BIT.UA at #SMM4H–HeaRD 2026: Towards Multi-Class Multilingual Clinical Entity Recognition with Multi-Head CRF Ensembles

Richard A. A. Jonker
IEETA, DETI, LASI,
University of Aveiro,
richard.jonker@ua.pt

Sérgio Matos
IEETA, DETI, LASI,
University of Aveiro, Aveiro, Portugal
aleixomatos@ua.pt

Abstract

This paper describes the BIT.UA system for the MultiClinNER shared task at #SMM4H–HeaRD 2026, targeting multilingual clinical named entity recognition across seven languages for three entity types (Disease, Procedure, Symptom). We extend the Multi-Head CRF architecture, originally developed for multi-class NER on Spanish clinical text, to the multilingual setting. To enable joint multi-entity training despite per-entity text variations in the dataset, we develop an adaptive text consolidation pipeline that preserves over 94% of annotations. Our central finding is that a single xlm-roberta-large model, trained jointly on all seven languages and three entity types, achieves competition rank 2 for five of seven languages, outperforming dedicated monolingual models by up to +6.94 F1 points, while requiring only a single set of weights. Ensembling multiple seeds of this model achieves rank 1 for those five languages, and combining it with monolingual models yields rank 1 for the remaining two. Code and models are publicly available at https://github.com/ieeta-pt/Multi-Head-CRF/tree/MultiClinNER and https://huggingface.co/collections/IEETA/multiclinner-models.

1 Introduction

Clinical named entity recognition (NER) is fundamental to extracting structured information from medical narratives, supporting tasks such as clinical decision-making and epidemiological research (Raza et al., 2022; Durango et al., 2023; Jonker et al., 2024a). While progress has been made for English clinical NER, most languages lack the annotated corpora and domain-specific language models needed to build competitive systems (Elvas et al., 2025).

The MultiClinNER shared task (Gallego-Donoso et al., 2026; Lopez-Garcia et al., 2026) addresses this gap by providing comparable clinical NER annotations across seven languages: Czech, Dutch, English, Italian, Romanian, Spanish, and Swedish, and for three entity types: Disease, Symptom, and Procedure.

Our system extends the Multi-Head CRF architecture (Jonker et al., 2024a), originally developed for multi-class biomedical NER on Spanish clinical notes. The central question we investigate is whether a single multilingual model can match or exceed the performance of dedicated monolingual systems, each with its own domain-specific model and hyperparameter tuning. This question has practical significance: maintaining separate NER pipelines for each language and entity type is expensive in terms of development effort, compute, and deployment complexity. If a single model can serve all languages competitively, it dramatically reduces the barrier to multilingual clinical NLP adoption. Our contributions are:

1. A systematic comparison showing that jointly training a model on 7 different languages always outperforms its monolingual counterpart.
2. A systematic comparison showing that a single multilingual model outperforms optimized monolingual models for five of seven languages.
3. An entity-level ensemble boosts performance, achieving rank 1 in the official competition for all seven languages across our submissions.

2 Related Work

The Multi-Head CRF model (Jonker et al., 2024a) was introduced for multi-class biomedical NER on Spanish clinical text, combining datasets from SympTEMIST (Lima-López et al., 2023b), MedProcNER (Lima-López et al., 2023a), DisTEMIST (Miranda-Escalada et al., 2022), and Phar-

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 41–48

maCoNER into a unified five-class corpus. The architecture places multiple CRF heads on a shared transformer backbone, enabling detection of overlapping entities across classes with single-model efficiency. This architecture was further shown to already be state of the art in the MultiCardioNER competition related to single entity cardiological NER (Jonker et al., 2024b). Our work extends this to the multilingual setting.

The MultiClinNER training data builds on the Spanish Clinical Case Corpus (Intxaurrondo and Krallinger, 2018) and CardioCCC (Lima-López et al., 2024), with multilingual versions created through machine translation and expert validation. Cross-lingual transfer via multilingual pre-trained models such as XLM-RoBERTa (Conneau et al., 2019) has shown strong results in general-domain NER but remains underexplored for clinical text. Our work evaluates whether such transfer can substitute for language-specific clinical backbones.

