MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition

Besnik Fetahu   Zhiyu Chen   Sudipta Kar   Oleg Rokhlenko   Shervin Malmasi
Amazon.com, Inc.   Seattle, WA, USA
{besnikf,zhiyuche,sudipkar,olegro,malmasi}@amazon.com

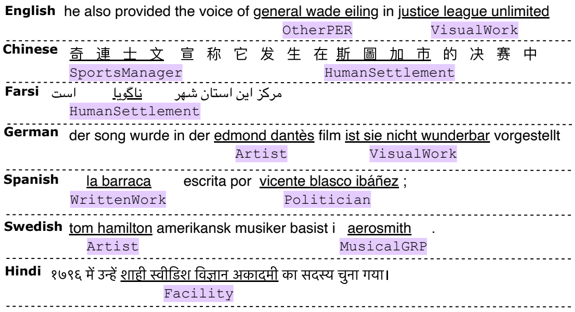
Abstract

We present MULTICONER V2, a dataset for fine-grained Named Entity Recognition covering 33 entity classes across 12 languages, in both monolingual and multilingual settings. This dataset aims to tackle the following practical challenges in NER: (i) effective handling of fine-grained classes that include complex entities like movie titles, and (ii) performance degradation due to noise generated from typing mistakes or OCR errors.

The dataset is compiled from open resources like Wikipedia and Wikidata, and is publicly available.[1] Evaluation based on the XLM-RoBERTa baseline highlights the unique challenges posed by MULTICONER V2: (i) the fine-grained taxonomy is challenging, where the scores are low with macro-F1=0.63 (across all languages), and (ii) the corruption strategy significantly impairs performance, with entity corruption resulting in 9% lower performance relative to non-entity corruptions across all languages. This highlights the greater impact of entity noise in contrast to context noise.

1 Introduction

Named Entity Recognition (NER) is a core task in Natural Language Processing that involves identifying entities and recognizing their type (e.g., *person* or *location*). Recently, Transformer-based NER approaches have achieved new state-of-the-art (SOTA) results on well-known benchmark datasets like CoNLL03 and OntoNotes (Devlin et al., 2019). Despite these strong results, as shown by Malmasi et al. (2022a), there remain a number of practical challenges that are not well represented by existing datasets, including low context, diverse domains, missing casing information, and text with grammatical and typographical errors. Furthermore, wide use of coarse-grained taxonomies often contribute to a lower utility of NER systems, as additional fine-grained disambiguation steps are needed in many practical systems. Furthermore, models trained on datasets such as CoNLL03 perform significantly worse on unseen entities or noisy texts (Meng et al., 2021).



English he also provided the voice of general wade eiling in justice league unlimited
OtherPER
VisualWork

Chinese 奇連士文宣称它发生在斯圖加市的决赛中
SportsManager
HumanSettlement

Farsi مرکز این استان شهر ناگویا است
HumanSettlement

German der song wurde in der edmond dantès film ist sie nicht wunderbar vorgestellt
Artist
VisualWork

Spanish la barraca escrita por vicente blasco ibáñez ;
WrittenWork
Politician

Swedish tom hamilton amerikansk musiker basist i aerosmith .
Artist
MusicalGRP

Hindi १७९६ में उन्हें शाही स्वीडिश विज्ञान अकादमी का सदस्य चुना गया।

Figure 1: Example sentences from MULTICONER V2.

1.1 Real-world Challenges in NER

Outside the news domain used in datasets like CoNLL03, there are many challenges for NER. We categorize the challenges (cf. Table 5 in Appendix A) typically encountered in NER by several dimensions: (i) available context around entities, (ii) named entity surface form complexity, (iii) frequency distribution of named entity types, (iv) multilinguality, (v) fine-grained entity types, and (vi) noisy text containing typing errors. MULTICONER V2 represents all of these challenges, with a focus on (v) and (vi).

Fine-Grained Entity Types Most datasets use coarse types (e.g., PERSON, LOCATION in CoNLL03), with each type subsuming a wide range of entities with different properties. For instance, for LOCATION, different entities can appear in different contexts and have diverging surface form token distributions, which can lead to challenges in terms of NER, e.g., *airports* as part of LOCATION are often named after notable persons, thus for sentences with low context, NER models can confuse them with PERSON entities.

We introduce a fine-grained named entity taxonomy with 33 classes, where coarse grained types are further broken down into fine-grained types.

[1] https://registry.opendata.aws/multiconer
https://huggingface.co/datasets/MultiCoNER/multiconer_v2
https://www.kaggle.com/datasets/cryptexcode/multiconer-2

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 2027–2051
December 6-10, 2023 ©2023 Association for Computational Linguistics

Additionally, we introduce types from the medical domain, presenting a further challenge for NER. Fine-grained NER poses challenges like distinguishing between different fine-grained types of a coarse type. For example, correctly identifying fine-grained types (e.g., SCIENTIST, ARTIST) of a PERSON entity requires effective representation of the *context* and often external world knowledge. Since two instances belonging to different types may have the same name, the context is needed to correctly disambiguate the types.

Noisy Text Existing NER datasets mostly consist of well-formed sentences. This creates a gap in terms of training and evaluation, and datasets assume that models are applied only on well-formed content. However, NER models are typically used in web settings, exposed to user-created content. We present a noisy test set in MULTICONER V2, representing typing errors affected by keyboard layouts. This noise, impacting named entity and context tokens, presents challenges for accurately identifying entity spans.

Contributions Our contributions through this work can be summarized as follows:

1. We created a taxonomy of 33 fine-grained NER classes across six coarse categories (PERSON, LOCATION, GROUP, PRODUCT, CREATIVEWORK, and MEDICAL). We use the taxonomy to develop a dataset for 12 languages based on this taxonomy. Our experiments show that NER models have a large performance gap (≈14% macro-F1) in identifying fine-grained classes versus coarse ones.

2. We implement techniques to mimic web text errors, such as typos, to enhance and test NER model resilience to noise.

3. Experimentally we show that noise can significantly impact NER performance. On average, across all languages, the macro-F1 score for corrupted entities is approximately 9% lower than that for corrupted context, emphasizing the greater influence of entity noise as opposed to context noise.

MULTICONER V2 has been used as the main dataset for the MULTICONER V2 shared task (Fetahu et al., 2023) at SemEval-2023.

2 MULTICONER V2 Dataset Overview

MULTICONER V2 was designed to address the challenges described in §1.1. It contains 12 languages, including multilingual subsets, with a total of 2.3M instances with 32M tokens and a total of 2.2M entities (1M unique). While focused on different challenges, we adopted the data construction process described of the original MULTICONER (v1) corpus Malmasi et al. (2022a) and details about the process are provided in Appendix B. While both the v1 and v2 datasets focus on complex entities, MULTICONER V2 introduces a fine-grained taxonomy, adds realistic noise as well as 5 new languages.

Next we describe the fine-grained NER taxonomy construction and noise generation processes.

2.1 NER Taxonomy

MULTICONER V2 builds on top of the WNUT 2017 (Derczynski et al., 2017) taxonomy entity types, with the difference that it groups GROUP and CORPORATION types into one,[2] and introduces the MEDICAL type. Next, we divide coarse classes into fine-grained sub-types. Table 1 shows the 33 fine-grained classes, grouped across 6 coarse grained types. Our taxonomy is inspired by DBpedia's ontology (Auer et al., 2007), where a subset of types are chosen as fine-grained types that have support in encyclopedic resources.

This taxonomy allows us to capture fine-grained differences in entity types, where the entity context plays a crucial role in distinguishing the different types, as well as cases where external knowledge can help where context may be insufficient (e.g.,PRIVATECORP vs. PUBLICCORP).

2.2 Languages and Subsets

MULTICONER V2 includes 12 languages: `Bangla` (BN), `Chinese` (ZH), `English` (EN), `Farsi` (FA), `French` (FR), `German` (DE), `Hindi` (HI), `Italian` (IT), `Portuguese` (PT), `Spanish` (ES), `Swedish` (SV), `Ukrainian` (UK). The chosen languages span diverse typology, writing systems, and range from well-resourced (e.g., EN) to low-resourced (e.g., FA, BN).

Monolingual Subsets Each of the 12 languages has their own subset with data from all domains, consisting of a total of 2.3M instances.

[2] The original definition is ambiguous in the difference between a CORPORATION and GROUP entity.

PER (Person)	LOC (Location)	GRP (Group)	PROD (Product)	CW (Creative Work)	MED (Medical)
ARTIST	FACILITY	AEROSPACEMANUFACTURER	CLOTHING	ARTWORK	ANATOMICALSTRUCTURE
ATHLETE	HUMANSETTLEMENT	CARMANUFACTURER	DRINK	MUSICALWORK	DISEASE
CLERIC	STATION	MUSICALGRP	FOOD	SOFTWARE	MEDICALPROCEDURE
POLITICIAN	OTHERLOC	ORG	VEHICLE	VISUALWORK	MEDICATION/VACCINE
SCIENTIST		PRIVATECORP	OTHERPROD	WRITTENWORK	SYMPTOM
SPORTSMANAGER		PUBLICCORP			
OTHERPER		SPORTSGRP			

Table 1: MULTICONER V2 fine-grained NER taxonomy. There are 33 fine-grained classes, grouped across 6 coarse types.

Multilingual Subset This subset is designed for evaluation of multilingual models, and should be used under the assumption that the language for each sentence is unknown. The training and development set from the monolingual splits are merged, while, for the test set, we randomly select at most 35,000 samples resulting in a total of 358,668 instances. The test set consists of a subset of samples that incorporate simulated noise.

2.3 Noise Generation

To emulate realistic noisy scenarios for NER, we developed different corruption strategies that reflect typing errors by users. We considered two main strategies: (i) replace characters in a token with neighboring characters (mimicking typing errors) on a given keyboard layout for a target language, and (ii) replace characters with visually similar characters (mimicking incorrectly visually recognized characters, e.g., OCR).

We applied the noise only on the test set[3] for a subset of languages (EN, ZH, IT, ES, FR, PT and SV). The noise can corrupt *context* tokens and *entity* tokens. For ZH, the corruption strategies are applied at the character level. We followed Wang et al. (2018)'s approach that automatically extracts visually similar characters using OCR tools and phonologically similar characters using ASR tools. We used their constructed dictionary[4] to match targeting characters.

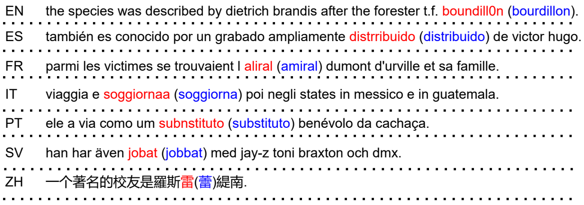
For the other six languages, we developed word-level corruption strategies based on common typing mistakes, utilizing language specific keyboard layouts.[5] Table 2 summarizes the designed corruptions and their probabilities of occurrence.

