	0.827
Ours: BERT-base (test)	**0.760**	0.770	0.750

Table 2: Task 5 final results. Competition mean: 0.729; median: 0.754. Our test F1 exceeds both (mean +0.031, median +0.006). Precision and recall for organizer baselines were not reported.

3 Task 1: Multilingual ADE Detection

3.1 Task and Data

Given a social media post in one of six languages, predict whether it contains an ADE mention. The metric is macro-averaged F1 (equal weight per language regardless of dataset size). Class imbalance ranges from 2.4% (Japanese) to 22.5% (French). Supplementary CADEC drug review data (machine-translated to German and French) was provided but excluded from the official metric. The test set contained **Farsi** (15,184 samples) absent from training data, unknown to participants until submission.

3.2 System: XLM-RoBERTa with Language-Balanced Sampling

We fine-tuned `xlm-roberta-base` (Conneau et al., 2020) as a single shared model across all six languages.

Language-balanced sampling. Standard class-weighted sampling lets high-resource languages dominate gradient updates proportional to dataset size (English: 17k samples, German: 1.9k — a 9$\times$ imbalance per epoch). We apply a combined sampler:

$$w_i = \lambda_{\text{lang}(i)} \times \lambda_{\text{class}(i)} \quad (1)$$

where λ_{lang} equalizes each language to $\approx$7,924 draws per epoch (total training samples divided by number of languages), and λ_{class} handles within-language imbalance. French F1 improved by +0.121 on the development set compared to class-only weighting, and overall macro F1 improved by +0.016. The effect is largest for languages dominated by higher-resource counterparts during standard sampling.

Per-language threshold tuning. We swept $\tau \in [0.05, 0.95]$ per language on dev sets (Table 3). The

Lang	τ	F1@0.5	F1@τ	Δ
en	0.86	0.756	0.761	+0.006
de	0.10	0.609	0.615	+0.007
fr	0.70	0.732	0.732	+0.000
ru	0.86	0.678	0.673	−0.005
zh	0.30	0.868	0.868	+0.000
ja	0.68	0.566	0.566	+0.000

Table 3: Per-language threshold tuning on dev sets.

final model was generally well-calibrated: most languages showed negligible gain from threshold shifting ($\Delta \leq 0.007$), and Russian was slightly hurt ($\Delta = -0.005$). The notable exception was German (τ=0.10, Δ= + 0.007): the model learns ADE patterns but under-estimates confidence for long forum posts, requiring a low threshold to recover recall. Chinese (τ=0.30) required downward pressure, consistent with over-confidence on structured Q&A text.

XLM-RoBERTa-Large. On an H100 GPU (Colab Pro), XLM-RoBERTa-Large (559M parameters) reached macro F1 = 0.687 after correcting a data leakage bug—worse than base (0.703). The underperformance is explained by data starvation: the larger model has more parameters to populate but low-resource languages (German: 1.9k, Japanese: 2.1k) do not provide sufficient gradient signal, causing German F1 to collapse from 0.615 to 0.484. This suggests a parameter-to-data ratio threshold below which upsizing hurts rather than helps in multilingual settings with unequal resource distribution.

Per-language expert models. Separate per-language models augmented with English→{German, French} translation via MarianMT (Junczys-Dowmunt et al., 2018) (Japanese translation discarded due to output quality failure) achieved macro F1 = 0.559 vs. 0.703 for the shared model. Cross-lingual transfer outweighs language-specific tuning when per-language data is scarce (<5k samples).

3.3 Task 1 Results

Our de_cadec (0.879) and fr_cadec (0.897) both exceed competition means (0.833 and 0.843 respectively). The overall score is heavily depressed by zero-shot Farsi (F1 = 0.257); the 6-language macro F1 of 0.608 better reflects performance on seen languages. Potential strategies for improving unseen-language robustness include multilingual data augmentation at training time and language-agnostic feature extraction; we leave these to future work.

