Enigma at #SMM4H–HeaRD 2026: Leveraging Multilingual Pre-trained Models for Clinical Named Entity Recognition

Sylvia Vassileva **Plamena Ilieva** **Teodor Kostadinov** **Monika Petkova**

Daniel Manchevski **Vitosh Doynov** **Ivan Koychev** **Svetla Boytcheva**

Faculty of Mathematics and Informatics
Sofia University St. Kliment Ohridski
Sofia, Bulgaria

Correspondence: svasileva@fmi.uni-sofia.bg

Abstract

This paper addresses the MultiClinAI challenge, subtask MultiClinNER, which focuses on clinical Named Entity Recognition (NER) across seven languages: Czech, Dutch, English, Italian, Romanian, Spanish, and Swedish. The main goal of MultiClinNER is to identify and extract clinical terms specifically related to diseases, procedures, and symptoms from discharge summaries. The paper explores a variety of state-of-the-art methods, both monolingual and multilingual, ranging from pretrained, zero-shot, domain-adapted transformers to fine-tuned transformer models, and demonstrates the benefits of ensemble modeling. Data augmentation through external resources significantly enhanced the models' ability to recognize clinical entities. Both monolingual and multilingual approaches showed complementary strengths depending on the language and entity type. The average F1 score achieved across the best models for each language and category is 0.6502.

1 Introduction

Automatic recognition of clinical entities in medical text is a fundamental task in biomedical natural language processing (NLP), enabling applications such as clinical decision support, pharmacovigilance, and population-level health analytics.

Clinical Named entity recognition (NER) models trained on large English clinical corpora like MIMIC-III[1], MIMIC-IV[2], i2b2 (Uzuner et al., 2011), n2c2 (Henry et al., 2020) have shown strong performance. Most clinical documentation is in languages other than English, creating a gap between NLP model capabilities and healthcare needs. The BioASQ[3], CLEF[4], and BioCreative challenge[5] ecosystems play a crucial role in benchmarking multilingual clinical NER. Evaluation results from recent challenges — DisTEMIST, which focuses on diseases (Miranda-Escalada et al., 2022); SympTEMIST, covering symptoms, signs, and findings (Lima-López et al., 2023); MultiCardioNER, addressing diseases and drugs (Lima-López et al., 2024); and MedProcNER, dedicated to procedures (Lima-López et al., 2023)—demonstrate that monolingual BERT-based models consistently achieve high accuracy across these tasks. In SympTEMIST, the top NER F1 score of 0.7477 was achieved by an ensemble of fine-tuned models, including BSC-Bio-Es, Roberta-Biomedical-Es, XLMR-Galen, BETO, and mBERT-Galen. MedProcNER's best F1 score, 0.7985, came from a transformer model with masked CRF and data augmentation. DisTEMIST's highest F1, 0.777, was reached by fine-tuning PlanTL-GOB-ES/roberta-base-biomedical-clinical-es. In MultiCardioNER, the highest F1 score of 0.8199 was achieved with an ensemble of RoBERTa-based models.

Multilingual transformers such as mBERT, XLM, and XLM-R allow cross-lingual NLP, but often do not cover clinical domains. Specialized models (e.g., BETO for Spanish, CamemBERT for French, medBERT.de for German) and cross-lingual transfer methods, such as zero-shot learning (e.g., GliNER (Zaratiana et al., 2024)), machine translation, and annotation projection, are used to improve clinical NER, with varying success depending on the language and domain. Recently, large language models (LLMs) have shown promising results in zero-shot clinical NER, although challenges related to entity boundaries and consistency of results remain (Lu et al., 2025), (Monajatipoor et al., 2024).