3 Methodology

Our approach consists of three components: (1) a data consolidation pipeline that unifies per-entity text variants into a single document for joint training, (2) the Multi-Head CRF model, and (3) an entity-level majority-vote ensemble over multiple model seeds. We train both language-specific models using monolingual backbones and a single multilingual model using xlm-roberta on all seven languages jointly. The following subsections describe each component in detail.

3.1 Data

The MultiClinNER training data contains 1,258 clinical case reports per language across seven languages, covering Disease, Symptom, and Procedure entities. For all languages except Spanish, each entity type has its own text version; annotators independently corrected machine-translated clinical concepts, sometimes producing conflicting corrections across entity types. This prevents direct multi-entity training with the Multi-Head CRF, which requires a single text with aligned spans for multiple entity types.

3.1.1 Adaptive Text Consolidation

We developed a pipeline that selects a single canonical text per document. For documents with divergent text variants, each variant is considered as a candidate anchor. Annotations from the other entity types are realigned using exact positional matching, falling back to contextual similarity (30-character window, SequenceMatcher threshold > 0.6). The anchor retaining the most annotations is selected. This partially reverses annotator corrections for non-anchor entity types, but is necessary to leverage the multi-head architecture for joint multi-entity prediction. Table 1 shows retention rates of 93.82–100%.

Table 1: Annotation consolidation statistics. Ret.: % preserved. Symptom/Disease/Procedure: anchor wins per entity type.

Lang.	Ret.(%)	Lost	S	D	P
ES	100.0	–	–	–	–
EN	99.59	316	279	836	143
NL	99.37	492	245	864	149
CZ	98.16	1,471	512	596	150
SV	97.13	2,255	403	527	328
IT	94.85	4,149	516	302	440
RO	93.82	4,868	604	348	306

Documents were split 830/428 for training/validation per language. A mixed multilingual dataset was created by concatenating all languages (552,567 annotations). At test time, inference ran on each entity-type text variant using the same model.

3.2 Model Architecture

We use the Multi-Head CRF (Jonker et al., 2024a): a shared transformer encoder with one CRF head per entity type, each decoding BIO sequences independently. Training minimizes the sum of negative log-likelihood losses across heads. We trained two model families: (i) language-specific models using monolingual backbones, and (ii) a single multilingual model using xlm-roberta (base for most validation, large for official submissions) on the mixed 7-language dataset. For each language, we swept over context size $\{32, 64\}$, CRF hidden layers $\{1, 3\}$, and augmentation (random/unknown token replacement or none). Fixed hyperparameters are listed in Table 8 (Appendix).

3.3 Experimental Setup

For final submissions, the top 5 configurations per language were retrained on the full dataset across 4 seeds, yielding up to 20 models per language. Seed diversity was prioritized since the F1 range across configurations was narrow ($\leq$1.74 points). Predictions were combined via entity-level majority voting, retaining spans in $\geq\lceil N/2 \rceil$ models. We submitted five runs per language: single best multi-

lingual (ML-1), single best monolingual (Lang-1), and ensembles of multilingual ($\sim$ 8 models, ML-ens), monolingual (11-20 models, Lang-ens), and all runs (All-ens). Details on how many models were trained per language can be seen in Table 13, in the Appendix.

4 Validation Experiments

We performed a large amount of validation experiments, and across 272 runs, the architecture proved robust to hyperparameter variation ($\sigma \leq 0.48$, $\Delta \leq 1.74$ F1, Table 2), as found in the original paper. In general the backbone model selected has a much higher impact than any hyperparameter tuning done, as shown in the Appendix (Table 7). The ablation tables in the Appendix further examine individual hyperparameter effects: context size 64 consistently outperforms 32 across all languages (Table 9), three CRF hidden layers generally improve over one (Table 10), augmentation via random or unknown token replacement yields small but consistent gains over no augmentation (Table 11), and training for 60 epochs helps for most languages except Dutch and Romanian where it leads to slight degradation (Table 12). Fixed training hyperparameters are listed in Table 8.

Table 2: Hyperparameter sensitivity on validation F1 (%).

