Discovery@FI at #SMM4H–HeaRD 2026: Ensemble Character Classifier for Multilingual Clinical NER

Petr Zelina and Vít Nováček
Faculty of Informatics, Masaryk University
xzelina2@fi.muni.cz

Abstract

We present a system for multilingual clinical named entity recognition (NER) submitted to the MultiClinNER subtask of MultiClinAI 2026, covering all seven languages and three entity classes (disease, symptom, procedure). Our approach trains one binary token classifier ensemble per entity class using cross-lingual fine-tuning of XLM-RoBERTa-large, with all languages handled jointly. We apply character-level ensembling over six models (two encoder variants × three cross-validation folds). This ensembling method provides more granular probability estimates than single-model classifiers, allowing for more flexible precision-recall trade-off tuning. The system achieves character-level F1 scores of 0.70–0.88 on the official test set.

1 Introduction

We present a language-universal system for multilingual clinical named entity recognition (NER), submitted to all seven languages and all three entity classes of the MultiClinNER subtask of MultiClinAI (Lopez-Garcia et al., 2026). We train one binary token classifier per entity class, each covering all languages simultaneously. Our system achieves competitive results across all languages and entity classes, likely ranking in the top half of participating systems on the main metrics. A detailed overview of all systems is available in the official overview paper (Gallego-Donoso et al., 2026).

The source code is available on GitHub[1]

2 System Description

Our system, illustrated in Figure 1, is based on a transformer encoder fine-tuned for token classification. Each model performs binary (positive/null) classification independently for a single entity class.

[1] https://github.com/ZepZep/ner_character_ensemble

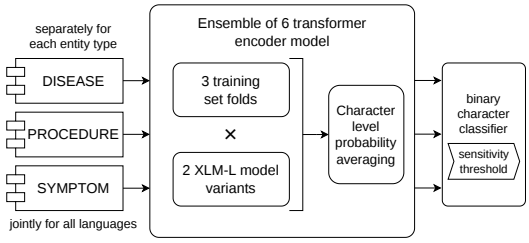

Figure 1: System overview diagram

At inference time, we use an ensemble of several models that may differ in encoder architecture, initialization, and their training set split. Decoding proceeds as follows:

1. Token-level class probabilities are redistributed into character-level probabilities to accommodate models with different tokenizers.

2. Character probabilities are averaged across all ensemble members, yielding an array of shape `[n_chars, n_classes]` per document.

3. Argmax is applied over the non-null classes, assigning a single candidate label per character.

4. If the probability of the winning class falls below a sensitivity threshold, the character is assigned the null class.

5. Character labels are coalesced into contiguous spans.

6. Span boundaries are expanded to full word boundaries.

In the binary classification setup used here, steps 3 and 4 reduce to thresholding the probability of the positive class directly.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 173–176

base model	disease	procedure	symptom
XLM-L	0.867 ± 0.001	0.865 ± 0.005	0.802 ± 0.006
XLM-L-M10K	0.866 ± 0.001	0.866 ± 0.003	0.801 ± 0.006
Fernet-M10K	0.724 ± 0.005	0.737 ± 0.007	0.664 ± 0.004

Table 1: Token F1 performance and standard deviation on the validation set. `M10K` means 10K steps of additional medical pretraining.

3 Experimental Setup

Ensemble. The final ensemble consists of 6 models: 3 cross-validation folds of the training data $\times$ 2 encoder variants. As the base encoder we use `FacebookAI/xlm-roberta-large` (Conneau et al., 2019) (multilingual, 561M parameters). The second variant is initialized from the same checkpoint but further pre-trained on a private dataset of Czech clinical notes (~200M words) using masked language modelling with HuggingFace `run_mlm.py` (April 2025 version), run for 10K steps on a single H100 GPU (batch size 96, bf16 precision, approximately 3 hours). Table 1 shows validation performance with different base models.

Training data. We train one system per entity class (DISEASE, SYMPTOM, PROCEDURE) on all seven languages simultaneously. The training texts for non-Spanish languages differ across entity classes. This is a known artifact of the translation-based dataset construction process, which complicates the training of a single multi-class model, as it would require first unifying the underlying texts and mapping the entities.

Fine-tuning. Each model is fine-tuned using full-context segmentation with a maximum of 512 tokens and a 32-token overlap between segments. Training uses the AdamW optimizer with a learning rate of 10^{-5}, linear decay, and 100 warmup steps, for 2,000 steps with a batch size of 8 (roughly 1 epoch). All 6 models are trained in parallel on a single H100 GPU, taking approximately 16 minutes per run (2 minutes 40 seconds amortized per model).

