DNT at #SMM4H–HeaRD 2026: Leveraging BERT-based Encoders and LLMs for Medical Information Extraction

Doan Nhat Tien, Dang Van Thin
University of Information Technology
Vietnam National University Ho Chi Minh City
22521463@gm.uit.edu.vn, thindv@uit.edu.vn

Abstract

This paper presents our systems for two tasks at #SMM4H-HeaRD 2026. For Task 1 (multilingual Adverse Drug Event detection), we fine-tune BERT-based multilingual models (InfoXLM and XLM-RoBERTa) and Qwen3.5-9B with ensemble methods, achieving 0.8584 macro F1 on the development set and 0.5304 F1 on unseen Farsi. For Task 7 (span detection of ClinicalImpacts and SocialImpacts in opioid narratives), DeBERTa-Large with simplified labeling achieves the best test performance (0.583 relaxed F1, 0.500 strict F1). Our analysis shows that LLMs excel on known languages in Task 1, while transformer-based models with simplified labeling generalize better for NER tasks.

1 Introduction

Social media has become an important source for monitoring health-related events and understanding real-world patient experiences. The #SMM4H shared tasks provide a platform for developing NLP systems to extract insights from user-generated health content (Lopez-Garcia et al., 2026).

We participate in two tasks. Task 1 focuses on multilingual Adverse Drug Event (ADE) detection across six languages, with the added challenge of an unseen Farsi test set for zero-shot cross-lingual evaluation. Task 7 addresses span detection of ClinicalImpacts and SocialImpacts in first-person opioid use narratives (Paul et al., 2016), where such accounts capture real-world consequences often underreported in clinical settings.

Our key findings are: (1) ensembles of diverse multilingual architectures generalize better to unseen scripts than any single model, including a 9B-parameter LLM, underscoring the value of architectural diversity for zero-shot cross-lingual transfer; (2) instruction-tuned LLMs outperform encoder-only models on seen languages in ADE classification, but fail to produce reliable structured outputs for token-level span detection; and (3) simplified labeling (collapsing BIO tags to entity-class labels) can match or exceed standard BIO tagging in NER settings where consecutive same-type spans are rare, suggesting that label-space complexity is a meaningful design choice in low-boundary-ambiguity datasets.

2 Task Description

2.1 Task 1: Multilingual ADE Detection

Task 1 is a binary classification problem: given a user-generated post, the system should predict whether it contains a mention of an ADE. The dataset includes six languages (German, French, Russian, English, Mandarin, Japanese) and is imbalanced with the positive class (ADE mentions) representing a small fraction of the data. Evaluation is based on unweighted macro F1-score.

A key challenge is the inclusion of Farsi in the test set, a language not present in the training data, which tests the models' zero-shot cross-lingual generalization capabilities.

2.2 Task 7: Span Detection of Clinical and Social Impacts

Task 7 is framed as a named entity recognition problem where the goal is to identify and extract text spans corresponding to two entity types:

- **ClinicalImpacts**: Health-related consequences of opioid use, including physical symptoms (e.g., withdrawal, overdose, seizures), mental health effects (e.g., depression, anxiety), and general health deterioration.
- **SocialImpacts**: Societal and interpersonal consequences, including employment issues (e.g., job loss), legal problems (e.g., arrests, charges), relationship impacts, and other social ramifications.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 36–40

The task requires systems to output predictions in BIO format, where each token is tagged as either 'O' (outside), 'B-EntityType' (beginning of an entity), or 'I-EntityType' (inside an entity). Evaluation uses both strict F1 (exact span and type match) and relaxed F1 (token-level overlap) metrics (Dey et al., 2025).

3 System Description

3.1 Task 1 Systems

3.1.1 Data Preprocessing

We combine training and validation data from SMM4H 2026 Task 1 and CADEC translated datasets (Karimi et al., 2015). To address class imbalance, we upsample the positive class by 3x. The data is split using stratified sampling based on both label and language (15% validation), ensuring that each language and class is proportionally represented in the training and validation sets.