Noise Sampling We introduce synthetic noise into 30% of the test set. For each corrupted sentence, the number of corrupted tokens δ is sampled from a truncated Poisson distribution ($\lambda = 1$ and $\delta \leq 3$). Since the minimum of δ can be zero, we use $\delta' = \delta + 1$ as the final corruption count.

Corruption Strategy	Prob.
Adds a neighbor letter on keyboard next to a random token letter.	0.1
Replaces a random word letter with its keyboard-neighbor letter.	0.2
Replaces a random word letter with another visually similar one.	0.2
Skips a random word letter.	0.2
Swaps two random consecutive word letters.	0.1
Repeats a random word letter.	0.2

Table 2: Corruption strategies and the probability with which they are applied.

Corruptions can apply to entity or non-entity (context) tokens. For ZH, we corrupt a context or entity character with 15% and 85% probability respectively. Using the corruption dictionary, we replace the original character at random with a visually or phonologically similar character from identified candidates. For the other languages, the probability of corrupting entity tokens is set to 80%. This is lower given that each letter in entities can be corrupted and the number of candidates in entities is more than that in ZH. On a word chosen for corruption, we perform a corruption action based on probabilities shown in Table 2. Examples of corrupted sentences are shown in Figure 2.



EN	the species was described by dietrich brandis after the forester t.f. boundill0n (bourdillon).
ES	también es conocido por un grabado ampliamente distrribuido (distribuido) de victor hugo.
FR	parmi les victimes se trouvaient l aliral (amiral) dumont d'urville et sa famille.
IT	viaggia e soggiornaa (soggiorna) poi negli states in messico e in guatemala.
PT	ele a via como um subnstituto (substituto) benévolo da cachaça.
SV	han har även jobat (jobbat) med jay-z toni braxton och dmx.
ZH	一个著名的校友是羅斯雷(蕾)緹南.

Figure 2: Examples of corrupted sentences. The red tokens are the corrupted versions of the blue tokens.

[3] To emulate a realistic scenario, where models are trained on clean text, but are exposed directly to user input.

[4] https://github.com/wdimmy/Automatic-Corpus-Generation

[5] We extended the keyboard layouts to include 7 languages: https://github.com/ranvijaykumar/typo.

3 NER Model Performance

We assess if MultiCoNER v2 is challenging (cf. §1.1) by training and testing with XLM-RoBERTa (Conneau et al., 2020, XLMR).

Metrics We measure the model performance using Precision (P), Recall (R), and F1, where for F1 we distinguish between *micro/macro* averages.

3.1 Results

Table 3 shows the results obtained for the XLMR baseline. The results show the micro-F1 scores achieved on the individual NER coarse classes (See Appendix C for fine-grained performance).

Across all subsets, XLMR achieves the highest performance of micro-F1=0.77 for HI, and lowest micro-F1=0.61 for EN, ES and FA. The reason for the higher F1 scores on HI (or BN second highest score) is due to the smaller test set size, since the data is translated and is more scarce. For the rest, the F1 scores are typically lower, this is is mainly associated with the larger test sizes.

Lang.	Macro-F1	Micro-F1
BN	0.68	0.74
DE	0.67	0.71
EN	0.53	0.61
ES	0.53	0.61
FA	0.52	0.61
FR	0.56	0.64
HI	0.71	0.77
IT	0.58	0.63
UK	0.57	0.66
SV	0.57	0.70
PT	0.54	0.65
ZH	0.58	0.63
MULTI	0.63	0.69

Table 3: XLMR performance on MultiCoNER v2.

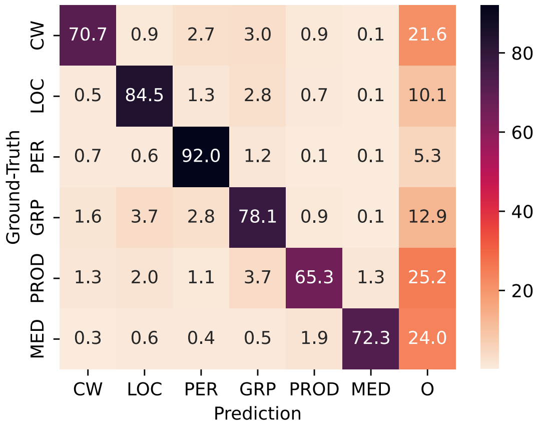


Figure 3: Coarse grained errors on the MULTI test set.

3.2 Impact of Fine-Grained Taxonomy

Performance varies greatly across fine-grained types within a coarse type (Table 10). Using MULTI as a representative baseline, we note a gap of up to 34% in F1 scores, as seen with CreativeWork and Person, indicating the challenges of introducing fine-grained types not seen at the coarse level (Malmasi et al., 2022b).

The gap can be attributed to two primary components: (i) the incorrect classification of entities under inappropriate sub-categories or broad categories, and (ii) recall problems where certain entities are missed (assigned the O tag).

Misclassifications Figure 3 shows the distribution of predicted coarse tags for MULTI. Most of the misclassifications are for the Group class, which is often misclassified with Location or Person. Yet, this does not reveal how the models handle the specific classes within a coarse type.

To understand the challenges presented by our taxonomy for NER models, we further inspect misclassifications within the coarse type Group in Figure 7 of Appendix D. There are several types that are misclassified, e.g., PublicCORP as ORG, or PrivateCORP with ORG or PublicCORP. On the other hand, MusicalGRP is often misclassified with types from other coarse types, such as Artist or VisualWork. This highlights some of the issues and challenges that NER models face when exposed to more fine-grained types.

Recall Issues Figure 3 shows the rate of entities marked with the O tag. Consistent recall issues affect the Medical, CreativeWork, and Product classes, often with more than 20% of entities being missed. Across languages the recall can vary, with even higher portions of missed entities (e.g., FA, ZH, EN with more than 30+% of Medical entities missed, cf. Figure 6).

3.3 Impact of Corruption Strategies

Figure 4 shows the impact of noise level on the macro-F1 score, where we see a strong negative correlation between noise level and performance (Spearman's correlation across different languages is $\widehat{\rho} = -0.998$). For ZH, the impact is higher than for the rest of languages and the macro-F1 decreases faster: with 50% of test data corrupted, the score decreases approximately by 9%, while for the rest it drops only between 4% and 5.5%.

Table 4 shows the impact of corruption strategies on *context* and *entity* tokens, targeting exclusively the respective tokens. For both variants, the number of corruptions per sampled with $\lambda = 1$

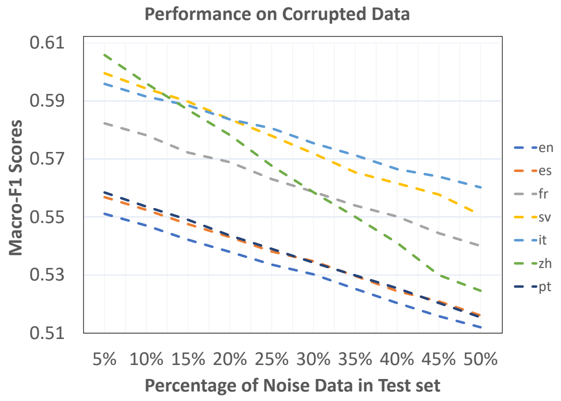


Figure 4: Noise impact on performance. X-axis shows the rate of noise, and Y-axis shows the macro-F1 score. Our final test set applies noise to 30% of sentences.

Corruption Target	EN	ES	FR	IT	PT	SV	UK	ZH
Context	0.48	0.51	0.50	0.54	0.51	0.55	0.54	0.55
Entity	0.42	0.42	0.45	0.46	0.42	0.44	0.41	0.48

Table 4: macro-F1 scores when all the test data is corrupted on *context* or *entity* tokens.

and $\delta \leq 4$ hyper-parameters. Results show that entity token corruption has a notably detrimental effect on performance, in contrast to context token corruption. On average across all languages, the macro-F1 score for corrupted entities is approximately 9% lower than the performance on corrupted context, emphasizing the higher impact of entity noise compared to context noise. Finally, Table 11 shows the baseline's performance breakdown on the clean and noisy subsets of the test set.

4 Conclusions and Future Work

We presented MULTICONER V2, a new large-scale dataset that represents a number of current challenges in NER. The results from XLMR indicate that our dataset is difficult, with a macro-F1 score of only 0.63 on the MULTI test set.

The results presented illustrate that the MULTICONER V2, with its detailed NER taxonomy and noisy test sets, poses substantial challenges to large pre-trained language models' performance. It is our hope that this resource will help further research for building better NER systems. This dataset can serve as a benchmark for evaluating NER methods that infuse external entity knowledge, as well as assessing NER model robustness on noisy scenarios.

We envision several directions for future work. The extension of MULTICONER V2 to additional languages is the most straightforward direction for future work. Expansion to other more challenging domains, such as search queries (Fetahu et al., 2021), can help explore fine-grained NER within even shorter inputs.

Finally, evaluating the performance of this data using zero-shot and few-shot cross-lingual models (Fetahu et al., 2022) or even unsupervised NER models (Iovine et al., 2022) can help us better understand the role of data quantity in building effective NER systems.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer.

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, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gazetteer enhanced named entity recognition for code-mixed web queries. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1677–1681.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2022. Dynamic gazetteer integration in multilingual models for cross-lingual and cross-domain named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2777–2790.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. SemEval-2023 Task 2: Fine-grained Multilingual Named Entity Recognition (MultiCoNER 2). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Andrea Iovine, Anjie Fang, Besnik Fetahu, Oleg Rokhlenko, and Shervin Malmasi. 2022. Cyclener: an unsupervised training approach for named entity recognition. In *Proceedings of the ACM Web Conference 2022*, pages 2916–2924.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. SemEval-2022 task 11: Multilingual complex named entity recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437, Seattle, United States. Association for Computational Linguistics.

Stephen Mayhew, Tatiana Tsygankova, and Dan Roth. 2019. ner and pos when nothing is capitalized. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6256–6261, Hong Kong, China. Association for Computational Linguistics.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. GEMNET: effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1499–1512. Association for Computational Linguistics.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for chinese spelling check. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2517–2527.

Appendix

A NER Challenges & Motivation

Table 5 lists the key challenges that are introduced in MULTICONER V2 for NER model. The challenges range from low context sentences, complex entities with additionally fine-grained NER taxonomy, to syntactic issues such as lowercasing as well as noisy context and entity tokens.

B Dataset Construction

This section provides a detailed description of the methods used to generate our dataset. Namely, how sentences are selected from Wikipedia, and additionally how the sentences are tagged with the corresponding named entities and how the underlying fine-grained taxonomy is constructed. Additionally, the different data splits are described, as well as the license of our dataset.

Table 6 shows the detailed stats about our proposed MULTICONER V2 dataset.

B.1 Fine-Grained NER Taxonomy

We use our fine-grained NER taxonomy, namely, the entities associated with each type as part of the automated data generation process. To generate entity associations to our fine-grained taxonomy we make use of Wikidata. The entity type graph in Wikidata is noisy, containing *cycles*, and often types are associated with different parent types, thus, making it challenging to use the data as is for generating fine-grained and clean taxonomies for NER. We carry out a series of steps that we explain in the following.