Inference. The ensemble sensitivity threshold was set to 0.8 based on partial validation performance search. Further threshold analysis may yield additional gains. Inference runs all 6 models in parallel, achieving a combined throughput of approximately 21K tokens/second. This could be further sped up by optimizing or parallelizing the per-character decoding process run in Python, which currently blocks the GPU inference.

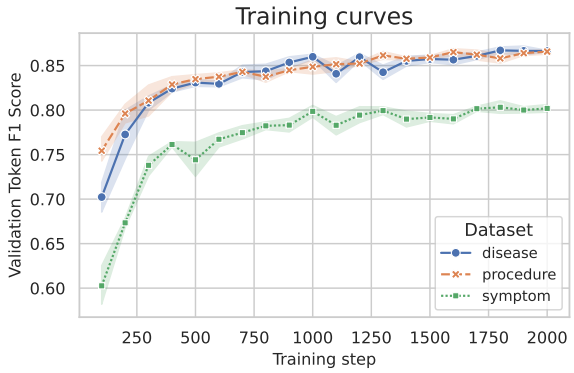

Figure 2: Validation token F1 performance during training (unused fold). Aggregated over both XLM-L variants and the dataset folds. Colored area shows 95% CI. 2000 training steps correspond roughly to 1 epoch.

Czech model experiment. We also evaluated `fav-kky/FERNET-C5` (Lehečka and Švec, 2021) (Czech, 164M parameters) as an alternative second variant, but its validation F1 reached only ~83% of the XLM score, so we did not use it for the final predictions.

4 Results

The official test set evaluation metrics are shown in Table 2. These include *strict* exact match metrics as well as more granular *char*acter level metrics. We were also provided with a rank in each metric; however, the total number of participating systems was not published at the time of writing.

We also provide a similar performance breakdown on the full training set in Table 3. To obtain these numbers, we used the official evaluation script[2].

5 Discussion

Precision-recall trade-off. Our system ranks higher in recall than in precision across most language-entity combinations (Table 2). This reflects the sensitivity threshold of 0.8 tuned on validation F1, which could be raised to improve character precision as well as the F1 score.

The recall-oriented operating point makes the system a natural fit for the retrieval stage of a two-stage pipeline, where candidate spans are passed to a high-precision re-ranker. For example, this could be an instruction-tuned LLM asked only to answer *yes/no* for each candidate, avoiding the token-generation overhead and hallucination risk

[2] `https://github.com/nlp4bia-bsc/MultiClinAIEval/blob/fernando/src/evaluate.py`

		es	en	ro	sv	it	nl	cz
disease	strict_f1	0.728 (10)	0.737 (13)	0.686 (8)	0.662 (7)	0.669 (10)	0.661 (9)	0.639 (9)
	char_f1	0.830 (10)	0.837 (9)	0.811 (6)	0.789 (6)	0.799 (10)	0.775 (7)	0.768 (9)
	char_recall	0.867 (11)	0.855 (7)	0.844 (4)	0.812 (3)	0.832 (2)	0.800 (4)	0.797 (5)
	char_precision	0.796 (13)	0.820 (17)	0.781 (12)	0.767 (9)	0.769 (14)	0.752 (11)	0.741 (13)
procedure	strict_f1	0.717 (9)	0.690 (10)	0.691 (8)	0.687 (7)	0.664 (8)	0.673 (8)	0.658 (8)
	char_f1	0.829 (9)	0.822 (9)	0.823 (5)	0.815 (5)	0.804 (7)	0.801 (4)	0.805 (7)
	char_recall	0.879 (7)	0.848 (2)	0.852 (3)	0.849 (2)	0.846 (2)	0.831 (1)	0.841 (2)
	char_precision	0.785 (13)	0.797 (12)	0.796 (10)	0.783 (9)	0.765 (13)	0.772 (14)	0.771 (11)
symptom	strict_f1	0.624 (8)	0.627 (13)	0.593 (8)	0.598 (8)	0.561 (11)	0.550 (10)	0.542 (13)
	char_f1	0.767 (7)	0.759 (10)	0.753 (5)	0.748 (8)	0.741 (9)	0.710 (6)	0.707 (11)
	char_recall	0.825 (6)	0.781 (6)	0.786 (3)	0.788 (3)	0.791 (2)	0.743 (4)	0.767 (3)
	char_precision	0.718 (12)	0.739 (14)	0.722 (11)	0.712 (11)	0.696 (13)	0.680 (18)	0.655 (14)

Table 2: Performance on the test set. Number in parenthesis shows our system's rank in that metric. Colors are normalized in each row. Languages are sorted from best performing to worst.