This paper explores transformer-based ap-

[1] https://doi.org/10.13026/2kv6-1s83
[2] https://doi.org/10.13026/7qgp-kc16
[3] http://bioasq.org/
[4] https://clef2026.clef-initiative.eu/conference/
[5] https://www.ncbi.nlm.nih.gov/research/bionlp/biocreative9

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 113–120

proaches for the MultiClinAI challenge part of the SMM4H workshop (Lopez-Garcia et al., 2026), subtask 1: MultiClinNER [6], combining domain adaptation, cross-lingual learning, and enriched datasets to improve clinical NER across seven European languages (Czech, Dutch, English, Italian, Romanian, Spanish, and Swedish). Our code is available publicly on GitHub [7].

2 Data

The MultiClinNER shared task provides expert-validated clinical NER data for seven languages: Spanish, English, Italian, Dutch, Romanian, Swedish, and Czech (Gallego-Donoso et al., 2026). The multilingual versions of the corpus are produced via machine translation from the Spanish gold standard into each target language. All annotations are distributed in BRAT standoff format with character-level spans and one of three entity types: *Disease*, *Symptom*, or *Procedure*. The total annotations differ by language. Since entity types are provided in separate document sets, we treat the task as three independent labeling problems and train a dedicated model for each.

3 Methods

Our approach addresses the entity NER task separately for each category and language. We fine-tune a separate encoder model for each category and language due to differences in their training datasets and overlapping entities across categories.

The high-level NER pipeline implemented for different categories and languages is presented in Figure 1. It consists of four main steps with slight variations in the implementation in different languages:

Splitting into chunks - since the discharge summaries are quite long and exceed the maximum length for the standard encoder models, we use sentence splitting or a sliding window to process each part of the summary and then combine the results.

Entity token classification - we fine-tune BERT-based models on the token classification task using BIO tagging, where each token is classified as Beginning, Inside, or Outside an entity. For some experiments, we add a CRF layer on top of the classifier.

Multi-model ensembling - we optionally apply ensembling of model predictions by using a majority or weighted voting on the span-level predictions. Depending on the language, a different number of models is used in the ensemble due to training time constraints.

Post-processing - we apply some minor post-processing like cleaning up punctuation and whitespaces, and predictions shorter than 2 characters.

3.1 Chunking Strategies

For chunking, we use two different strategies for the different languages:

Sentence splitting - we apply it to Spanish and Czech. For Spanish, we use the SPACCC Sentence Splitter [8]. For Czech, we use punctuation heuristics with a curated list of Czech abbreviations (e.g., *mudr.*, *např.*).

Sliding window - we use a sliding window approach for English, Romanian, Italian, and Dutch. Token-level logits for words that appear in multiple windows are aggregated via mean pooling.

3.2 Token Classification

For fine-tuning the BERT-based models, we use the train set provided by the organizers. We retain sentences without entities to prevent recall bias. We use the same token classification fine-tuning across all languages, with additional variations applied to Czech. The models and hyperparameters for fine-tuning are listed in Appendix Sections B and C.

For Dutch, we applied filtering on the training data before model training: We removed documents shorter than 100 characters, dropped annotations longer than 10 words (which we considered likely projection errors), and excluded annotations that contained only punctuation or had invalid character offsets. This resulted in removing about 2% of the labeled mentions.

We experiment with data augmentation for Czech, using up-sampling of sentences with entities, as well as morphological synonym augmentation based on a medical dictionary. For Czech, we also experiment with additional **CRF** layer (Lafferty et al., 2001) that enforces valid BIO transitions via Viterbi decoding, preventing invalid label sequences (e.g., `I-X` after `O`).