Type Cycles. Wikidata type graph contains cycles, which are not suitable for taxonomy generation. We first break the cycles through a depth-first search approach, by following the children of a given type, and whenever a given type has children that are already visited (encountered during the DFS), those edges are cut out.

Entity to Fine-Grained Type Association: Our taxonomy consists of 6 coarse types, PERSON, LOCATION, GROUP, MEDICAL, PRODUCT, CREATIVEWORK, and 33 of fine-grained types. The coarse types represent classes that we aim to have in our taxonomy, and with the exception of MEDICAL and PRODUCT, which at the same time represent contributions and enrichment of the current NER data landscape, represent common types encountered in NER research. We generate the entity to type association at the fine-grained level, given that the association at the coarse level can be inferred automatically.

For each of the fine-grained types we *manually identify* specific Wikidata types, from where we then query the associated entities. Wikidata types and entity type associations are done by the Wikidata community, and thus, often there are redundant entity types (with similar semantics). From the initially identified types and their entities, we expand our set of Wikidata types that are associated to a fine-grained type in our taxonomy, by additionally taking into account other Wikidata types that the entities are associated with. From the identified Wikidata types for a fine-grained type in our taxonomy, we traverse the entire Wikidata type graph, and obtain all the entities that are associated with those types and all of their children.

Finally, this resulted in 16M Wikidata entities, whose distribution across the different coarse and fine-grained types is shown in Table 7 and 8.

Quality Filtering: Finally, once we have obtained all the possible Wikidata entities that are associated with the fine-grained types in our NER taxonomy, we need to ensure that the associations are not *ambiguous*.

We tackle the problem ambiguity, namely Wikidata entities being associated with different types, causing them to be in different NER classes at the same time. To resolve such ambiguous cases, we compute the distribution of entities that are associated with a set of fine-grained NER type, and based on a small sample analysis, manually determine with which NER type they should be associated.

B.2 Sentence Extraction

MULTICONER V2 is constructed from encyclopedic resources such as Wikipedia, with focus on sentences with low context, sampled using heuristics that identify sentences that represent the NER challenges we target in Table 5. Figure 5 shows the steps from our data construction workflow, which for a given language it extracts sentences from the corresponding Wikipedia locale, and applies a series of steps to generate the NER data. This process is performed for the following languages: FA, FR, EN, ES, IT, PT, SV. For the rest of languages

Challenge	Description
Fine-grained Entities	The entity type can be different based on the context. For example, a creative work entity "`Harry Potter and the Sorcerer's Stone`" could be s a book or a film, depending on the context.
Noisy NER	Gazetteer based models would not work for typos (e.g., "`sony xperia`" → "`somy xpria`") or spelling errors (e.g., "`ford cargo`" → "`f0rd cargo`") in entities, degrading significantly their performance.
Ambiguous Entities and Contexts	Some NEs are ambiguous: they are not always entities, e.g., "`Inside Out`", "`Among Us`", and "`Bonanza`" may refer to NEs (a movie, video game, and TV show) in some contexts, but not in others. Such NEs often resemble regular syntactic constituents.
Surface Features	Capitalization/punctuation features are large drivers of success in NER (Mayhew et al., 2019), but short inputs (ASR, queries) often lack such surface features. An uncased evaluation is needed to assess model performance.

Table 5: Challenges addressed by MULTICONER V2.

Class	**Split**	**EN**	**DE**	**FA**	**FR**	**ES**	**UK**	**SV**	**HI**	**BN**	**ZH**	**IT**	**PT**	**Multi**
PER	train	9,294	5,508	8,006	9,295	8,360	6,441	7,695	3,609	3,778	4,862	10,387	8,241	85,476
	dev	481	280	413	483	442	341	445	174	194	239	548	447	4,487
	test	137,681	11,299	115,868	141,401	125,379	96,864	111,157	5,736	6,935	9,095	160,598	120,413	180,080
CW	train	4,084	2,466	3,661	5,438	3,606	2,907	3,714	1,646	1,981	2,264	5,048	3,839	40,654
	dev	215	127	184	268	183	146	200	90	103	112	267	206	2,101
	test	62,126	4,777	53,034	84,952	55,459	43,291	54,806	2,804	3,640	4,369	79,873	58,245	87,030
GRP	train	4,224	2,815	3,209	3,745	3,632	3,204	3,459	2,273	2,227	2,696	3,416	3,788	38,688
	dev	218	177	180	195	195	151	194	143	122	145	173	200	2,093
	test	60,026	4,418	38,807	52,987	50,259	39,709	46,929	3,897	3,651	4,715	46,271	48,994	73,226
LOC	train	4,353	2,269	5,086	4,723	4,651	5,458	7,176	2,487	2,457	2,470	4,446	4,794	50,370
	dev	197	117	267	242	230	294	370	133	127	129	248	250	2,604
	test	67,901	5,306	70,907	73,373	72,996	84,643	111,879	7,172	7,375	6,170	68,564	70,923	117,257
PROD	train	1,935	1,571	2,049	1,946	1,989	2,258	1,989	1,420	1,384	1,529	1,770	1,927	21,767
	dev	109	78	107	100	100	117	112	74	67	73	86	101	1,124
	test	27,580	1,643	18,212	28,274	28,469	30,071	22,686	1,611	1,493	1,869	22,887	21,115	35,545
MED	train	1,559	1,322	1,651	1,230	1,669	1,688	1,381	1,435	1,396	1,407	1,376	1,850	17,964
	dev	76	62	85	64	81	86	70	70	63	75	76	88	896
	test	22,491	1,434	15,287	17,208	23,812	20,796	13,702	1,979	1,919	1,781	19,029	21,062	29,553
Total	train	16,778	9,785	16,321	16,548	16,453	16,429	16,363	9,632	9,708	9,759	16,579	16,469	170,824
	dev	871	512	855	857	854	851	856	514	507	506	858	854	8,895
	test	249,980	20,145	219,168	249,786	246,900	238,296	231,190	18,399	19,859	20,265	247,881	229,490	358,668

Table 6: MULTICONER V2 dataset statistics for the different languages for the Train/Dev/Test splits. For each NER class we show the total number of entity instances per class on the different data splits. The bottom three rows show the total number of sentences for each language.

Coarse Class	#Entities	ratio
PERSON	9,586,716	57%
CREATIVEWORK	3,146,607	18.7%
LOCATION	3,010,448	17.9%
GROUP	607,420	3.5%
PRODUCT	396,611	2.4%
MEDICAL	29,968	0.18%

Table 7: Gazetteer stats at the coarse named entity type level.

(BN, ZH, DE, HI), we apply machine translation to obtain the data, as described below.

In the following we explain the required steps to generate the final output of sentences tagged with the corresponding entity types.

1. **Sentence Extraction:** From the cleaned Wikipedia pages, we extract sentences, which contain hyperlinks that point to entities that can be mapped into their corresponding Wikidata entity equivalent. In this way we ensure that the entities in a sentence are marked with high quality to their corresponding fine-grained types (cf. Table 8).
2. **Sentence Filtering:** We remove sentences that are shorter than 28 characters and longer than 128 characters. In this way, we ensure that the sentences contain complete clauses,[6] and furthermore, by filtering out longer sentences (more than 128 characaters) we ensure that the remaining sentences have low context, making NER more challenging. Finally, we additionally filter out sentences that contain capitalized nouns (proper nouns), which are not interlinked

[6] i.e., sentences shorter than 28 characters may contain only entity names, and as such are not suitable for NER

Coarse Class	Fine-Grained Class	#Entities	ratio
PERSON	OTHERPER	4,343,697	25.89%
	SCIENTIST	1,901,438	11.333%
	ARTIST	1,392,664	8.301%
	CLERIC	134,078	0.799%
	SPORTSMANAGER	48,741	0.291%
	ATHLETE	1,003,554	5.981%
	POLITICIAN	762,544	4.545%
LOCATION	FACILITY	2,470,918	14.727%
	OTHERLOC	256,644	1.53%
	STATION	196,113	1.169%
	HUMANSETTLEMENT	86,773	0.517%
GROUP	ORG	392,390	2.339%
	MUSICALGRP	106,643	0.636%
	SPORTSGRP	83,902	0.5%
	PUBLICCORP	12,570	0.075%
	PRIVATECORP	1,888	0.011%
	CARMANUFACTURER	9,528	0.057%
	AEROSPACEMANUFACTURER	499	0.003%
PRODUCT	OTHERPROD	213,848	1.275%
	VEHICLE	144,555	0.862%
	FOOD	17,791	0.106%
	CLOTHING	11,604	0.069%
	DRINK	8,813	0.053%
CREATIVEWORK	WRITTENWORK	1,366,197	8.143%
	ARTWORK	575,665	3.431%
	VISUALWORK	540,773	3.223%
	MUSICALWORK	539,878	3.218%
	SOFTWARE	87,017	0.519%
MEDICAL	DISEASE	14,627	0.087%
	MEDICATION/VACCINE	5,947	0.035%
	MEDICALPROCEDURE	4,286	0.026%
	ANATOMICALSTRUCTURE	4,186	0.025%
	SYMPTOM	922	0.005%

Table 8: Gazetteer stats at the coarse named entity type level.

to entities. This is done for languages that follow capitalization rules for proper nouns (e.g. EN and ES). This presents an important step, since if a proper noun is not interlinked to an entity, however it represents one, this will artificially inflate false positive scores for NER models.

B.3 Sentence Translation

For the languages, for which we cannot automatically extract sentences and their corresponding tagged entities (e.g. DE uses capitalization for all nouns, thus, we cannot accurately distinguish nouns against proper nouns), we apply automatic translation to generate two portions of our data. For four languages (BN, DE, HI, ZH), Wikipedia sentences are translated from the EN data.

We use the Google Translation API[7] to perform our translations. The input texts may contain known entity spans or slots.

To prevent the spans from being translated, we leveraged the notranslate attribute to mark them and prevent from being translated. The translation quality in the different languages such as BN, ZH, DE and HI is very high, with over 90% translation accuracy (i.e., accuracy as measured by human annotators in terms of the translated sentence retaining the semantic meaning and as well have a correct syntactic structure in the target language).

B.4 Data Splits

To ensure that obtained NER results on this dataset are *reproducible*, we create three predefined sets for training, development and testing. The entities in each class are distributed following a root square normalization, which takes into account the actual distribution of a given entity class. Additionally we set a minimum number of entities from a class that need to appear in each split. This combination allows us not to heavily bias the data splits towards the more popular classes (e.g. WRITTENWORK). Table 6 shows detailed statistics for each of the 13 subtasks and data splits.

Training Data For the training data split, we limit the size to approx. 16k sentences. The number of instances was chosen to be comparable to well-known NER datasets such as CoNLL03 (Sang and De Meulder, 2003). Note that in the case of the Multi subset, the training split contains all the instances from the individual language splits.