		es	en	ro	sv	it	nl	cz
disease	strict_f1	0.749	0.724	0.722	0.730	0.721	0.716	0.722
	char_f1	0.868	0.840	0.848	0.854	0.854	0.845	0.844
	char_recall	0.892	0.885	0.888	0.888	0.887	0.878	0.884
	char_precision	0.845	0.800	0.811	0.822	0.823	0.814	0.807
procedure	strict_f1	0.779	0.735	0.749	0.747	0.744	0.729	0.735
	char_f1	0.880	0.849	0.857	0.858	0.858	0.845	0.847
	char_recall	0.912	0.894	0.901	0.898	0.899	0.884	0.885
	char_precision	0.851	0.809	0.817	0.820	0.820	0.810	0.812
symptom	strict_f1	0.705	0.684	0.639	0.674	0.656	0.656	0.650
	char_f1	0.829	0.809	0.776	0.803	0.808	0.795	0.790
	char_recall	0.875	0.871	0.846	0.861	0.868	0.852	0.854
	char_precision	0.787	0.756	0.716	0.753	0.756	0.745	0.736

Table 3: Performance on the full train set. Colors are normalized in each row.

that make LLMs difficult to use as standalone NER systems.

Spanish vs. translated languages. Table 3 reveals a pronounced performance gap between Spanish (es) and all other languages. We attribute this primarily to the translation-based construction of the multilingual corpus: machine-translated notes occasionally lose or alter entity mentions, resulting in fewer or wrongly annotated entities in the target-language versions.

For instance, using wrong abbreviations, inserting meta-level translation comments into the annotated entity (*this abbreviation has a different meaning in Czech*), or translating *Patient with no history of interest* as *Patient with no interests*.

These inconsistencies introduce noise for non-Spanish languages, making the learning signal weaker and evaluation scores lower.

Interestingly, this sharp boundary does not appear in the test results (Table 2), suggesting that the test set may have undergone more rigorous validation or manual correction.

Ensemble. One practical advantage of ensembling is that averaging across models produces smoother, better-calibrated probability distributions than any single model alone (Lakshminarayanan et al., 2017). We observed that when a model is trained on a small dataset over multiple epochs, token probabilities tend to collapse towards 0 or 1, leaving little useful signal for threshold tuning. Ensemble averaging mitigates this effect by converting model disagreements into continuous probabilities (Lakshminarayanan et al., 2017).

Character-level decoding also simplifies ensemble construction: models can be freely combined regardless of their tokenizer, and individual models can be added or removed without retraining the ensemble. A natural extension would be to weight each model by its validation performance, enabling principled mixing of stronger and weaker classifiers without degrading overall accuracy.

Subset difficulty. Both the training curves (Figure 2) and the absolute test metrics indicate that *symptom* is consistently the hardest entity class. This likely reflects the inherently vague and context-dependent nature of symptom mentions

compared to the more standardised language used for diseases and procedures.

Czech (cz) and Dutch (nl) consistently rank lowest across entity classes. Both languages belong to different families from Spanish and have fewer active speakers, which may partly explain their lower scores. However, Swedish (sv) is similarly distant from Spanish and has similar number of speakers to Czech, yet achieves better performance.

The exact number of competing systems per language–entity slice was not available at the time of writing and likely varies across slices. Therefore, rank comparisons should be interpreted with caution.

6 Conclusion

We presented an ensemble system for multilingual clinical NER, covering all seven languages and all three entity classes of the MultiClinNER subtask. The system combines cross-lingual fine-tuning of XLM-RoBERTa-large with character-level ensemble decoding and sensitivity thresholds. Its fast inference and tunable precision-recall trade-off make it a strong candidate for the retrieval stage of a multi-stage extraction pipeline.

Based on the provided rankings and number of participating systems estimated from our higher precision ranks, our system places roughly in the middle of the field, with better performance on recall-oriented metrics.

Analysis of training scores points to translation noise as the main source of the performance gap between Spanish and the other languages, while the test results suggest this gap narrows when evaluation data quality is higher.

Future work could explore unified multi-class training (once the cross-lingual text alignment issues are resolved), speed up the character-level Python decoder, and utilize a more diverse set of models for the ensemble.

Acknowledgments

Supported by the European Union's Horizon research and innovation programme under grant agreement no. 101057048 (IDEA4RC). Supported by the Grant Agency of Masaryk University under grant no. MUNI/A/1873/2025 (SV26-CODAIS) Funded by the Technology Agency of the Czech Republic under the grant no. TQ12000018 (DecideHealtlh)

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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: Systems, Evaluation, and Datasets. 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.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jan Lehečka and Jan Švec. 2021. Comparison of czech transformers on text classification tasks. In *Statistical Language and Speech Processing*, pages 27–37, Cham. Springer International Publishing.

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.