Language	Dev F1	Test F1
English (en)	0.761	0.649
German (de)	0.615	0.602
French (fr)	0.732	0.631
Russian (ru)	0.673	0.488
Chinese (zh)	0.868	0.759
Japanese (ja)	0.567	0.520
Farsi (fa)†	—	0.257
de_cadec	—	0.879
fr_cadec	—	0.897
Overall macro	**0.703**	**0.497**
6-lang (excl. Farsi)	0.703	0.608

Table 4: Task 1 per-language results. †Farsi absent from all training data (zero-shot transfer only). Competition mean: 0.547; median: 0.580; de_cadec mean: 0.833; fr_cadec mean: 0.843.

4 Analysis

Prompting ceiling. Single-sentence classification with a 7B model plateaus near F1 ≈ 0.59 regardless of prompt sophistication, and larger cloud models did not consistently exceed this in our experiments. One possible explanation is that single-sentence context is insufficient to resolve ambiguous cases that require surrounding document context; however, we note this is based on a limited set of models and a single task, and further investigation is needed to confirm this hypothesis.

Base vs. large under data constraints. In both tasks, the base model outperformed the large variant. For Task 5, reduced MAX_LEN for Large discards information from longer sentences. For Task 1, data starvation in low-resource languages causes performance collapse in the larger model, particularly for German (0.615 base vs. 0.484 large). A shared model architecture that equalizes per-language data exposure appears more robust than simply scaling model capacity.

Language-balanced sampling. Equalizing gradient contributions across languages improved overall macro F1 by 0.016 and French F1 by 0.121. The effect is largest for languages dominated by higher-resource counterparts during standard sampling, confirming that the bottleneck for low-resource languages in a shared multilingual model is gradient competition rather than representational capacity.

BERT+LLM ensembles. When one model has a highly bimodal probability distribution, combining it with a weaker model introduces noise. Practitioners should verify probability calibration before ensemble combinations.

5 Conclusion

We presented systems for Tasks 5 and 1 of #SMM4H-HeaRD 2026. BiomedBERT-base with threshold tuning achieved F1 = 0.760 on Task 5 (above competition mean), outperforming the organizer's larger baseline using the base-sized model. XLM-RoBERTa-base with language-balanced sampling achieved macro F1 = 0.497 overall in Task 1 (0.608 on the 6 seen languages), with performance substantially reduced by zero-shot Farsi. Our results highlight the value of calibrated probability thresholds, language-balanced gradient weighting, and caution against model upsizing and LLM prompting reliance under limited data.

Limitations

Our Task 1 system was unaware of the Farsi test subset until submission; future work should include zero-shot language robustness testing. LLM prompting experiments were limited by free-tier API quotas, preventing full-dataset evaluation of larger models. The language-balanced sampler hyperparameter was set heuristically; principled tuning may yield further improvements. BiomedBERT experiments used a single random seed. The prompting ceiling observation is based on a small set of models tested on one dataset and one task; broader conclusions would require more systematic evaluation.

Ethical Considerations

Task 5 involves sentences from PubMed articles describing patient cohorts in SARS-CoV-2 sequencing studies. All data used is drawn from publicly available published literature; no new patient data was collected or accessed. Task 1 involves social media posts that may contain personal health disclosures. The dataset was provided by the shared task organizers under their data use agreement; we did not access any user-identifiable information and performed no re-identification. ADE detection systems of the type described here carry potential for both benefit (pharmacovigilance, drug safety monitoring) and harm (false positives stigmatizing drug use, false negatives missing safety signals). Deployment in clinical or regulatory contexts would require substantially more rigorous evaluation than reported here.

Acknowledgments

This work was conducted as part of undergraduate research at Thapar Institute of Engineering & Technology. The authors thank the #SMM4H-HeaRD 2026 organizers for providing the shared task datasets and evaluation infrastructure.

References

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

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.