Development Data We randomly sample around ∼800 instances,[8] a reasonable amount of data for assessing model generalizability.

Test Data Finally, the testing set represents the remaining instances that are not part of the training or development set. To avoid exceedingly large test sets, we limit the number of instances in the test set to be around at most 250k sentences (cf. Table 6). The only exception is for the Multi, which was generated from the language specific test splits, and was downsampled to contain at most 35k from each monolingual test set, resulting in a total of 358k instances. The larger test set sizes are ideal to assess the generalizability and robustness of models on unseen entities.

B.5 License, Availability, and File Format

The dataset is released under a CC BY-SA 4.0 license, which allows adapting the data. Details

[7] https://cloud.google.com/translate

[8] With the exception of the translated languages, DE, HI, BN, and ZH, where the data is scarce.

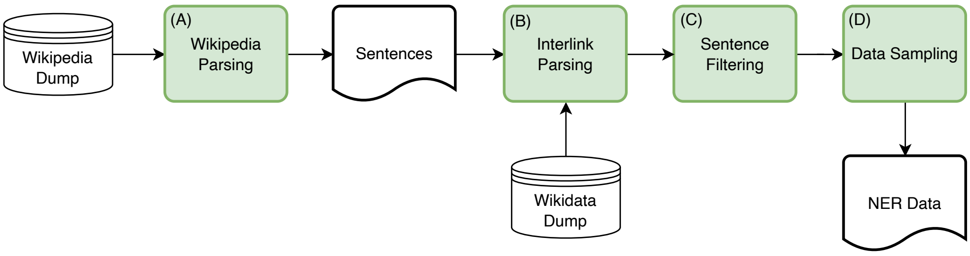


Figure 5: An overview of the different steps involved in extracting the MULTICONER V2 data from Wikipedia dumps, similar to MULTICONER.

about the license are available on the Creative Commons website.[9]

The data is distributed using the commonly used BIO tagging scheme in CoNLL03 format (Sang and De Meulder, 2003). The complete dataset will be available for download.[1]

B.6 Noise Generation

Original Character	Visually Similar Ones	Phonetically Similar Ones
干	千,午,汗	肝.敢.甘
防	肪,访,妨	方,房,芳

Table 9: Examples of visually and phonetically similar ZH characters.

Finally, as part of the data construction, we generate a subset of corrupted test set (30%) that adds noise to either the context or entity tokens as described in in §2.3. The purpose of the noisy test set is to assess the robustness of NER models, trained on clean data, in spotting correctly the named entities that may contain either context or entity noise. The noise mimics typing noise or noise that may be attributed to OCR, where characters are misrecognized with visually similar characters.

For all the languages the noise is applied at the token level, where random characters within a token are replaced with either a character that comes from typing errors (neighboring characters on a given keyboard layout for a specific language) or visually similar characters. For ZH is applied at character level, this is due to the language structure. Table 9 shows examples of visually similar and phonetically similar characters in ZH.

Figure 2 shows examples of corrupted sentences from seven different languages.

C Fine Grained NER Performance

Performance on Fine-grained Classes: Table 10 shows the results for the different fine-grained classes and languages. We note a high divergence in terms of performance for the different fine-grained types within a coarse type.

Performance on Noise and Clean Test Subsets: From Table 11, we can see that the simulated noise can have a huge impact on the performance of the baseline model. On average, the Macro-F1 score on noisy subset is about 10% lower than that on the clean subset. In this section, we study the impact of different corruption strategies, and amount of corruption on the test set. Note that, differently from the fixed amount of noise of 30% applied on the test set, here we assess the range of noise with up to 50% of the test data.

D Fine Grained NER Error Analysis

Here we show the errors that the XLMR baseline makes at the fine-grained type level. Figure 7 shows the misclassifications of the baseline approach. Here, we notice that within the GROUP type there are several fine-grained types that are often confused with other types within GROUP, e.g. PUBLICCORP often misclassified as ORG, or PRIVATECORP often misclassified with ORG or PUBLICCORP. On the other hand, MUSICALGRP is often misclassified with other fine-grained types from other coarse types, such as ARTIST or VISUALWORK.

The error analysis highlights some of the issues that NER models face when exposed to more fine-grained types. Furthermore, through this analysis, we provide insights on the challenges presented by the newly proposed NER taxonomy and areas of improvement in terms of NER performance.

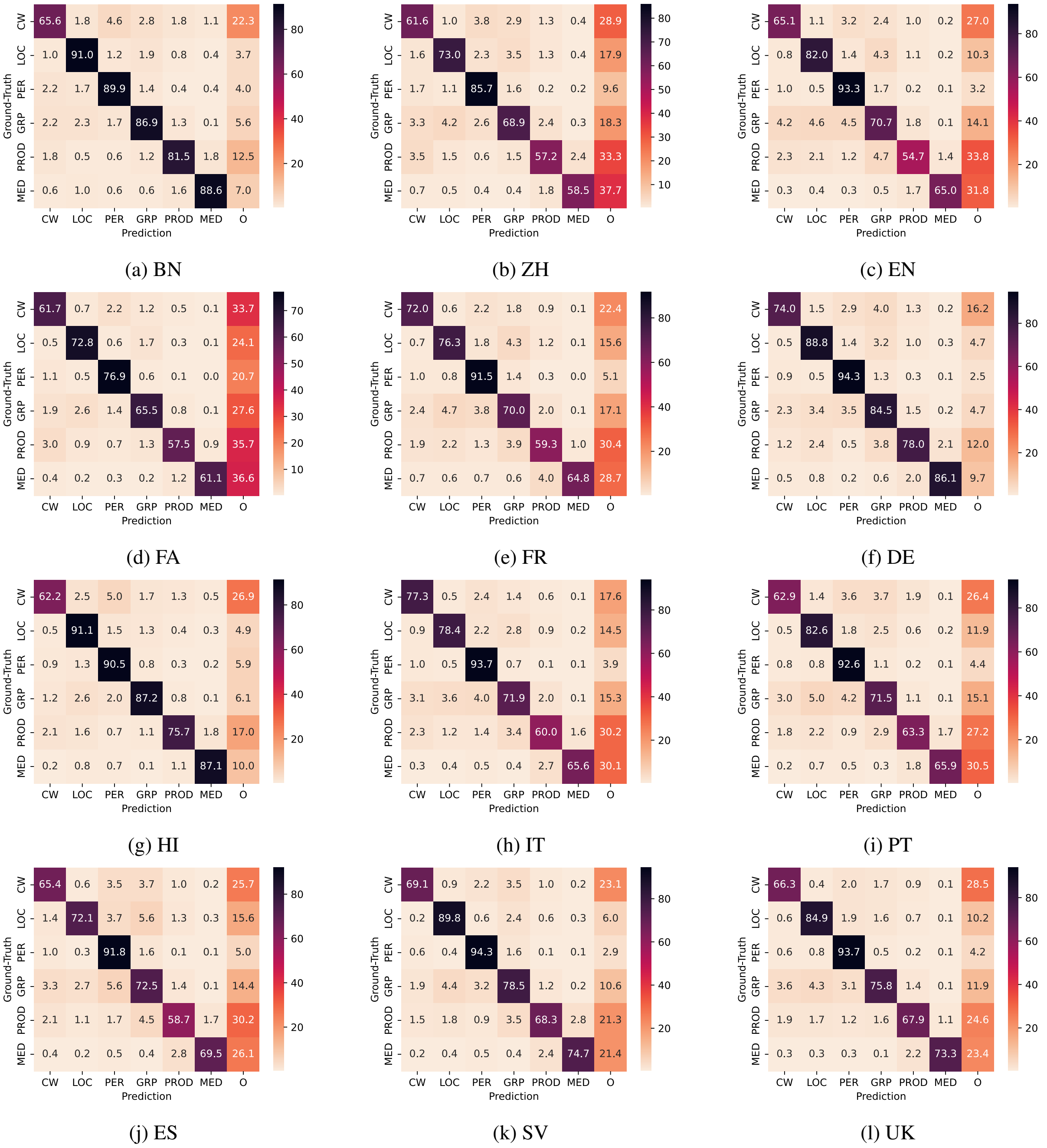
[9] https://creativecommons.org/licenses/by-sa/4.0

		DE	EN	ES	FR	SV	IT	ZH	PT	FA	BN	HI	UK	MULTI
MED	AnatomicalStructure	0.79	0.59	0.59	0.52	0.64	0.60	0.63	0.58	0.47	0.77	0.81	0.67	0.67
	Disease	0.81	0.61	0.62	0.59	0.65	0.56	0.66	0.63	0.54	0.82	0.85	0.66	0.68
	Medication/Vaccine	0.81	0.67	0.65	0.65	0.67	0.65	0.62	0.68	0.64	0.81	0.81	0.74	0.73
	MedicalProcedure	0.77	0.53	0.54	0.55	0.55	0.58	0.58	0.57	0.52	0.79	0.84	0.54	0.63
	Symptom	0.55	0.40	0.37	0.49	0.40	0.41	0.41	0.32	0.44	0.79	0.83	0.46	0.55
PER	Artist	0.74	0.72	0.69	0.76	0.72	0.81	0.68	0.72	0.72	0.69	0.71	0.69	0.76
	Athlete	0.75	0.73	0.68	0.70	0.66	0.83	0.72	0.64	0.49	0.71	0.77	0.76	0.74
	Cleric	0.56	0.44	0.50	0.54	0.50	0.61	0.39	0.55	0.42	0.71	0.65	0.54	0.56
	Politician	0.58	0.49	0.50	0.54	0.59	0.48	0.52	0.54	0.53	0.64	0.69	0.49	0.57
	Scientist	0.41	0.37	0.34	0.42	0.34	0.40	0.46	0.31	0.29	0.55	0.31	0.42	0.42
	SportsManager	0.58	0.49	0.54	0.51	0.36	0.63	0.54	0.47	0.48	0.65	0.25	0.57	0.58
	OtherPER	0.48	0.41	0.40	0.43	0.47	0.42	0.44	0.45	0.38	0.52	0.54	0.44	0.48
LOC	Facility	0.69	0.58	0.53	0.60	0.67	0.66	0.60	0.57	0.53	0.70	0.70	0.57	0.64
	HumanSettlement	0.88	0.83	0.79	0.78	0.91	0.82	0.75	0.83	0.75	0.86	0.88	0.85	0.84
	OtherLOC	0.54	0.45	0.38	0.43	0.88	0.43	0.57	0.65	0.30	0.81	0.75	0.55	0.57
	Station	0.78	0.69	0.66	0.71	0.73	0.66	0.82	0.68	0.75	0.87	0.83	0.70	0.73
PROD	Food	0.67	0.47	0.50	0.49	0.58	0.48	0.57	0.53	0.52	0.65	0.75	0.55	0.57
	Drink	0.66	0.45	0.55	0.50	0.59	0.53	0.43	0.58	0.48	0.82	0.87	0.56	0.60
	Software	0.77	0.59	0.72	0.64	0.70	0.68	0.55	0.70	0.60	0.82	0.82	0.73	0.74
	Vehicle	0.68	0.42	0.48	0.42	0.54	0.48	0.59	0.43	0.49	0.68	0.76	0.55	0.54
	Clothing	0.46	0.49	0.40	0.48	0.49	0.37	0.48	0.36	0.27	0.34	0.84	0.44	0.51
	OtherPROD	0.62	0.40	0.44	0.49	0.58	0.48	0.52	0.58	0.53	0.58	0.65	0.51	0.57
GRP	MusicalGRP	0.73	0.57	0.63	0.64	0.70	0.76	0.61	0.67	0.60	0.68	0.87	0.74	0.72
	PrivateCORP	0.64	0.19	0.33	0.38	0.14	0.36	0.70	0.00	0.26	0.95	0.91	0.13	0.54
	CarManufacturer	0.69	0.51	0.59	0.57	0.55	0.62	0.59	0.52	0.64	0.70	0.76	0.65	0.63
	AerospaceManufacturer	0.69	0.41	0.32	0.44	0.16	0.21	0.56	0.16	0.62	0.08	0.17	0.26	0.54
	PublicCORP	0.61	0.49	0.61	0.55	0.53	0.59	0.51	0.70	0.55	0.70	0.80	0.67	0.67
	ORG	0.67	0.52	0.54	0.52	0.59	0.52	0.61	0.58	0.55	0.82	0.84	0.62	0.61
	SportsGRP	0.86	0.73	0.70	0.71	0.78	0.75	0.78	0.74	0.82	0.90	0.94	0.81	0.78
CW	ArtWork	0.46	0.38	0.25	0.37	0.25	0.50	0.46	0.07	0.10	0.06	0.12	0.22	0.43
	VisualWork	0.73	0.61	0.62	0.79	0.71	0.87	0.63	0.64	0.73	0.71	0.74	0.69	0.77
	WrittenWork	0.76	0.58	0.59	0.66	0.61	0.56	0.64	0.56	0.48	0.74	0.76	0.60	0.64
	MusicalWork	0.77	0.67	0.59	0.59	0.67	0.76	0.54	0.67	0.53	0.61	0.71	0.55	0.70
	macro-F1	0.67	0.53	0.53	0.56	0.57	0.58	0.58	0.54	0.52	0.68	0.72	0.57	0.63
	micro-F1	0.71	0.61	0.61	0.64	0.70	0.70	0.63	0.65	0.61	0.74	0.77	0.66	0.69

