Vinland_Vector at #SMM4H-HeaRD 2026: Multilingual ADE Detection and Query-Augmented Clinical NER for English

Nirjhar Das, Rathijit Aich, Mahfuzulhoq Chowdhury
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
nirjhardasami@gmail.com, aichrathijit@gmail.com,
mahfuz@cuet.ac.bd

Abstract

In this paper, we address Task 1 on adverse drug event (ADE) detection and Task 8 on MultiClinNER at SMM4H-HeaRD 2026. ADE detection is formulated as a multilingual binary classification problem over social media posts spanning German, French, Russian, English, Mandarin and Japanese, with zero-shot on Farsi. Using XLM-RoBERTa-Large with a dual-pooling head, combined with stratified sampling, language-conditioned inputs, translation-based augmentation, and calibrated ensembling, our model achieves a macro F1 score of 0.6088, surpassing both the competition mean (0.5465) and median (0.5798). Our work in MultiClinNER targets clinical NER for English text. Using GLiNER-large with sliding-window inference, query augmentation, and calibrated thresholds, it achieves strict F1 scores of 0.7591 (Disease), 0.7263 (Procedure), and 0.6733 (Symptom), outperforming a PubMedBERT baseline across all entities.

1 Introduction

Social media has become a major source of patient-reported health information, including discussions of adverse drug events (ADEs) and clinical conditions. However, such data are largely unstructured and noisy, limiting their utility (Sarker and et al., 2015).

In this work, we present our systems for two shared task at SMM4H-HeaRD 2026: ADE detection (Task 1) and clinical Named Entity Recognition (Task 8) under the MultiClinAI shared task (Gallego-Donoso et al., 2026) and the SMM4H-HeaRD evaluation framework (Lopez-Garcia et al., 2026). ADE detection focuses on multilingual binary classification of social media posts across six languages, with additional zero-shot evaluation on Farsi, while NER targets entity extraction from English clinical text, a setting characterized by domain-specific terminology and high linguistic variability (Yadav and Bethard, 2018).

For ADE detection, we employ XLM-RoBERTa-Large with data-centric strategies to address class imbalance, multilingual variability, and calibration. For clinical NER, we adopt a GLiNER-based approach with enhancements for long-context inference and entity-level optimization.

Our core contribution is to show that data-centric optimization governs multilingual classification performance, while model formulation governs structured extraction, revealing complementary failure modes in clinical NLP systems.

2 Related Works

Multilingual Transformer models such as XLM-RoBERTa (Conneau and et al., 2020) and mDeBERTa-v3 (He and et al., 2021) have become the standard cross-lingual NLP, demonstrating strong zero-shot and few-shot transfer. In ADE detection prior SMM4H shared tasks (Weissenbacher and et al., 2019; Magge and et al., 2021) show that fine-tuned transformers outperform traditional methods. Techniques such as Focal Loss (Lin and et al., 2017), R-Drop (Huang et al., 2021), and layer-wise learning rate decay (Sun et al., 2019) have been proposed to improve robustness, while NLLB-200 (NLLB Team, 2022) enables effective zero-shot transfer to low-resource languages like Farsi. In biomedical NER, domain-specific models such as BioBERT (Lee and et al., 2020) and PubMedBERT (Gu and et al., 2020) achieve strong performance by leveraging large-scale clinical corpora. More recently, query-based approaches like GLiNER (Zaratiana and et al., 2024) improve flexibility across entity types but require careful tuning.

3 Dataset

3.1 Task 1: ADE Detection

The dataset consists of training, development, and test splits across multiple language–domain combinations. Table 1 presents the training and develop-

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 211–215

Language	Train	Dev	% Pos
German (de)	1,482	634	Neg. skewed
French (fr)	977	419	Neg. skewed
Russian (ru)	10,754	2,670	Pos. skewed
English (en)	17,974	902	Moderate
Mandarin (zh)	2,248	379	Moderate
Japanese (ja)	14,208	3,045	Moderate

Table 1: Training and development statistics for Task 1

Language	Test	%	In Train?
Farsi (fa)	15,184	34.9%	No (zero-shot)
English (en)	11,712	26.9%	Yes
Russian (ru)	9,293	21.3%	Yes
Japanese (ja)	3,045	7.0%	Yes
Mandarin (zh)	1,144	2.6%	Yes
German (de)	1,105	2.5%	Yes
French (fr)	1,104	2.5%	Yes
German CADEC	87	0.2%	Yes (shift)
French CADEC	87	0.2%	Yes (shift)

Table 2: Test distribution for Task 1

ment data, while Table 2 shows the test distribution. Notably, the test set is highly asymmetric: Farsi accounts for 15,184 of 43,571 examples (34.9%) despite being absent from training and development, constituting a strict zero-shot setting.