Tokenizer artifact XLM-RoBERTa's SentencePiece byte-fallback emits overlapping tokens at

[6] https://temu.bsc.es/MultiClinAI/

[7] https://github.com/svassileva/enigma_multiclinai

[8] SPACCC Sentence Splitter

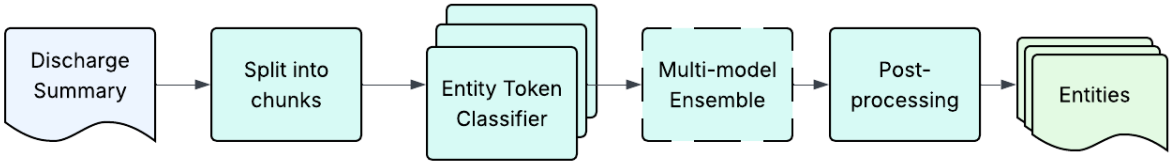

Figure 1: The high-level NER pipeline used for different categories and languages. Multi-model ensembling is an optional step.

identical character offsets ($\sim$4 per Czech clinical document), causing contradictory B$\rightarrow$B training signal and spurious short predicted spans. During *training*, overlapping tokens are assigned label -100 and excluded from the loss. During *inference*, such tokens extend rather than split the current span. As a *post-filter*, spans of $\leq$1 character are discarded, and 2-character spans are retained only if fully uppercase (valid abbrv. such as *DM*, *HT*).

Morphological synonym augmentation - Czech Czech is morphologically rich: the same medical concept may appear in many inflected surface forms (e.g., *hypertenze* $\rightarrow$ *hypertenzí*, *hypertenzi*). We build a Czech medical synonym dictionary (1,399 entries, 887 morphological families) and generate augmented documents by replacing rare entity surface forms (frequency $\leq$5) with morphological variants, yielding 2,161 - 2,271 new documents per entity type - more than tripling the training set size.

LLM paraphrase augmentation - Czech We additionally generated paraphrases using GPT-4.1-mini following the BioSynNER methodology (Lima López et al., 2026), prompting the model to rewrite clinical sentences while preserving entity spans. This provided an additional 0.002-0.004 F1 gain on the development set beyond morphological augmentation.

3.3 Ensembling Strategy

We experiment with two different ensembling strategies for different languages at the span level. We perform experiments with varying numbers of fine-tuned models in the ensemble.

For Italian, we combine the four model predictions via span-level agreement majority voting - a candidate span is accepted if at least K=2 models predict it, irrespective of span boundaries. This majority-agreement strategy balances precision and recall without tuning model weights.

For Dutch, we perform majority voting over 5 model predictions, which yields the best results in test-set experiments. A span is predicted if at least 3 models agree on it. We also experimented with averaging the model outputs; however, this approach can miss entities predicted by a single model that may still be correct. Therefore, we take the union of all model predictions in part of our experiments. When two predictions overlap, we keep the longer one.

In the case of Czech, we combine four model checkpoints via span-level weighted voting (weights proportional to development F1). We use XLM-RoBERTa, two checkpoints of RobeCzech trained with different data augmentation methods, and GPT-4.1-mini zero-shot.

3.4 Post-processing

Predictions were passed through a deterministic post-processing stage that: normalized span boundaries to valid document offsets, dropped duplicates, removed trailing punctuation, filtered invalid punctuation-only spans, and merged overlapping or immediately adjacent spans when the gap consisted only of whitespace or simple connector characters (hyphen, dash, or slash). In addition, we removed Dutch-only stopwords and numbers.

4 Experiments and Results

We split the train set into two parts - train and validation (80%/20%), and conduct initial experiments using the validation set. The best performing models on the test set for different languages and categories are shown in Table 1.

4.1 Fine-tuning Improvements

In our Czech experiments, **Morphological augmentation** is the single most impactful intervention, adding $\approx$4 F1 points over the non-augmented baseline (0.657 $\rightarrow$ 0.710 - 0.712). **CRF decoding** adds a small consistent gain over softmax. Despite being Czech-specific, RobeCzech slightly