Table 10: XLM-RoBERTa baseline results on the MultiCoNER v2 test set as measured by the F1 score for the different NER tags.

		EN		ES		FR		SV		IT		ZH		PT	
		clean	noisy	clean	noisy	clean	noisy	clean	noisy	clean	noisy	clean	noisy	clean	noisy
MED	AnatomicalStructure	0.61	0.55	0.62	0.51	0.54	0.47	0.67	0.57	0.62	0.56	0.70	0.44	0.60	0.53
	Disease	0.63	0.55	0.65	0.54	0.60	0.55	0.69	0.57	0.57	0.53	0.73	0.50	0.65	0.60
	Medication/Vaccine	0.70	0.62	0.68	0.58	0.68	0.56	0.71	0.57	0.68	0.57	0.67	0.54	0.72	0.61
	MedicalProcedure	0.55	0.49	0.57	0.47	0.58	0.48	0.60	0.42	0.61	0.52	0.65	0.44	0.59	0.53
	Symptom	0.43	0.32	0.41	0.26	0.54	0.35	0.45	0.30	0.42	0.37	0.52	0.23	0.35	0.22
PER	Artist	0.74	0.69	0.70	0.65	0.77	0.73	0.73	0.69	0.82	0.78	0.73	0.58	0.73	0.68
	Athlete	0.74	0.70	0.69	0.66	0.71	0.66	0.67	0.64	0.84	0.80	0.76	0.63	0.66	0.61
	Cleric	0.46	0.39	0.52	0.47	0.55	0.51	0.51	0.47	0.63	0.58	0.46	0.26	0.58	0.48
	Politician	0.50	0.46	0.52	0.46	0.55	0.51	0.60	0.56	0.50	0.45	0.57	0.42	0.56	0.50
	Scientist	0.37	0.37	0.35	0.31	0.43	0.41	0.35	0.31	0.41	0.38	0.50	0.35	0.32	0.27
	SportsManager	0.50	0.46	0.55	0.53	0.52	0.48	0.36	0.35	0.64	0.61	0.62	0.36	0.48	0.44
	OtherPER	0.43	0.38	0.41	0.37	0.45	0.40	0.49	0.44	0.43	0.40	0.49	0.33	0.46	0.43
LOC	Facility	0.62	0.50	0.57	0.45	0.63	0.53	0.71	0.57	0.69	0.60	0.65	0.51	0.60	0.49
	HumanSettlement	0.85	0.77	0.82	0.73	0.81	0.72	0.93	0.87	0.85	0.76	0.80	0.62	0.85	0.77
	Station	0.73	0.61	0.68	0.59	0.75	0.62	0.78	0.60	0.70	0.58	0.88	0.67	0.71	0.60
	OtherLOC	0.49	0.36	0.41	0.29	0.48	0.32	0.96	0.65	0.47	0.32	0.60	0.51	0.69	0.55
PROD	Food	0.49	0.43	0.53	0.42	0.51	0.45	0.62	0.49	0.49	0.44	0.66	0.38	0.56	0.44
	Drink	0.48	0.38	0.60	0.43	0.55	0.37	0.63	0.48	0.58	0.41	0.52	0.31	0.64	0.38
	Software	0.62	0.53	0.76	0.63	0.68	0.57	0.74	0.61	0.70	0.61	0.59	0.48	0.74	0.61
	Vehicle	0.44	0.36	0.49	0.43	0.44	0.37	0.57	0.45	0.51	0.43	0.68	0.36	0.47	0.33
	Clothing	0.51	0.44	0.43	0.34	0.50	0.43	0.53	0.39	0.40	0.29	0.61	0.30	0.42	0.23
	OtherPROD	0.42	0.36	0.47	0.38	0.51	0.43	0.61	0.50	0.50	0.44	0.58	0.41	0.61	0.52
GRP	MusicalGRP	0.60	0.50	0.66	0.55	0.68	0.57	0.73	0.63	0.78	0.71	0.67	0.49	0.69	0.61
	PrivateCORP	0.20	0.16	0.31	0.36	0.40	0.32	0.11	0.19	0.36	0.33	0.77	0.48	0.00	0.00
	CarManufacturer	0.54	0.42	0.63	0.51	0.61	0.47	0.58	0.48	0.66	0.52	0.65	0.51	0.54	0.46
	AerospaceManufacturer	0.45	0.32	0.37	0.22	0.48	0.36	0.18	0.12	0.23	0.17	0.00	0.00	0.20	0.09
	PublicCORP	0.52	0.41	0.65	0.51	0.58	0.48	0.57	0.45	0.63	0.51	0.60	0.40	0.73	0.63
	ORG	0.57	0.42	0.58	0.45	0.57	0.44	0.64	0.49	0.55	0.45	0.68	0.50	0.62	0.48
	SportsGRP	0.77	0.66	0.74	0.62	0.76	0.61	0.81	0.70	0.77	0.69	0.84	0.67	0.77	0.68
CW	ArtWork	0.39	0.37	0.26	0.22	0.39	0.31	0.28	0.19	0.54	0.41	0.00	0.08	0.07	0.05
	VisualWork	0.63	0.55	0.65	0.57	0.81	0.75	0.74	0.64	0.89	0.84	0.67	0.55	0.67	0.59
	WrittenWork	0.61	0.51	0.62	0.53	0.70	0.57	0.65	0.52	0.58	0.50	0.69	0.55	0.59	0.47
	MusicalWork	0.69	0.63	0.61	0.53	0.62	0.55	0.70	0.62	0.78	0.72	0.57	0.49	0.69	0.61

Table 11: XLM-RoBERTa baseline results on the MultiCoNER v2 test set as measured by the *macro* F1 score for the different NER tags.



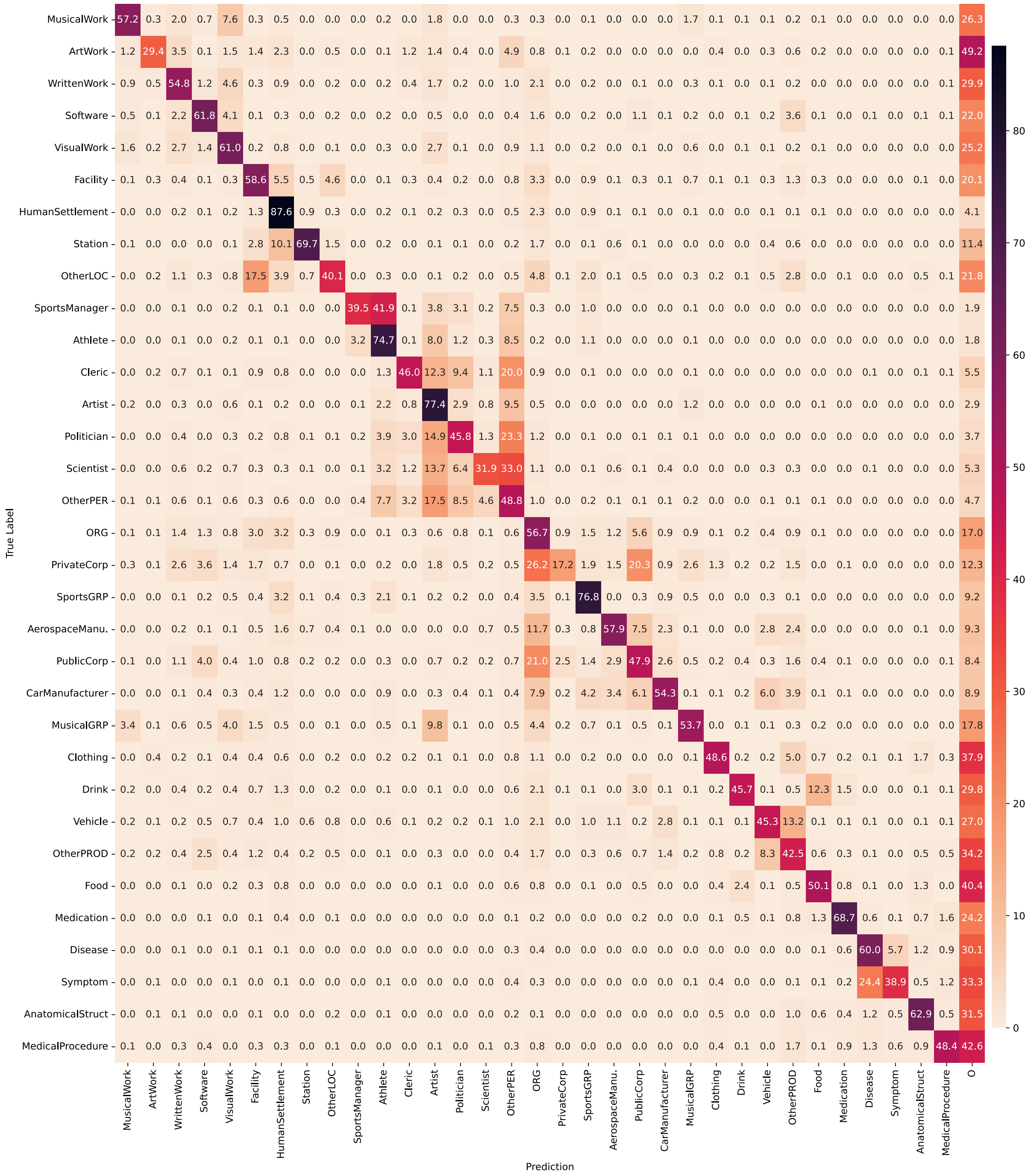
(a) BN (b) ZH (c) EN

(d) FA (e) FR (f) DE

(g) HI (h) IT (i) PT

(j) ES (k) SV (l) UK

Figure 6: Error analysis for the baseline approach. The plots on the x-axis show the predict coarse class, while the y-axis shows the ground-truth class. The scores represent the ratio of entities of a specific class and its predicted class. The diagonal represent correct class predictions.