3.2 Task 8: Named Entity Recognition

We use the dataset provided by (Lima López et al., 2026). The training data contains 79,316 annotations across three entity types in english language, as summarized in Table 3.

Entity Type	Annotations	TSV Lines
Disease	25,118	25,119
Procedure	26,733	26,734
Symptom	27,465	27,466
Total	79,316	79,319

Table 3: Training statistics of the MultiClinNER English dataset.

The test set contains annotations across the same entity types, as shown in Table 4.

Example: The patient reported **chest pain** (Symptom), underwent **physical examination** (Procedure), and had **hypertension** (Disease).

4 Methodology

Figure 1 shows the overall system architecture of task 1 and task 8.

4.1 Task 1: ADE Detection

Our system is designed around data-centric and multilingual robustness hypotheses depicted in Figure 1.

Entity Type	Test Instances
Disease	5,554
Procedure	5,550
Symptom	5,553
Total	16,657

Table 4: Test set statistics of the MultiClinNER English dataset.

Hyperparameter	Value
Base model	XLM-RoBERTa-Large
Max sequence length	192
Batch size / accumulation	8 / 4 (32)
Epochs	3
Learning rate	2×10^{-5}
LLRD decay	0.9
Weight decay	0.01
Gradient clip	1.0
Focal γ	2.0
R-Drop α	0.5
Dropout heads	3 ($p = 0.1$)
Seeds	[42, 2024]

Table 5: Training configuration for Task 1

4.1.1 Model Architecture

We use XLM-RoBERTa-Large (Conneau and et al., 2020) with a dual-pooling representation, concatenating the [CLS] embedding with mean-pooled token representations. This captures both global semantics and token-level evidence. The pooled representation is passed through a two-layer MLP with LayerNorm and dropout. Multi-sample dropout (Inoue, 2019) is applied during training.

4.1.2 Training Strategy

Table 5 summarizes the training configuration. We train for 3 epochs using AdamW with layer-wise learning rate decay (Sun et al., 2019), Focal Loss (Lin and et al., 2017), and R-Drop (Huang et al., 2021). The embedding layer is frozen for the first epoch. Mixed-precision training and gradient clipping are applied.

Farsi is treated as a zero-shot language. We translate inputs into English using NLLB-200 (NLLB Team, 2022) and perform inference on translated text. We apply fixed per-language thresholds and temperatures. Grid-search calibration overfit to small development sets and was replaced with stable manually tuned values. However, validation-stage observations suggested that translation and threshold tuning provided the most noticeable gains.