Language	Model	Disease F1	Procedure F1	Symptom F1
Czech	RobeCzech + morph. aug.	**0.6552**	-	-
	XLM-R + CRF	0.6381	0.6552	0.5960
	XLM-R + RobeCzech (weighted)	0.6148	0.6492	0.5949
	RobeCzech + oversampling	-	**0.6620**	**0.6070**
English	XLM-R	**0.7452**	**0.7019**	**0.6641**
	XLM-R + GliNER_multi-v2.1	0.7376	-	-
	GliNER_multi-v2.1	0.1697	-	-
Spanish	CLIN-X-ES	**0.6845**	0.6734	0.5451
	RigoBERTa-Clinical	0.6834	**0.6764**	**0.5454**
Italian	4-model ensemble (majority)	**0.7055**	**0.3873**	-
	medbit-r3-plus	0.3484	-	0.5958
	bioBIT	-	-	**0.6133**
Dutch	5-model ensemble (majority)	**0.7119**	**0.7083**	**0.6055**
	3-model ensemble (majority)	0.6868	0.6855	0.5823
	4-model ensemble (union)	0.5336	0.5757	0.4815
Romanian	XLM-R	**0.6920**	**0.7087**	**0.6302**
	XLM-R + GliNER	0.6902	-	-
	GliNER	0.1072	-	-

Table 1: Overall results on the test set for all languages and categories using strict F1.

lags XLM-RoBERTa under morphological augmentation alone (0.698 vs. 0.710), but benefits more from LLM paraphrases, reaching 0.714 as the best single model. **Oversampling** (OS$\times$1) yields no additional gain beyond morphological augmentation, though it helps without augmentation. The **tokenizer artifact fix** improves the performance between 13-14% based on the entity type as shown in Table 2.

4.2 Ensemble Performance

In general, we see performance improvement of ensembles over individual models.

For Italian disease entities, the 4-model ensemble (bioBIT + medbit-r3-plus + umberto + xlmr) achieves strict F1 = 0.706, outperforming a single model (MedBIT-r3-plus, F1 = 0.348) by a substantial margin. The difference is explained by an issue in determining the entity boundaries, which was resolved for the ensemble. The character-level F1 difference between the ensemble and the single model is smaller: 0.7832 vs 0.7706.

For Dutch, majority voting with 5 models (MedRoBERTa.nl + dragon-bert + robbert-v2-dutch-ner + wikineural-multilingual-ner + xlmr-ner-hrl) performed best, with an average F1 across the three categories of 0.67. When comparing a 3-model ensemble (MedRoBERTa.nl + robbert-v2-dutch-base + dragon-bert) using majority voting and a 4-model (xlmr + MedRoBERTa.nl + robbert-v2-dutch + dragon-bert) one using the union of all predictions, the former scored 0.65 on average, while the latter only scored 0.53.

For Czech, the weighted voting achieves F1 = 0.738, outperforming the best single model by 2.4 points. Including GPT-4.1-mini increases recall (0.719 $\rightarrow$ 0.730) without substantially hurting precision, since 65.8% of its correct finds are entities missed by all fine-tuned models.

5 Conclusion

We developed separate models for each entity category and language and experimented with using different BERT-based models for token classification and ensembling their predictions. The ensembles outperform the individual models that comprise them, and we experimented with majority and weighted span-level voting in different languages. Furthermore, we performed different experiments for different languages, including morphological and GPT-4.1-mini augmentation, as well as identified a fix for the XLM-R tokenizer, which improved the model performance.

Limitations

Experiments were conducted on the MultiClinNER dataset, and the results may not generalize directly to other datasets. The ensembling methods are improving the strict F1 scores; however, they may not be feasible in a clinical setting due to computational cost and inference time constraints.

Acknowledgments

This work was partially supported by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria [Grant Project No. BG-RRP-2.004-0008]. The research was partially performed under the Project i-METODE, No BG05SFPR001-3.004-0012-C01, funded under the procedure "Support for the development of project-based doctoral study" from the Programme "Education 2021-2027", co-financed by the European Union. Views and opinions expressed are, however, those of the author only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: grammar and spelling check. After using these tool(s)/ service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

Joeran S Bosma, Koen Dercksen, Luc Builtjes, Romain André, Christian Roest, Stefan J Fransen, Constant R Noordman, Mar Navarro-Padilla, Judith Lefkes, Natália Alves, and 1 others. 2025. The dragon benchmark for clinical nlp. *NPJ Digital Medicine*, 8(1):289.