3.1.2 BERT-based Models

We fine-tune two multilingual BERT-based encoders selected based on their strong performance in cross-lingual benchmarks:

InfoXLM: Microsoft/infoxlm-large (570M parameters), a cross-lingual model pre-trained on 100 languages (Chi et al., 2021). Fine-tuning hyperparameters: learning rate 5×10^{-5}, batch size 16, gradient accumulation steps 4, 5 epochs, max sequence length 256.

XLM-RoBERTa: FacebookAI/xlm-roberta-large (570M parameters), a multilingual version of RoBERTa (Conneau et al., 2020). Same fine-tuning configuration as InfoXLM.

For confidence score extraction, BERT-based models output softmax probabilities directly from the final classification layer.

3.1.3 Large Language Model

We use **Qwen3.5-9B-Base**, a large language model with 9 billion parameters (Bai et al., 2023). While smaller LLMs (1B–4B parameters) are more commonly adopted for text classification tasks, we deliberately opt for a larger model to fully leverage the available hardware resources and maximize task performance. This model was further selected based on our prior positive experience with the Qwen model family in related NLP tasks. To adapt the model for classification, we employ instruction-tuning with LoRA (Low-Rank Adaptation) (Hu et al., 2022) to efficiently fine-tune the model (see Appendix for the prompt templates).

LoRA configuration: $r = 8$, $\alpha = 32$, dropout 0.1. Training: learning rate 5×10^{-5}, batch size 32, gradient accumulation steps 2, 5 epochs, bfloat16 precision.

For confidence scores, we extract logits from the first generated token using `output_scores=True` and `return_dict_in_generate=True`, select logits for tokens "0" and "1", apply softmax, and use the probability of the predicted label as the confidence score.

3.1.4 Ensemble Methods

We investigate three ensemble strategies combining predictions from all three models:

Hard Voting: Majority vote across models (label $= 1$ if ≥ 2 models predict 1).

Soft Voting: Averages confidence-weighted probabilities. For each model, we convert $(label, confidence)$ to probability of class 1: $P(1|model) = confidence$ if $label = 1$, else $1 - confidence$. The final probability: $P(1) = \frac{1}{3}\sum_m P(1|m)$, with threshold 0.5 for binary decision.

Weighted Soft Voting: Similar to soft voting but uses development set F1-scores as weights (InfoXLM: 0.8370, XLM-R: 0.8360, Qwen: 0.8543).

All three ensemble methods achieve similar macro F1-scores on the development set (0.8583–0.8584). We select soft voting for test set submission to leverage confidence scores for uncertainty quantification.

3.2 Task 7 Systems

3.2.1 Approach 1: Instruction Fine-tuning with Qwen3.5-9B-Base

Following the same rationale as in Task 1, we continue to use this Qwen model, adapting it via LoRA ($r = 8$, $\alpha = 32$, dropout 0.1) with a custom prompt template that outputs enumerated tokens to avoid ambiguity. Training: max sequence length 768, batch size 8 with gradient accumulation of 2, learning rate 5×10^{-5}, 5 epochs, bfloat16 precision.

The prompt requires the model to output the exact index of each token with its label, formatted as `<Index> : <Token> : <Label>`, which is critical for handling cases where identical tokens appear multiple times in the same text (see Appendix for the prompt templates).

3.2.2 Approach 2: DeBERTa-Large with BIO Labeling

Following the baseline approach in (Dey et al., 2025), we adopt DeBERTa-Large (24 layers, 1024 hidden dimensions, 16 attention heads, 355M parameters) (He et al., 2021) for token classification with standard BIO tagging. The label set includes `O`, `B-ClinicalImpacts`, `I-ClinicalImpacts`, `B-SocialImpacts`, and `I-SocialImpacts` (Ramshaw and Marcus, 1995).