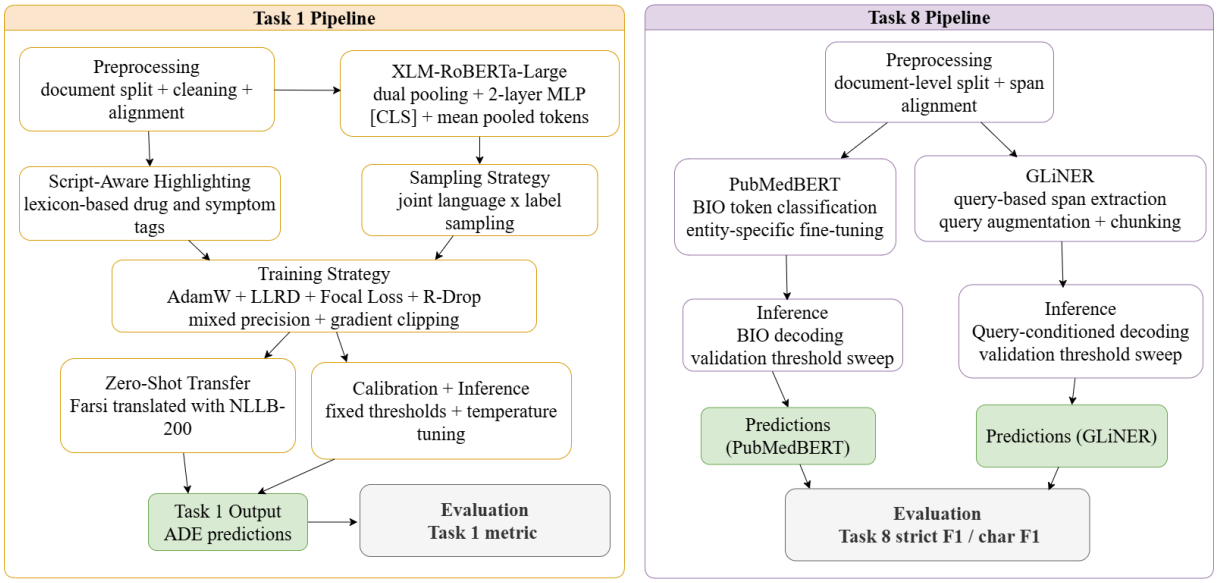

Figure 1: Task 1 and 8 methodology

4.2 Task 8: Named Entity Recognition

4.2.1 PubMedBERT-Based Method

We formulate NER as a token-level BIO tagging task using PubMedBERT (Gu and et al., 2020), training one model per entity type (Disease, Procedure, Symptom). Table 6 shows the training details. Documents are split at the document level (85/15). Long texts are processed using overlapping windows (length 256, stride 64). Tokens are aligned using offsets and labeled with BIO tags.

Parameter	Value
Max sequence length	256
Stride	64
Learning rate	2×10^{-5}
Epochs	4
Batch size	4
Grad accumulation	2
Warmup ratio	0.1
Weight decay	0.01
Seed	42

Table 6: PubMedBERT configuration.

At inference, predictions are decoded into spans, mapped to character offsets, and merged across overlapping windows.

4.2.2 GLiNER-Based Method

We adopt a query-based NER formulation using GLiNER-BioMed (Yazdani et al., 2025), training separate models for each entity type to reduce inter-class ambiguity. Table 7 summarizes the training configuration.

Documents are split into training and validation sets (85/15) and processed using overlapping chunks (Disease: 192/48, others: 128/32). Only entity spans fully contained within chunk boundaries are retained. Query augmentation is applied during training by randomly sampling synonymous labels. For example, **Disease:** disease, disorder, diagnosis, lesion. **Procedure:** procedure, intervention, surgery. **Symptom:** symptom, manifestation, clinical sign. At inference time, a fixed canonical query is used for prediction. Predicted spans are aggregated across chunks via deduplication and filtered using entity-specific confidence thresholds.

Parameter	Value
Max length	192 (Disease) / 128 (Others)
Stride	48 (Disease) / 32 (Others)
Learning rate	5×10^{-6}
Batch size	1
Max steps	3000 (Disease) / 2000 (Others)
Weight decay	0.01
Threshold	tuned
Seed	42

Table 7: GLiNER training and inference configuration.

At inference, spans are predicted, merged across chunks, and filtered using entity-specific thresholds.

5 Results & Analysis

5.1 Task 1: ADE Detection

Table 8 presents the ADE detection results. The final system (RoBERTa + calibration + overfit optimization) achieves a macro F1 of 0.6088, outperforming the competition mean (0.5465) and median (0.5798). Compared to the baseline (0.6056), the

System	en	de	fr	ja	ru	zh	fa	de_c	fr_c	Macro F1
RoBERTa + per-language calibration (grid search)	.7475	.6211	.7090	.5224	.5524	.8354	.4340	.7708	.7677	0.6056
RoBERTa + calibration + overfit optimization	.7205	.6947	.6754	.6258	.5456	.8297	.3770	.7629	.7423	**0.6088**
Competition mean	.6845	.6640	.6814	.5342	.5327	.8044	.3670	.8328	.8430	0.5465
Competition median	.7011	.6559	.6961	.5490	.5504	.8210	.3797	.8598	.8829	0.5798

Table 8: Per-language F1 and macro F1 for ADE detection. de_c = de_cadec, fr_c = fr_cadec (Task 1)

Entity	Model	Threshold	Strict			Char		
			P	R	F1	P	R	F1
Disease	GLiNER	0.55	0.7459	0.7728	**0.7591**	0.8212	0.8453	**0.8331**
	PubMedBERT	0.30	0.7364	0.6898	0.7123	0.8187	0.7668	0.7919
Procedure	GLiNER	0.60	0.7686	0.6884	**0.7263**	0.8507	0.7561	**0.8006**
	PubMedBERT	0.50	0.7019	0.6273	0.6625	0.8088	0.7229	0.7634
Symptom	GLiNER	0.50	0.6457	0.7035	**0.6733**	0.7292	0.7872	**0.7571**
	PubMedBERT	0.30	0.6770	0.6081	0.6407	0.7689	0.6907	0.7277

Table 9: Comparison of GLiNER and PubMedBERT across entity types (Task 8) with calibrated thresholds

gain is modest (+0.0032) but indicates improved generalization under limited and skewed data.