Tommaso Mario Buonocore, Claudio Crema, Alberto Redolfi, Riccardo Bellazzi, and Enea Parimbelli. 2023. Localizing in-domain adaptation of transformer-based biomedical language models. *Journal of Biomedical Informatics*, 144:104431.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *arXiv preprint arXiv:2109.03570*.

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.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

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.

Sam Henry, Yanshan Wang, Feichen Shen, and Ozlem Uzuner. 2020. The 2019 national natural language processing (nlp) clinical challenges (n2c2)/open health nlp (ohnlp) shared task on clinical concept normalization for clinical records. *Journal of the American Medical Informatics Association: JAMIA*, 27(10):1529–1537.

John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2022. Clin-x: pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain. *Bioinformatics*, 38(12):3267–3274.

Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco-Sánchez, Jan Rodríguez-Miret, and Martin Krallinger. 2023. Overview of symptemist at biocreative viii: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*, page 11.

Salvador Lima-López, Eulàlia Farré-Maduell, Jan Rodríguez-Miret, Miguel Rodríguez-Ortega, Livia Lilli, Jacopo Lenkowicz, Giovanna Ceroni, Jonathan Kossof, Anoop Shah, Anastasios Nentidis, and 1 others. 2024. Overview of multicardioner task at bioasq 2024 on medical specialty and language adaptation of clinical ner systems for spanish, english and italian. In *CLEF (Working Notes)*, pages 8–27.

Salvador Lima López, Judith Rosell, Jan Rodríguez Miret, Fernando Gallego-Donoso, and Martin Krallinger. 2026. MultiClinAI shared task training set + MultiClinNER and MultiClinCorpus test (only texts) and background sets.

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.

Qiuhao Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. 2025. Large language models struggle in token-level clinical named entity recognition. In *AMIA Annual Symposium Proceedings*, volume 2024, page 748.

Guillermo López-García, José M. Jerez, Nuria Ribelles, Emilio Alba, and Francisco J. Veredas. 2021. Transformers for clinical coding in spanish. *IEEE Access*, 9:72387–72397.

Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. *CLEF (Working Notes)*, 3180:179–203.

Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Melika Emami, Fazlolah Mohaghegh, Mozhdeh Rouhsedaghat, and Kai-Wei Chang. 2024. Llms in biomedicine: A study on clinical named entity recognition. *arXiv preprint arXiv:2404.07376*.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. `https://github.com/musixmatchresearch/umberto`.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. Robeczech: Czech roberta, a monolingual contextualized language representation model. In *Text, Speech, and Dialogue*, pages 197–209, Cham. Springer International Publishing.

Guillem García Subies, Álvaro Barbero Jiménez, and Paloma Martínez Fernández. 2025. Clintext-sp and rigoberta clinical: a new set of open resources for spanish clinical nlp. *arXiv preprint arXiv:2503.18594*.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Stella Verkijk and Piek Vossen. 2025. Creating, anonymizing and evaluating the first medical language model pre-trained on dutch electronic health records: Medroberta. nl. *Artificial Intelligence in Medicine*, page 103148.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376.

A Tokenizer Artifact Fix Experiments

The SentencePiece byte-fallback fix brought substantial precision gains (Table 2), confirming that overlapping token positions were generating a large number of spurious short entity spans. Disease precision improved from 0.565 to 0.708; symptom from 0.564 to 0.695.