Lang.	Best	Worst	Δ	σ	n
ES	80.71	79.93	0.78	0.22	24
EN	73.87	72.97	0.90	0.25	24
IT	70.62	69.71	0.91	0.24	22
CZ	69.98	68.84	1.15	0.28	33
RO	67.58	66.34	1.24	0.27	25
SV	70.51	69.19	1.31	0.32	24
NL	68.98	67.24	1.74	0.48	22

Table 3 compares three strategies. A single xlm-roberta-base model trained on all seven languages (XLM-Mix) outperforms the best optimized monolingual backbone for 5/7 languages, despite using a single fixed configuration rather than per-language hyperparameter search. Only Spanish and English, which have strong clinical-domain backbones as well as being much higher resource languages, do not follow this finding. Notably, training multilingually consistently improves over training the same xlm-roberta-base backbone per-language (XLM-Mono), demonstrating genuine cross-lingual transfer rather than simply a better pre-trained model.

Table 4 reports results for the mixed setting. The xlm-roberta-large model achieved a peak F1 of 74.1 during training but suffered from gradient instability across runs, particularly at longer training durations. The large backbone was not extensively hyperparameter-searched due to computational constraints; further tuning may improve stability and performance.

Table 3: Validation F1 (%): best monolingual after hyperparameter search (Mono, best of 14–33 configs) vs. xlm-roberta-base per-language (XLM-M) and on all languages (XLM-Mix), both single fixed configs. †Clinical pre-training.

			XLM-RoBERTa	
Lang.	**Backbone**	**Mono**	**XLM-M**	**XLM-Mix**
ES	roberta-es-clin.†	**80.71**	77.75	78.01
EN	PubMedBERT-b.†	**73.87**	71.42	72.59
IT	medBIT†	70.62	69.75	**71.71**
SV	bert-base-sv	70.51	70.63	**72.48**
CZ	RobeCzech-b.	69.98	69.59	**71.14**
NL	MedRoBERTa.nl†	68.98	69.42	**71.31**
RO	bert-base-ro	67.58	67.77	**69.21**

Table 4: Multilingual model size variation on the mixed validation set. The large model achieved a peak F1 of 74.1 during training but suffered from gradient instability across runs.

Model	Ep.	Best	Mean	Worst	n
xlm-roberta-b.	10	72.65	72.46	72.20	9
xlm-roberta-b.	30	72.20	65.80	56.42	10
xlm-roberta-l.	10	72.37	68.37	64.45	5

Table 5 shows ensemble gains over single models. The improvement scales inversely with baseline performance: +3.05 for Romanian vs. +0.78 for Spanish. The largest jump occurs between Top-1 and Top-5, with diminishing returns thereafter.

Table 5: Ensemble size effect on validation F1 (%).

Lang.	Top-1	Top-5	Top-10	Top-20	Δ
ES	80.71	81.26	81.31	**81.49**	+0.78
EN	73.87	74.64	74.93	**74.94**	+1.07
Mix	72.65	73.51	73.62	**74.11**	+1.46
SV	70.51	72.14	72.31	**72.41**	+1.90
CZ	69.98	71.79	71.72	**71.92**	+1.94
IT	70.62	72.09	72.57	**73.03**	+2.41
NL*	68.98	68.74	69.82	**71.38**	+2.40
RO	67.58	69.80	70.13	**70.63**	+3.05

*Convergence failures: 3/5, 6/10, 14/20 active.

5 Results

Our most important result is the performance of ML-1 (MultiLingual-1): a single xlm-roberta-large

checkpoint handling three entity types across all seven languages simultaneously. Table 6 shows that this single model achieves competition rank 2 for five languages (CZ, IT, NL, RO, SV), outperforming the best dedicated monolingual model (Lang-1), by substantial margins: +4.70 F1 on Czech, +4.87 on Italian, +5.96 on Dutch, and +6.94 on Romanian. This is achieved with a single set of weights and substantially less computational overhead at inference compared to the ensemble submissions.

Table 6: Official competition results (strict F1, %, macro-averaged across entity types). Formatting shows average rank (rounded): **bold+uline** = 1st, **bold** = 2nd, uline = 3rd. Per-entity breakdown in Appendix Table 14.

Lang.	Lang-1	Lang-ens	ML-1	ML-ens	All-ens
CZ	64.73	67.49	**69.43**	**70.75**	69.08
EN	73.69	75.67	75.20	**76.19**	**76.65**
ES	79.03	**79.92**	77.76	78.66	**80.56**
IT	66.93	69.52	**71.80**	**72.55**	71.15
NL	63.11	65.24	**69.07**	**70.11**	67.88
RO	66.56	68.83	**73.50**	**74.22**	70.75
SV	66.14	68.01	**71.33**	**72.25**	70.50

Lang-1/ens: single/ensemble monolingual; **ML-1/ens**: single/ensemble xlm-roberta-large; **All-ens**: all runs combined.