Performance varies across languages. Japanese shows the largest improvement (0.5224 → 0.6258), benefiting from joint sampling and stable calibration. High-resource languages such as Mandarin remain stable (0.8354 → 0.8297), suggesting limited gains with sufficient data. Zero-shot Farsi performance remained comparatively lower, likely due to translation ambiguity, script differences, and limited multilingual clinical vocabulary coverage. Overall, results indicate that improvements are primarily driven by data-centric strategies, rather than architectural changes.

5.2 Task 8: Named Entity Recognition

We evaluate models using strict F1 and character F1 following the official evaluation script.[1] Table 9 summarizes results across entity types. GLiNER consistently outperforms PubMedBERT on both metrics. It achieves higher strict F1 for Disease (0.7591 vs. 0.7123), Procedure (0.7263 vs. 0.6625), and Symptom (0.6733 vs. 0.6407), indicating improved boundary detection and reduced span fragmentation. Character-level scores show a similar pattern, suggesting better partial span alignment as well. Overall, these results highlight the advantage of query-based span extraction over token-level BIO tagging, particularly for handling complex and variable-length clinical entities

[1] https://github.com/nlp4bia-bsc/MultiClinAIEval

6 Conclusion

We propose a multilingual ADE detection system based on XLM-RoBERTa-Large with data-centric enhancements, achieving a macro F1 of 0.6088 and outperforming both the competition mean and median. Results show that sampling and calibration are the primary drivers of cross-lingual robustness, with calibration being particularly critical for generalization. For clinical NER, we compare PubMedBERT and GLiNER, where GLiNER consistently outperforms across all entity types in both strict and character-level F1. Despite PubMedBERT's biomedical specialization, GLiNER achieved stronger performance, potentially because its instruction-style entity representation generalizes better to heterogeneous multilingual clinical expressions.

Overall, our findings highlight the importance of data-centric design and flexible modeling approaches for multilingual and clinical NLP.

Limitations

For ADE detection, pseudo-labeling for Farsi was ineffective due to the absence of in-language supervision and the propagation of noisy labels from translation-based predictions. Aggressive threshold tuning further caused overfitting , reducing generalization. For clinical NER, we use separate models per entity, limiting scalability and cross-entity modeling. GLiNER performance is also sensitive to query design and threshold tuning, with suboptimal choices reducing extraction quality.

References

Alexis Conneau and et al. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. The multiclinai shared task on multilingual clinical corpus construction and concept extraction. In *Proceedings of SMM4H-HeaRD*.

Yu Gu and et al. 2020. Domain-specific language model pretraining for biomedical natural language processing. ArXiv preprint arXiv:2007.15779.

Pengcheng He and et al. 2021. DeBERTaV3: improving deberta using electra-style pre-training. ArXiv preprint arXiv:2111.09543.

Po-Sen Huang, Robert Logan, Mostofa Parekh, and Ming-Wei Chang. 2021. R-Drop: regularized dropout for neural networks. In *NeurIPS*, pages 10946–10959.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. ArXiv preprint arXiv:1905.09788.

Jinhyuk Lee and et al. 2020. BioBERT: a pre-trained biomedical language representation model. *Bioinformatics*, 36(4):1234–1240.

Salvador Lima López, Judith Rosell, Jan Rodríguez Miret, Fernando Gallego-Donoso, and Martin Krallinger. 2026. Multiclinai shared task training data.

Tsung-Yi Lin and et al. 2017. Focal loss for dense object detection. In *ICCV*, pages 2980–2988.

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.

Arjun Magge and et al. 2021. Deepademiner: a pharmacovigilance pipeline for adverse drug event extraction. *JAMIA Open*, 4(1).

NLLB Team. 2022. No language left behind: scaling machine translation. ArXiv preprint arXiv:2207.04672.

Abeed Sarker and et al. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification. In *CCL*, pages 194–206.

Davy Weissenbacher and et al. 2019. Overview of the fourth smm4h shared tasks. In *Proceedings of the SMM4H Workshop*, pages 21–30.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of COLING*, pages 2145–2158.

Anthony Yazdani, Ihor Stepanov, and Douglas Teodoro. 2025. GLiNER-biomed: efficient biomedical named entity recognition. ArXiv preprint arXiv:2504.00676.

Urchade Zaratiana and et al. 2024. GLiNER: generalist model for named entity recognition. In *NAACL*.