	MusicalWork	ArtWork	WrittenWork	Software	VisualWork	Facility	HumanSettlement	Station	OtherLOC	SportsManager	Athlete	Cleric	Artist	Politician	Scientist	OtherPER	ORG	PrivateCorp	SportsGRP	AerospaceManu.	PublicCorp	CarManufacturer	MusicalGRP	Clothing	Drink	Vehicle	OtherPROD	Food	Medication	Disease	Symptom	AnatomicalStruct	MedicalProcedure	O
MusicalWork	57.2	0.3	2.0	0.7	7.6	0.3	0.5	0.0	0.0	0.0	0.2	0.0	1.8	0.0	0.0	0.3	0.3	0.0	0.1	0.0	0.0	0.0	1.7	0.1	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0	0.0	26.3
ArtWork	1.2	29.4	3.5	0.1	1.5	1.4	2.3	0.0	0.5	0.0	0.1	1.2	1.4	0.4	0.0	4.9	0.8	0.1	0.2	0.0	0.0	0.0	0.0	0.4	0.0	0.3	0.6	0.2	0.0	0.0	0.0	0.0	0.1	49.2
WrittenWork	0.9	0.5	54.8	1.2	4.6	0.3	0.9	0.0	0.2	0.0	0.2	0.4	1.7	0.2	0.0	1.0	2.1	0.0	0.2	0.0	0.1	0.0	0.3	0.1	0.0	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.1	29.9
Software	0.5	0.1	2.2	61.8	4.1	0.1	0.3	0.0	0.2	0.0	0.2	0.0	0.5	0.0	0.0	0.4	1.6	0.0	0.2	0.0	1.1	0.1	0.2	0.0	0.1	0.2	3.6	0.1	0.0	0.1	0.0	0.1	0.1	22.0
VisualWork	1.6	0.2	2.7	1.4	61.0	0.2	0.8	0.0	0.1	0.0	0.3	0.0	2.7	0.1	0.0	0.9	1.1	0.0	0.2	0.0	0.1	0.0	0.6	0.0	0.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0	0.0	25.2
Facility	0.1	0.3	0.4	0.1	0.3	58.6	5.5	0.5	4.6	0.0	0.1	0.3	0.4	0.2	0.0	0.8	3.3	0.0	0.9	0.1	0.3	0.1	0.7	0.1	0.1	0.3	1.3	0.3	0.0	0.0	0.0	0.1	0.0	20.1
HumanSettlement	0.0	0.0	0.2	0.1	0.2	1.3	87.6	0.9	0.3	0.0	0.2	0.1	0.2	0.3	0.0	0.5	2.3	0.0	0.9	0.1	0.1	0.0	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	4.1
Station	0.1	0.0	0.0	0.0	0.1	2.8	10.1	69.7	1.5	0.0	0.2	0.0	0.1	0.1	0.0	0.2	1.7	0.0	0.1	0.6	0.1	0.0	0.0	0.0	0.0	0.4	0.6	0.0	0.0	0.0	0.0	0.0	0.0	11.4
OtherLOC	0.0	0.2	1.1	0.3	0.8	17.5	3.9	0.7	40.1	0.0	0.3	0.0	0.1	0.2	0.0	0.5	4.8	0.1	2.0	0.1	0.5	0.0	0.3	0.2	0.1	0.5	2.8	0.0	0.1	0.0	0.0	0.5	0.1	21.8
SportsManager	0.0	0.0	0.1	0.0	0.2	0.1	0.1	0.0	0.0	39.5	41.9	0.1	3.8	3.1	0.2	7.5	0.3	0.0	1.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.9
Athlete	0.0	0.0	0.1	0.0	0.2	0.1	0.1	0.0	0.0	3.2	74.7	0.1	8.0	1.2	0.3	8.5	0.2	0.0	1.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.8
Cleric	0.0	0.2	0.7	0.1	0.1	0.9	0.8	0.0	0.0	0.0	1.3	46.0	12.3	9.4	1.1	20.0	0.9	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	5.5
Artist	0.2	0.0	0.3	0.0	0.6	0.1	0.2	0.0	0.0	0.1	2.2	0.8	77.4	2.9	0.8	9.5	0.5	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	2.9
Politician	0.0	0.0	0.4	0.0	0.3	0.2	0.8	0.1	0.1	0.2	3.9	3.0	14.9	45.8	1.3	23.3	1.2	0.0	0.1	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.7
Scientist	0.0	0.0	0.6	0.2	0.7	0.3	0.3	0.1	0.0	0.1	3.2	1.2	13.7	6.4	31.9	33.0	1.1	0.0	0.1	0.6	0.1	0.4	0.0	0.0	0.0	0.3	0.3	0.0	0.0	0.1	0.0	0.0	0.0	5.3
OtherPER	0.1	0.1	0.6	0.1	0.6	0.3	0.6	0.0	0.0	0.4	7.7	3.2	17.5	8.5	4.6	48.8	1.0	0.0	0.2	0.1	0.1	0.1	0.2	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	4.7
ORG	0.1	0.1	1.4	1.3	0.8	3.0	3.2	0.3	0.9	0.0	0.1	0.3	0.6	0.8	0.1	0.6	56.7	0.9	1.5	1.2	5.6	0.9	0.9	0.1	0.2	0.4	0.9	0.1	0.0	0.0	0.0	0.0	0.0	17.0
PrivateCorp	0.3	0.1	2.6	3.6	1.4	1.7	0.7	0.0	0.1	0.0	0.2	0.0	1.8	0.5	0.2	0.5	26.2	17.2	1.9	1.5	20.3	0.9	2.6	1.3	0.2	0.2	1.5	0.0	0.1	0.0	0.0	0.0	0.0	12.3
SportsGRP	0.0	0.0	0.1	0.2	0.5	0.4	3.2	0.1	0.4	0.3	2.1	0.1	0.2	0.2	0.0	0.4	3.5	0.1	76.8	0.0	0.3	0.9	0.5	0.0	0.0	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	9.2
AerospaceManu.	0.0	0.0	0.2	0.1	0.1	0.5	1.6	0.7	0.4	0.1	0.0	0.0	0.0	0.0	0.7	0.5	11.7	0.3	0.8	57.9	7.5	2.3	0.1	0.0	0.0	2.8	2.4	0.0	0.0	0.0	0.0	0.1	0.0	9.3
PublicCorp	0.1	0.0	1.1	4.0	0.4	1.0	0.8	0.2	0.2	0.0	0.3	0.0	0.7	0.2	0.2	0.7	21.0	2.5	1.4	2.9	47.9	2.6	0.5	0.2	0.4	0.3	1.6	0.4	0.1	0.0	0.0	0.0	0.1	8.4
CarManufacturer	0.0	0.0	0.1	0.4	0.3	0.4	1.2	0.0	0.0	0.0	0.9	0.0	0.3	0.4	0.1	0.4	7.9	0.2	4.2	3.4	6.1	54.3	0.1	0.1	0.2	6.0	3.9	0.1	0.1	0.0	0.0	0.0	0.0	8.9
MusicalGRP	3.4	0.1	0.6	0.5	4.0	1.5	0.5	0.0	0.1	0.0	0.5	0.1	9.8	0.1	0.0	0.5	4.4	0.2	0.7	0.1	0.5	0.1	53.7	0.0	0.1	0.1	0.3	0.2	0.0	0.0	0.0	0.0	0.0	17.8
Clothing	0.0	0.4	0.2	0.1	0.4	0.4	0.6	0.0	0.2	0.0	0.2	0.2	0.1	0.1	0.0	0.8	1.1	0.0	0.2	0.0	0.0	0.0	0.1	48.6	0.2	0.2	5.0	0.7	0.2	0.1	0.1	1.7	0.3	37.9
Drink	0.2	0.0	0.4	0.2	0.4	0.7	1.3	0.0	0.2	0.0	0.1	0.0	0.1	0.0	0.0	0.6	2.1	0.1	0.1	0.0	3.0	0.1	0.1	0.2	45.7	0.1	0.5	12.3	1.5	0.0	0.0	0.1	0.1	29.8
Vehicle	0.2	0.1	0.2	0.5	0.7	0.4	1.0	0.6	0.8	0.0	0.6	0.1	0.2	0.2	0.1	1.0	2.1	0.0	1.0	1.1	0.2	2.8	0.1	0.1	0.1	45.3	13.2	0.1	0.1	0.1	0.0	0.1	0.1	27.0
OtherPROD	0.2	0.2	0.4	2.5	0.4	1.2	0.4	0.2	0.5	0.0	0.1	0.0	0.3	0.0	0.0	0.4	1.7	0.0	0.3	0.6	0.7	1.4	0.2	0.8	0.2	8.3	42.5	0.6	0.3	0.1	0.0	0.5	0.5	34.2
Food	0.0	0.0	0.1	0.0	0.2	0.3	0.8	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.6	0.8	0.0	0.1	0.0	0.5	0.0	0.0	0.4	2.4	0.1	0.5	50.1	0.8	0.1	0.0	1.3	0.0	40.4
Medication	0.0	0.0	0.0	0.1	0.0	0.1	0.4	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.5	0.1	0.8	1.3	68.7	0.6	0.1	0.7	1.6	24.2
Disease	0.0	0.0	0.1	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.6	60.0	5.7	1.2	0.9	30.1
Symptom	0.0	0.1	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.4	0.3	0.0	0.0	0.0	0.0	0.0	0.1	0.4	0.0	0.0	0.1	0.1	0.2	24.4	38.9	0.5	1.2	33.3
AnatomicalStruct	0.0	0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	1.0	0.6	0.4	1.2	0.5	62.9	0.5	31.5
MedicalProcedure	0.1	0.0	0.3	0.4	0.0	0.3	0.3	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.1	0.3	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.1	0.0	1.7	0.1	0.9	1.3	0.6	0.9	48.4	42.6

Figure 7: Fine-grained NER predictions for the XLM-RoBERTa baseline on the `EN` test set.