Ensembling multiple seeds of the multilingual model (ML-ens) lifts performance by a further 0.9–1.3 F1 points, achieving rank 1 for all five languages where ML-1 ranks second. For English and Spanish, where stronger clinical-domain backbones exist, the combined ensemble (All-ens) achieves rank 1 by leveraging complementary monolingual and multilingual models.

Even for English and Spanish, where the multilingual model does not top the leaderboard, it remains competitive (76.19, rank 2, and 78.66 F1, rank 6, respectively), with gaps of only 0.46 and 1.90 points to the best strategy. We attribute this gap entirely to the availability of clinical-domain backbones for these large-resource languages, which provide domain-specific priors that a general-purpose xlm-roberta-large cannot match. For the five languages lacking comparable backbones, cross-lingual transfer from joint training more than compensates, confirming the validation finding (Table 3).

Across all languages and strategies, Disease and Procedure entities are consistently easier to detect than Symptoms (Appendix Table 14). For ML-1, the average F1 across languages is 75.15 for Disease, 72.57 for Procedure, and 67.67 for Symptom. This pattern is consistent with prior work on the underlying Spanish corpora (Jonker et al., 2024a) and likely reflects the greater variability in how symptoms are expressed in clinical text compared to the more standardized terminology used for diseases and procedures.

From a deployment perspective, the ML-1 model offers a compelling trade-off: a single checkpoint that can be deployed for any of the seven target languages, handling all three entity types simultaneously, with no language-specific infrastructure required. Compared to maintaining seven separate monolingual pipelines, the multilingual approach reduces both engineering complexity and computational cost while delivering superior performance for the majority of languages.

6 Conclusion

We presented the BIT.UA system for MultiClinNER, extending the Multi-Head CRF architecture to multilingual clinical NER. Our central finding is that a single xlm-roberta-large model, jointly trained on three entity types across seven languages, outperforms dedicated monolingual models for five of seven languages while using a single set of weights, reducing deployment complexity from seven separate pipelines to one. Ensembling provides further gains of +0.78 to +3.05 F1, and combining multilingual with monolingual models achieves rank 1 for all seven languages.

The text consolidation pipeline we developed enables joint training despite per-entity text variations, retaining over 94% of annotations. The architecture is robust to hyperparameter choices ($\Delta \leq 1.74$ F1), suggesting that backbone selection and training strategy matter more than fine-grained tuning.

These results support a practical recommendation: for clinical NER in languages lacking specialized biomedical backbones, a single multilingual model trained on diverse clinical data provides a strong baseline that is difficult to beat with monolingual approaches, at a fraction of the development and inference cost. For high-resource languages with established clinical backbones (such as English and Spanish), combining multilingual and monolingual models via ensembling remains the best strategy. Future work should explore whether the approach generalizes to additional entity types and languages, and whether domain-adaptive pretraining of multilingual models on clinical text can close the gap for high-resource languages.

Limitations

Our text consolidation partially reverses expert translation corrections for non-anchor entity types, potentially training on lower-quality text. The multilingual xlm-roberta-large model was not extensively hyperparameter-searched due to computational constraints, and suffered from gradient instability during training. Our approach handles inter-class entity overlap but not intra-class overlap. Finally, evaluation was performed on machine-translated clinical text; performance on native clinical records may differ.

Acknowledgments

This work was funded by FEDER - Fundo Europeu de Desenvolvimento Regional funds through Programa Regional do Centro, within project CENTRO2030-FEDER-02595400 and by the Foundation for Science and Technology (FCT) through the contract https://doi.org/10.54499/UID/00127/2025. Richard A. A. Jonker is funded by the FCT doctoral grant PRT/BD/154792/2023, with DOI identifier https://doi.org/10.54499/PRT/BD/154792/2023.

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.

María C Durango, Ever A Torres-Silva, and Andrés Orozco-Duque. 2023. Named entity recognition in electronic health records: a methodological review. *Healthcare informatics research*, 29(4):286–300.