For token alignment, special tokens receive label -100, the first subword of each input word receives the original word label, and subsequent subwords receive an 'I-' label if the original label started with 'B-', otherwise the same label.

Training: max sequence length 256, batch size 8 with gradient accumulation of 2, learning rate 5×10^{-5}, weight decay 0.01, warmup ratio 0.1, 5 epochs, label smoothing 0.1.

3.2.3 Approach 3: DeBERTa-Large with Simplified Labeling

This approach uses a simplified label set with only three labels: O, ClinicalImpacts, SocialImpacts. The motivation is to reduce the label space and ease optimization: BIO tagging requires the model to learn both entity type and boundary position simultaneously, which may be unnecessarily difficult if the corpus contains few or no consecutive same-type entities. Each input word receives a single label, and all subwords receive the same label, eliminating BIO boundary constraints. Training hyperparameters are identical to Approach 2. During inference, predicted spans are post-processed to recover BIO format. Note that this approach cannot distinguish between consecutive entities of the same-type impacts. This approach is included primarily as an exploratory experiment rather than a standard solution.

3.3 Implementation Details

All experiments were conducted using the Transformers framework (Wolf et al., 2020). Training was performed on a workstation equipped with an NVIDIA RTX Pro 6000 Blackwell GPU with 96GB of VRAM. The complete source code for both tasks is publicly available at https://github.com/TienDoan274/SMM4H.

4 Results

4.1 Task 1 Results

On the development set, Qwen achieves the highest individual F1 (0.8543), benefiting from its larger capacity, while all ensemble methods further improve performance to 0.8583–0.8584, confirming that the three models learn complementary decision boundaries. On the test set, a notable divergence emerges: Qwen dominates on known languages (0.7853) but falls behind the soft voting ensemble on Farsi (0.4865 vs. 0.5304), an untrained language. This suggests that the diversity of an ensemble model provides more robust generalization to unseen scripts compared to a single model.

Method	Precision	Recall	F1-Score
XLM-RoBERTa	0.8484	0.8246	0.8360
InfoXLM	0.8266	0.8482	0.8370
Qwen	0.8806	0.8320	0.8543
Hard Voting	0.8753	0.8430	0.8583
Weighted Soft	0.8764	0.8423	0.8584
Soft Voting	0.8764	0.8423	**0.8584**

Table 1: Macro F1-scores for Task 1 on development set

Method	Overall	Known Languages	Farsi
InfoXLM	0.5195	0.6420	0.3668
Soft Voting	0.5987	0.7161	**0.5304**
Qwen	**0.6623**	**0.7853**	0.4865
Team Mean	0.5465	0.6971	0.3670
Team Median	0.5798	0.7145	0.3797

Table 2: Macro F1-scores for Task 1 on test set. Comparison with team mean and team median. Best results shown in bold.

4.2 Task 7 Results

On the development set, Qwen performs substantially worse than both DeBERTa variants, suggesting that generative models struggle to produce consistent structured outputs for span detection. Between the two DeBERTa variants, BIO labeling achieves higher relaxed F1 on development while simplified labeling leads on strict F1. On the test set, the simplified scheme marginally outperforms BIO on both metrics and exceeds the team mean and median, likely due to the dataset rarely containing consecutive same-type entities, which makes B-/I- boundary distinctions less informative in this specific setting.

Approach	Strict F1	Relaxed F1
Qwen3.5-9B + LoRA	0.2604	0.2961
DeBERTa + BIO	0.4881	**0.6170**
DeBERTa + Non-BIO	**0.5000**	0.5873

Table 3: Performance comparison for Task 7 on the development set

Approach	Strict F1	Relaxed F1
DeBERTa + BIO	0.4840	0.5640
DeBERTa + Non-BIO	**0.5000**	**0.5830**
Team Mean	0.46	0.55
Team Median	0.48	0.58

Table 4: Performance comparison for Task 7 on the test set. Comparison with team mean and team median. Best results shown in bold.