Entity	System	P	R	F1
Disease	Ensemble (pre-fix)	0.565	0.674	0.615
	Ensemble (post-fix)	0.708	0.675	**0.691**
Procedure	Ensemble (pre-fix)	0.617	0.685	0.649
	Ensemble (post-fix)	0.634	0.692	**0.662**
Symptom	Ensemble (pre-fix)	0.564	0.629	0.595
	Ensemble (post-fix)	0.695	0.582	**0.634**

Table 2: Test set impact of the SentencePiece tokenizer artifact fix.

B Models

The list of models used for different languages is shown in Table 3.

C Hyperparameters

The hyperparameters used for different languages and models are listed in Table 4. Training was performed in Google Colab with an A100 or an L4 GPU, depending on the model size.

Model	Multi-lingual	Domain	Languages
bioBIT (Buonocore et al., 2023)	No	Biomedical	Italian
medBIT-r3-plus (Buonocore et al., 2023)	No	Clinical	Italian
umberto (Parisi et al., 2020)	No	General	Italian
xlm-roberta-base/large (Conneau et al., 2020)	Yes	General	all
roberta-base-biomedical-clinical-es (Carrino et al., 2021)	No	Biomedical / Clinical	Spanish
CLIN-X-ES (Lange et al., 2022)	No	Clinical	Spanish
RigoBERTa-Clinical (Subies et al., 2025)	No	Clinical	Spanish
XLM-R_Galén (López-García et al., 2021)	No	Clinical	Spanish
GliNER_multi-v2.1 (Zaratiana et al., 2024)	Yes	General	English, Romanian
MedRoBERTa.nl (Verkijk and Vossen, 2025)	No	Biomedical	Dutch
RobBERT-v2-dutch-ner (Delobelle et al., 2020)	No	General	Dutch
Dragon-BERT (Bosma et al., 2025)	No	Mixed	Dutch
WikiNEURAL multilingual NER (Tedeschi et al., 2021)	Yes	General	Dutch
XLM-RoBERTa multilingual NER [9]	Yes	General	Dutch
robeczech-base (Straka et al., 2021)	No	General	Czech

Table 3: List of models used for fine-tuning on different languages

GliNER was fine-tuned on all three categories per language - English and Romanian. The GliNER labels were identified via manual testing on the base model: "disease or medical condition", "abnormality", "symptom or sign", "clinical procedure or clinical intervention", and "treatment". The prediction threshold was set to 0.65.

Language	Model	LR	BS	E	P
Italian	BioBIT	3×10^{-5}	16	10	bf16
	MedBIT-r3-plus	3×10^{-5}	16	10	bf16
	UmBERTo	3×10^{-5}	16	10	bf16
	XLM-RoBERTa	2×10^{-5}	16	10	bf16
Spanish	RigoBERTa-Clinical	2×10^{-5}	64	10	bf16
	Clin-X-ES	2×10^{-5}	64	10	bf16
English	XLM-RoBERTa	2×10^{-5}	8	5	bf16
	GliNER	5×10^{-5}	8	3	bf16
Romanian	XLM-RoBERTa	2×10^{-5}	8	5	bf16
	GliNER	5×10^{-5}	8	3	bf16
Dutch	XLM-RoBERTa	1×10^{-5}	8	12	fp16
	MedRoBERTa.nl	2×10^{-5}	8	7	fp16
	DRAGON BERT	2×10^{-5}	8	7	fp16
	RobBERT NER	2×10^{-5}	8	7	fp16
	WikiNEURAL multilingual NER	2×10^{-5}	8	7	fp16
	XLM-RoBERTa multilingual NER	2×10^{-5}	8	7	fp16
Czech	RobeCzech	2×10^{-5}	16	10	fp16
	XLM-RoBERTa	2×10^{-5}	16	10	fp16

Table 4: Hyperparameters used for training the models for different languages. LR - learning rate, Opt - optimizer, WD - weight decay, BS - batch size, E - epochs, P - precision, S - schedule, WR - warmup ratio. Common parameters for all models Opt - AdamW, WD - 0.1, S - linear, WR - 0.01.