Luis B. Elvas, Ana Almeida, and João C. Ferreira. 2025. Natural language processing in medical text processing: A scoping literature review. *International Journal of Medical Informatics*, 204:106049.

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.

Ander Intxaurrondo and M. Krallinger. 2018. Spaccc. Data set, version 2019-02-01.

Richard A A Jonker, Tiago Almeida, Rui Antunes, João R Almeida, and Sérgio Matos. 2024a. Multi-head crf classifier for biomedical multi-class named entity recognition on spanish clinical notes. *Database*, 2024:baae068.

Richard AA Jonker, Tiago Melo Almeida, and Sérgio Matos. 2024b. Bit. ua at multicardioner: Adapting a multi-head crf for cardiology. In *CLEF (Working Notes)*, pages 150–158.

Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2023a. Overview of medprocner task on medical procedure detection and entity linking at bioasq 2023. In *CLEF (Working Notes)*, pages 1–18.

Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco-Sánchez, Jan Rodríguez-Miret, and Martin Krallinger. 2023b. 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.

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.

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.

Shaina Raza, Deepak John Reji, Femi Shajan, and Syed Raza Bashir. 2022. Large-scale application of named entity recognition to biomedicine and epidemiology. *PLOS Digital Health*, 1(12):e0000152.

A Detailed Results and Ablations

Tables 7–12 report detailed ablation results from the 272 validation runs. In Table 7 we can see the different models tested for each language. Key findings from the ablations: context size 64 uniformly outperforms 32 (Table 9), with Dutch benefiting most (+1.32); three CRF hidden layers improve over one for 5/7 languages (Table 10), though English and Italian prefer a single layer; augmentation provides marginal gains (<1 F1 point) with no clear winner between random and unknown token replacement (Table 11); and training for 60 epochs helps most languages except Dutch and Romanian, which show signs of overfitting (Table 12). Table 13 shows how many models were used for each languages ensemble. The original target was 20 per language, however due to exploding gradients and increases of loss during training, we chose to remove certain runs. Table 14 provides the full per-entity competition results with ranks for all five submission strategies.

Table 7: Backbone comparison on validation F1 (%). †Clinical pre-training. Selected models in **bold**.

Lang.	Backbone	Best	Mean	σ	n
ES	**roberta-es-clin.**†	**80.71**	80.34	0.22	24
	bsc-bio-ehr-es†	80.10	79.64	0.18	14
	bert-base-spanish	76.73	–	–	1
EN	**PubMedBERT-b.**†	**73.87**	73.40	0.25	24
	PubMedBERT-l.†	73.68	73.61	0.10	2
	Bio_ClinicalBERT†	71.57	71.26	0.17	14
CZ	**RobeCzech-base**	**69.98**	69.56	0.28	33
	Czert-B-base	53.87	–	–	1
IT	**medBIT**†	**70.62**	70.07	0.24	22
	bert-base-it-xxl	70.17	69.77	0.24	16
NL	**MedRoBERTa.nl**†	**68.98**	68.05	0.48	22
	RobBERT-v2	68.44	68.11	0.22	16
RO	**bert-base-ro**	**67.58**	67.09	0.27	25
SV	**bert-base-sv**	**70.51**	69.94	0.32	24

Table 8: Fixed training hyperparameters.

Parameter	Lang-specific	Multilingual
Learning rate	4×10^{-5}	2×10^{-5}
Batch size	64	16
Warmup ratio	0.1	0.1
Weight decay	0.01	0.01
Max seq. length	512	512
Epochs	30–60	3

Table 9: Context size effect on validation F1 (%).

Lang.	Ctx 32	Ctx 64	Δ
NL	67.65	**68.98**	+1.32
CZ	69.45	**69.98**	+0.54
IT	70.15	**70.62**	+0.46
ES	80.29	**80.71**	+0.42
RO	67.21	**67.58**	+0.37
EN	73.53	**73.87**	+0.34
SV	70.19	**70.51**	+0.32

Table 10: Hidden layers per CRF head on validation F1 (%).

Lang.	1 layer	3 layers	Δ
NL	67.65	**68.98**	+1.32
RO	67.01	**67.58**	+0.57
SV	70.14	**70.51**	+0.37
ES	80.45	**80.71**	+0.26
CZ	69.80	**69.98**	+0.18
EN	**73.87**	73.58	−0.29
IT	**70.62**	70.25	−0.36

Table 11: Augmentation effect on validation F1 (%).