5 Conclusion

We presented NLP systems for multilingual ADE detection (Task 1) and opioid impact span detection (Task 7) at #SMM4H-HeaRD 2026.

For Task 1, Qwen3.5-9B fine-tuned with LoRA achieved the strongest overall performance, showing that instruction-tuned LLMs can effectively learn ADE detection patterns from class-imbalanced multilingual data. However, its advantage was less consistent on the unseen Farsi test set, where soft voting provided better cross-lingual generalization. This gap highlights the difficulty of zero-shot cross-lingual transfer to an unseen script: Farsi's right-to-left writing system and distinct vocabulary likely cause subword tokenizer over-fragmentation, compounded by domain-specific colloquialisms absent from training. The ensemble's advantage on Farsi demonstrates that combining diverse model architectures provides more robust generalization to unseen scripts than any single model (Zhou, 2012).

For Task 7, the simplified labeling scheme trades off boundary precision for learning simplicity by reducing the label space from five to three classes. Its marginal improvement over BIO on the test set suggests that consecutive same-type impacts are rare in this corpus. The substantial underperformance of Qwen compared to DeBERTa reflects the difficulty of generating consistent structured outputs via causal language modeling for token-level NER objectives. Improved prompt templates, constrained decoding, and task-specific loss functions remain promising directions to close this performance gap and advance the reliability of LLM-based span detection.

Future work will explore data augmentation, domain-adaptive pre-training, and constrained decoding to further improve cross-lingual and structured prediction performance.

6 Limitations

For both tasks, we did not apply any data augmentation techniques beyond class upsampling. This was a deliberate choice to preserve the original dataset and keep our focus on modeling approaches; however, augmentation could potentially expose models to richer linguistic patterns and improve generalization. The 3x upsampling strategy for class imbalance is a heuristic choice, and optimal ratios were not systematically explored. For Task 7, the simplified labeling scheme cannot distinguish consecutive same-type impact spans; this is an acceptable trade-off given the current corpus but would be unsuitable for datasets with denser entity distributions. The Qwen-based NER approach was evaluated with a single prompt template; different formulations may yield meaningfully different results.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 30 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Sumon Kanti Dey, Jeanne M. Powell, Azra Ismail, Jeanmarie Perrone, and Abeed Sarker. 2025. Inference gap in domain expertise and machine intelligence in

named entity recognition: Creation of and insights from a substance use-related dataset. *arXiv preprint arXiv:2508.19467*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81.

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, Farnoush Zeidi Kolehparcheh, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Amirali Rezaie Mianroodi, Roland Roller, Judith Rosell, and 10 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.

Michael J. Paul, Abeed Sarker, John S. Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L. Smith, and Graciela Gonzalez. 2016. Social media mining for public health monitoring and surveillance. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 468–479. World Scientific.

Lance A. Ramshaw and Mitch P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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. Association for Computational Linguistics.

Zhi-Hua Zhou. 2012. *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC.

Appendix

Task 1: ADE Detection

```
<|im_start|>system
You are an expert medical text analyzer.
Given a user-generated post, predict
whether the post contains a mention
of an Adverse Drug Event (ADE). An ADE
is a negative medical side effect
associated with a drug. Output '1' if
the post contains at least one mention
of an ADE, or '0' if it does not mention
any ADE.

Output ONLY 1 or 0.
<|im_end|>
<|im_start|>user
Post:
[text]
<|im_end|>
<|im_start|>assistant
```

Task 7: Span Detection

```
<|im_start|>system
You are an expert NER system for medical
text. Given an enumerated list of tokens
(Index: Token), identify ONLY the tokens
that belong to 'SocialImpacts' or
'ClinicalImpacts'.

Output ONLY the labelled tokens, one per
line, in the format:
<Index> : <token> : <Label>
If no token belongs to either label,
output: None
<|im_end|>
<|im_start|>user
Tokens:
{enumerated_tokens}
<|im_end|>
<|im_start|>assistant
```