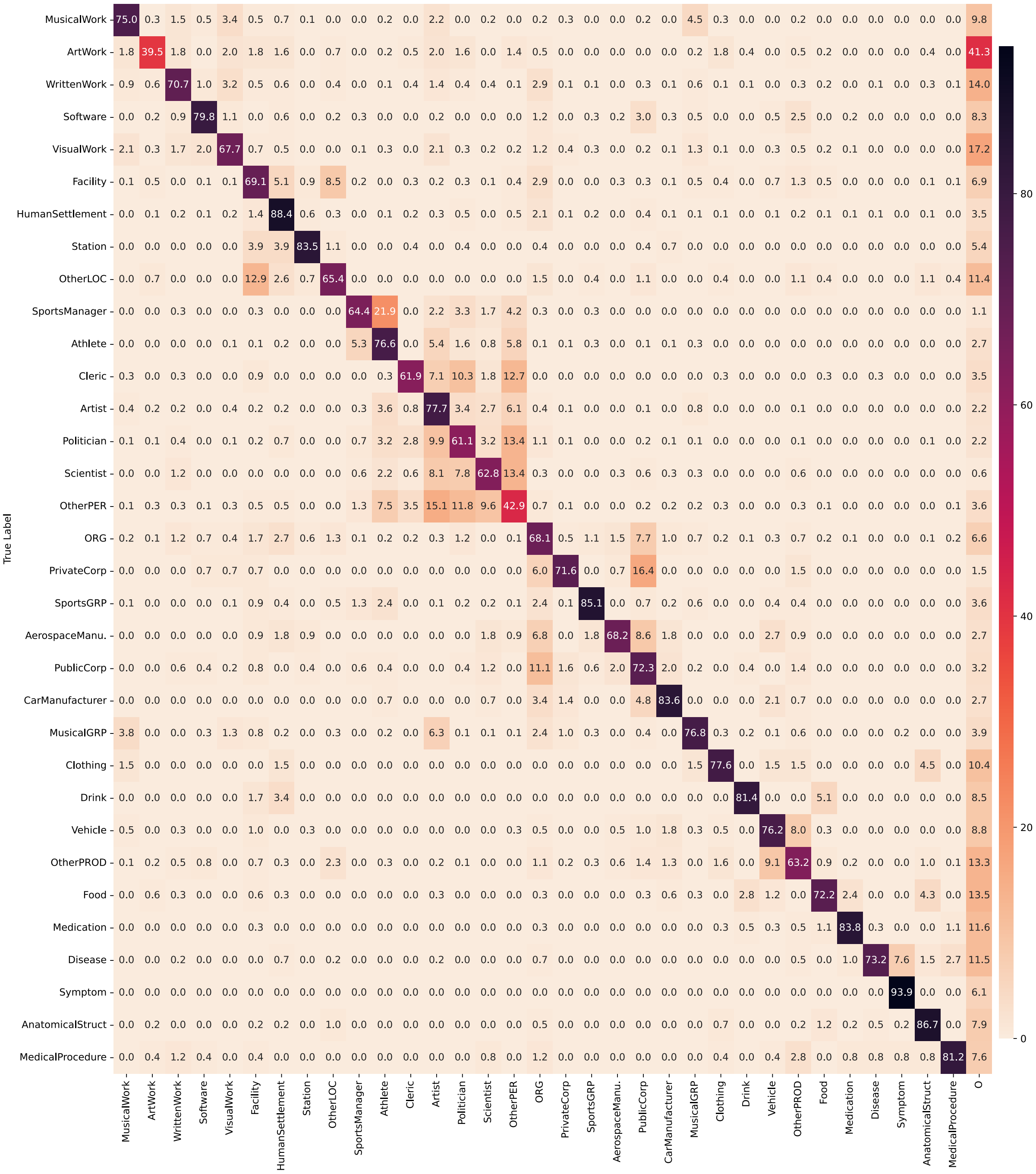
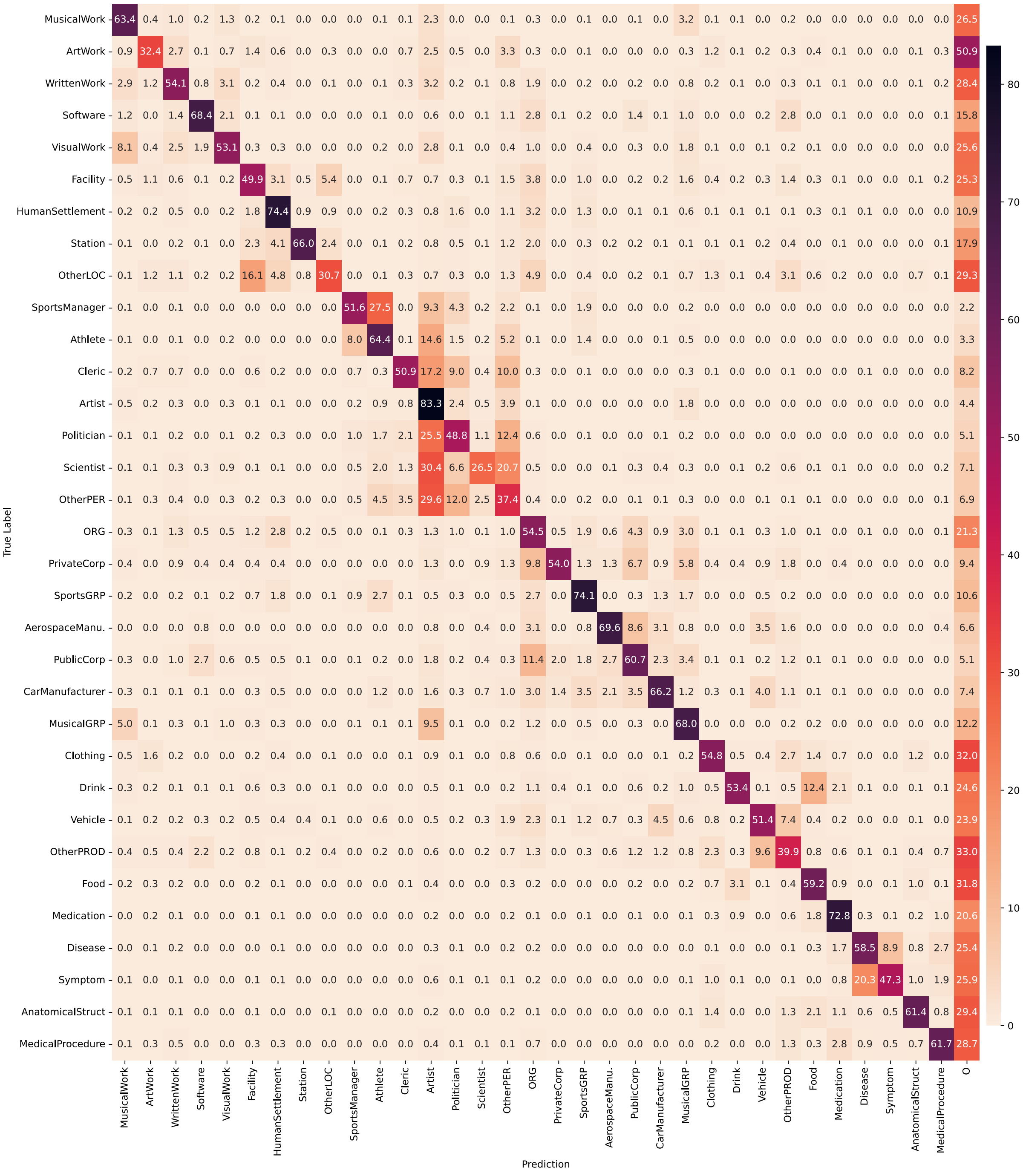


Figure 8: Fine-grained NER predictions for the XLM-RoBERTa baseline on the DE test set.