	Random		Unknown		None
Lang.	Best	Mean	Best	Mean	Best
ES	80.64	80.31	**80.71**	80.43	79.96
EN	73.76	73.35	**73.87**	73.47	73.65
CZ	**69.98**	69.54	69.84	69.60	69.89
IT	**70.62**	70.08	70.30	70.06	69.94
NL	**68.98**	67.94	68.65	68.16	68.70
RO	67.43	67.05	**67.58**	67.15	67.41
SV	**70.51**	69.93	70.40	69.94	70.16

Table 12: Training epochs effect on validation F1 (%).

Lang.	30 ep.	60 ep.	Δ
SV	69.98	**70.51**	+0.53
EN	73.40	**73.87**	+0.47
ES	80.44	**80.71**	+0.27
IT	70.41	**70.62**	+0.21
CZ	69.84	**69.98**	+0.15
RO	**67.58**	67.43	−0.15
NL	**68.98**	68.03	−0.95

Table 13: Number of models trained per language. For the mixed category, there are seven large models and one base model (which was included erroneously). Note that for certain languages, one of the seven large models failed during inference. The repository provides weights exclusively for the seven large models.

Language	**Ensemble size per language**
CZ	20
EN	20
ES	20
IT	14
NL	18
RO	18
SV	11
MIXED	7-8

Table 14: Full official competition results with per-entity breakdown (strict F1, %, overall corpus) and competition rank (#).

Lang.	Entity	Lang-1 F1	#	Lang-ens F1	#	ML-1 F1	#	ML-ens F1	#	All-ens F1	#
CZ	Disease	66.10	10	68.38	5	71.34	2	**72.47**	1	70.32	3
	Procedure	67.46	7	70.48	4	71.02	3	**72.72**	1	71.70	2
	Symptom	60.64	8	63.61	4	65.92	2	**67.07**	1	65.22	3
	Avg	*64.73*		*67.49*		*69.43*		***70.75***		*69.08*	
EN	Disease	77.70	7	79.45	3	79.21	4	79.91	2	**80.51**	1
	Procedure	72.01	11	74.14	4	73.91	5	75.02	2	**75.32**	1
	Symptom	71.37	5	73.43	3	72.49	4	73.63	2	**74.11**	1
	Avg	*73.69*		*75.67*		*75.20*		*76.19*		***76.65***	
ES	Disease	81.10	4	81.84	2	79.62	8	80.51	7	**82.43**	1
	Procedure	79.97	5	80.78	2	78.35	8	79.26	6	**81.33**	1
	Symptom	76.02	4	77.15	2	75.30	7	76.21	3	**77.93**	1
	Avg	*79.03*		*79.92*		*77.76*		*78.66*		***80.56***	
IT	Disease	70.09	10	72.58	4	74.41	2	**74.92**	1	73.76	3
	Procedure	68.43	7	70.86	4	72.64	2	**73.87**	1	72.44	3
	Symptom	62.28	8	65.12	4	68.34	2	**68.87**	1	67.26	3
	Avg	*66.93*		*69.52*		*71.80*		***72.55***		*71.15*	
NL	Disease	66.38	11	68.54	7	73.05	2	**73.74**	1	71.00	4
	Procedure	65.75	13	67.67	10	70.90	2	**71.84**	1	70.14	4
	Symptom	57.19	10	59.50	6	63.27	2	**64.75**	1	62.49	3
	Avg	*63.11*		*65.24*		*69.07*		***70.11***		*67.88*	
RO	Disease	67.89	14	69.82	9	76.32	2	**77.04**	1	71.76	5
	Procedure	69.46	10	71.20	6	74.12	2	**75.23**	1	73.41	3
	Symptom	62.34	10	65.47	5	70.05	2	**70.39**	1	67.09	4
	Avg	*66.56*		*68.83*		*73.50*		***74.22***		*70.75*	
SV	Disease	67.63	8	69.41	4	73.07	2	**73.76**	1	71.91	3
	Procedure	68.96	8	70.52	4	72.63	3	**73.77**	1	72.98	2
	Symptom	61.84	8	64.09	6	68.29	2	**69.21**	1	66.60	3
	Avg	*66.14*		*68.01*		*71.33*		***72.25***		*70.50*	