	MusicalWork	ArtWork	WrittenWork	Software	VisualWork	Facility	HumanSettlement	Station	OtherLOC	SportsManager	Athlete	Cleric	Artist	Politician	Scientist	OtherPER	ORG	PrivateCorp	SportsGRP	AerospaceManu.	PublicCorp	CarManufacturer	MusicalGRP	Clothing	Drink	Vehicle	OtherPROD	Food	Medication	Disease	Symptom	AnatomicalStruct	MedicalProcedure	O
MusicalWork	63.4	0.4	1.0	0.2	1.3	0.2	0.1	0.0	0.0	0.0	0.1	0.1	2.3	0.0	0.0	0.1	0.3	0.0	0.1	0.0	0.1	0.0	3.2	0.1	0.1	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	26.5
ArtWork	0.9	32.4	2.7	0.1	0.7	1.4	0.6	0.0	0.3	0.0	0.0	0.7	2.5	0.5	0.0	3.3	0.3	0.0	0.1	0.0	0.0	0.0	0.3	1.2	0.1	0.2	0.3	0.4	0.1	0.0	0.0	0.1	0.3	50.9
WrittenWork	2.9	1.2	54.1	0.8	3.1	0.2	0.4	0.0	0.1	0.0	0.1	0.3	3.2	0.2	0.1	0.8	1.9	0.0	0.2	0.0	0.2	0.0	0.8	0.2	0.1	0.0	0.3	0.1	0.1	0.0	0.0	0.1	0.2	28.4
Software	1.2	0.0	1.4	68.4	2.1	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.6	0.0	0.1	1.1	2.8	0.1	0.2	0.0	1.4	0.1	1.0	0.0	0.0	0.2	2.8	0.0	0.1	0.0	0.0	0.0	0.1	15.8
VisualWork	8.1	0.4	2.5	1.9	53.1	0.3	0.3	0.0	0.0	0.0	0.2	0.0	2.8	0.1	0.0	0.4	1.0	0.0	0.4	0.0	0.3	0.0	1.8	0.1	0.0	0.1	0.2	0.1	0.0	0.0	0.0	0.0	0.0	25.6
Facility	0.5	1.1	0.6	0.1	0.2	49.9	3.1	0.5	5.4	0.0	0.1	0.7	0.7	0.3	0.1	1.5	3.8	0.0	1.0	0.0	0.2	0.2	1.6	0.4	0.2	0.3	1.4	0.3	0.1	0.0	0.0	0.1	0.2	25.3
HumanSettlement	0.2	0.2	0.5	0.0	0.2	1.8	74.4	0.9	0.9	0.0	0.2	0.3	0.8	1.6	0.0	1.1	3.2	0.0	1.3	0.0	0.1	0.1	0.6	0.1	0.1	0.1	0.1	0.3	0.1	0.1	0.0	0.0	0.0	10.9
Station	0.1	0.0	0.2	0.1	0.0	2.3	4.1	66.0	2.4	0.0	0.1	0.2	0.8	0.5	0.1	1.2	2.0	0.0	0.3	0.2	0.2	0.1	0.1	0.1	0.1	0.2	0.4	0.0	0.1	0.0	0.0	0.0	0.1	17.9
OtherLOC	0.1	1.2	1.1	0.2	0.2	16.1	4.8	0.8	30.7	0.0	0.1	0.3	0.7	0.3	0.0	1.3	4.9	0.0	0.4	0.0	0.2	0.1	0.7	1.3	0.1	0.4	3.1	0.6	0.2	0.0	0.0	0.7	0.1	29.3
SportsManager	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	51.6	27.5	0.0	9.3	4.3	0.2	2.2	0.1	0.0	1.9	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2
Athlete	0.1	0.0	0.1	0.0	0.2	0.0	0.0	0.0	0.0	8.0	64.4	0.1	14.6	1.5	0.2	5.2	0.1	0.0	1.4	0.0	0.0	0.1	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.3
Cleric	0.2	0.7	0.7	0.0	0.0	0.6	0.2	0.0	0.0	0.7	0.3	50.9	17.2	9.0	0.4	10.0	0.3	0.0	0.1	0.0	0.0	0.0	0.3	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	8.2
Artist	0.5	0.2	0.3	0.0	0.3	0.1	0.1	0.0	0.0	0.2	0.9	0.8	83.3	2.4	0.5	3.9	0.1	0.0	0.0	0.0	0.0	0.0	1.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.4
Politician	0.1	0.1	0.2	0.0	0.1	0.2	0.3	0.0	0.0	1.0	1.7	2.1	25.5	48.8	1.1	12.4	0.6	0.0	0.1	0.0	0.0	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.1
Scientist	0.1	0.1	0.3	0.3	0.9	0.1	0.1	0.0	0.0	0.5	2.0	1.3	30.4	6.6	26.5	20.7	0.5	0.0	0.0	0.1	0.3	0.4	0.3	0.0	0.1	0.2	0.6	0.1	0.1	0.0	0.0	0.0	0.2	7.1
OtherPER	0.1	0.3	0.4	0.0	0.3	0.2	0.3	0.0	0.0	0.5	4.5	3.5	29.6	12.0	2.5	37.4	0.4	0.0	0.2	0.0	0.1	0.1	0.3	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.1	6.9
ORG	0.3	0.1	1.3	0.5	0.5	1.2	2.8	0.2	0.5	0.0	0.1	0.3	1.3	1.0	0.1	1.0	54.5	0.5	1.9	0.6	4.3	0.9	3.0	0.1	0.1	0.3	1.0	0.1	0.0	0.1	0.0	0.0	0.1	21.3
PrivateCorp	0.4	0.0	0.9	0.4	0.4	0.4	0.4	0.0	0.0	0.0	0.0	0.0	1.3	0.0	0.9	1.3	9.8	54.0	1.3	1.3	6.7	0.9	5.8	0.4	0.4	0.9	1.8	0.0	0.4	0.0	0.0	0.0	0.0	9.4
SportsGRP	0.2	0.0	0.2	0.1	0.2	0.7	1.8	0.0	0.1	0.9	2.7	0.1	0.5	0.3	0.0	0.5	2.7	0.0	74.1	0.0	0.3	1.3	1.7	0.0	0.0	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	10.6
AerospaceManu.	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.4	0.0	3.1	0.0	0.8	69.6	8.6	3.1	0.8	0.0	0.0	3.5	1.6	0.0	0.0	0.0	0.0	0.0	0.4	6.6
PublicCorp	0.3	0.0	1.0	2.7	0.6	0.5	0.5	0.1	0.0	0.1	0.2	0.0	1.8	0.2	0.4	0.3	11.4	2.0	1.8	2.7	60.7	2.3	3.4	0.1	0.1	0.2	1.2	0.1	0.1	0.0	0.0	0.0	0.0	5.1
CarManufacturer	0.3	0.1	0.1	0.1	0.0	0.3	0.5	0.0	0.0	0.0	1.2	0.0	1.6	0.3	0.7	1.0	3.0	1.4	3.5	2.1	3.5	66.2	1.2	0.3	0.1	4.0	1.1	0.1	0.1	0.0	0.0	0.0	0.0	7.4
MusicalGRP	5.0	0.1	0.3	0.1	1.0	0.3	0.3	0.0	0.0	0.1	0.1	0.1	9.5	0.1	0.0	0.2	1.2	0.0	0.5	0.0	0.3	0.0	68.0	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0	0.0	12.2
Clothing	0.5	1.6	0.2	0.0	0.0	0.2	0.4	0.0	0.1	0.0	0.0	0.1	0.9	0.1	0.0	0.8	0.6	0.0	0.1	0.0	0.0	0.1	0.2	54.8	0.5	0.4	2.7	1.4	0.7	0.0	0.0	1.2	0.0	32.0
Drink	0.3	0.2	0.1	0.1	0.1	0.6	0.3	0.0	0.1	0.0	0.0	0.0	0.5	0.1	0.0	0.2	1.1	0.4	0.1	0.0	0.6	0.2	1.0	0.5	53.4	0.1	0.5	12.4	2.1	0.1	0.0	0.1	0.1	24.6
Vehicle	0.1	0.2	0.2	0.3	0.2	0.5	0.4	0.4	0.1	0.0	0.6	0.0	0.5	0.2	0.3	1.9	2.3	0.1	1.2	0.7	0.3	4.5	0.6	0.8	0.2	51.4	7.4	0.4	0.2	0.0	0.0	0.1	0.0	23.9
OtherPROD	0.4	0.5	0.4	2.2	0.2	0.8	0.1	0.2	0.4	0.0	0.2	0.0	0.6	0.0	0.2	0.7	1.3	0.0	0.3	0.6	1.2	1.2	0.8	2.3	0.3	9.6	39.9	0.8	0.6	0.1	0.1	0.4	0.7	33.0
Food	0.2	0.3	0.2	0.0	0.0	0.2	0.1	0.0	0.0	0.0	0.0	0.1	0.4	0.0	0.0	0.3	0.2	0.0	0.0	0.0	0.2	0.0	0.2	0.7	3.1	0.1	0.4	59.2	0.9	0.0	0.1	1.0	0.1	31.8
Medication	0.0	0.2	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.2	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.3	0.9	0.0	0.6	1.8	72.8	0.3	0.1	0.2	1.0	20.6
Disease	0.0	0.1	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.3	0.1	0.0	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.3	1.7	58.5	8.9	0.8	2.7	25.4
Symptom	0.1	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.6	0.1	0.1	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.1	1.0	0.1	0.0	0.1	0.0	0.8	20.3	47.3	1.0	1.9	25.9
AnatomicalStruct	0.1	0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.0	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.1	1.4	0.0	0.0	1.3	2.1	1.1	0.6	0.5	61.4	0.8	29.4
MedicalProcedure	0.1	0.3	0.5	0.0	0.0	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.4	0.1	0.1	0.1	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	1.3	0.3	2.8	0.9	0.5	0.7	61.7	28.7

Figure 9: Fine-grained NER predictions for the XLM-RoBERTa baseline on the ES test set.

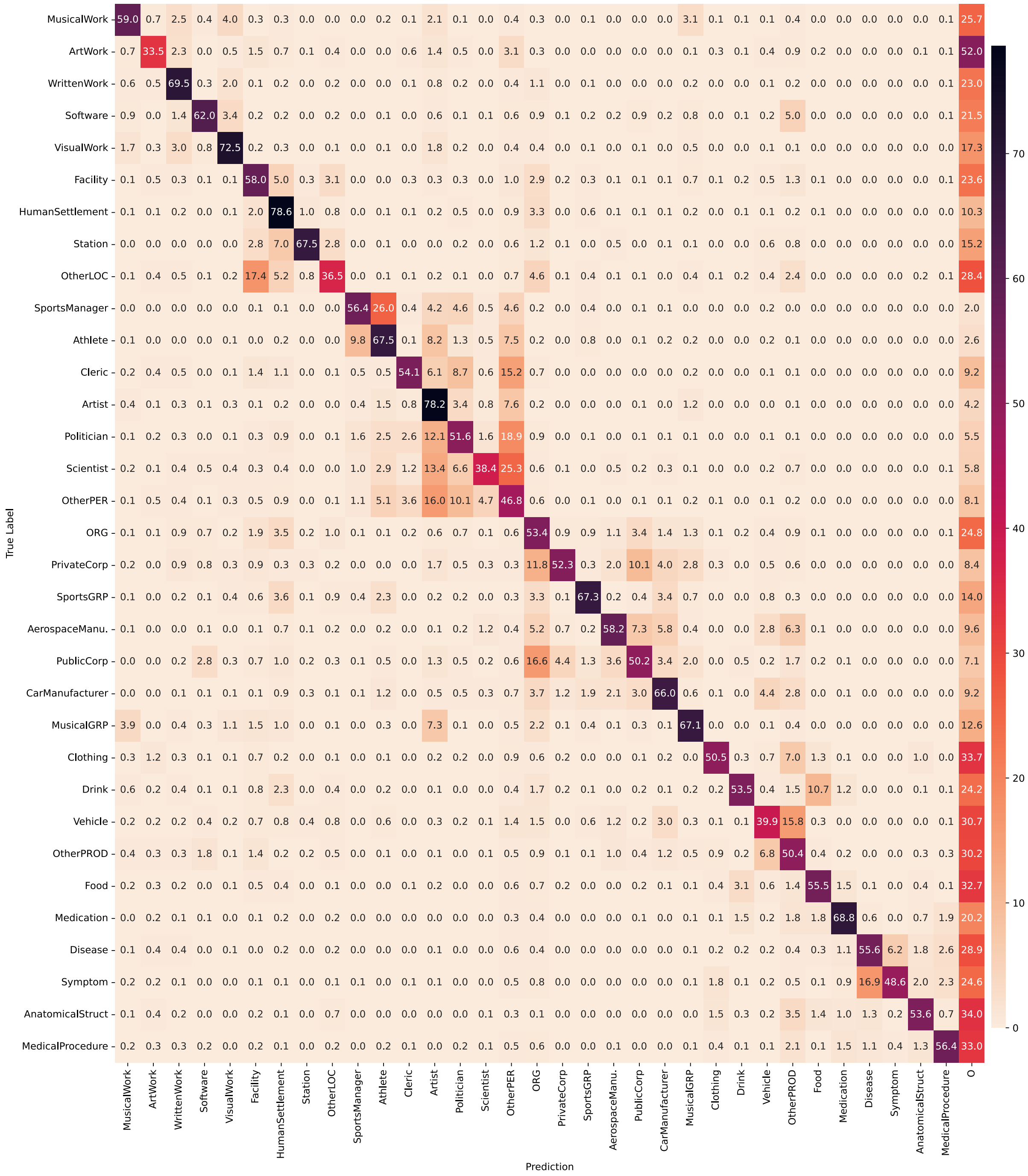


Figure 10: Fine-grained NER predictions for the XLM-RoBERTa baseline on the `FR` test set.

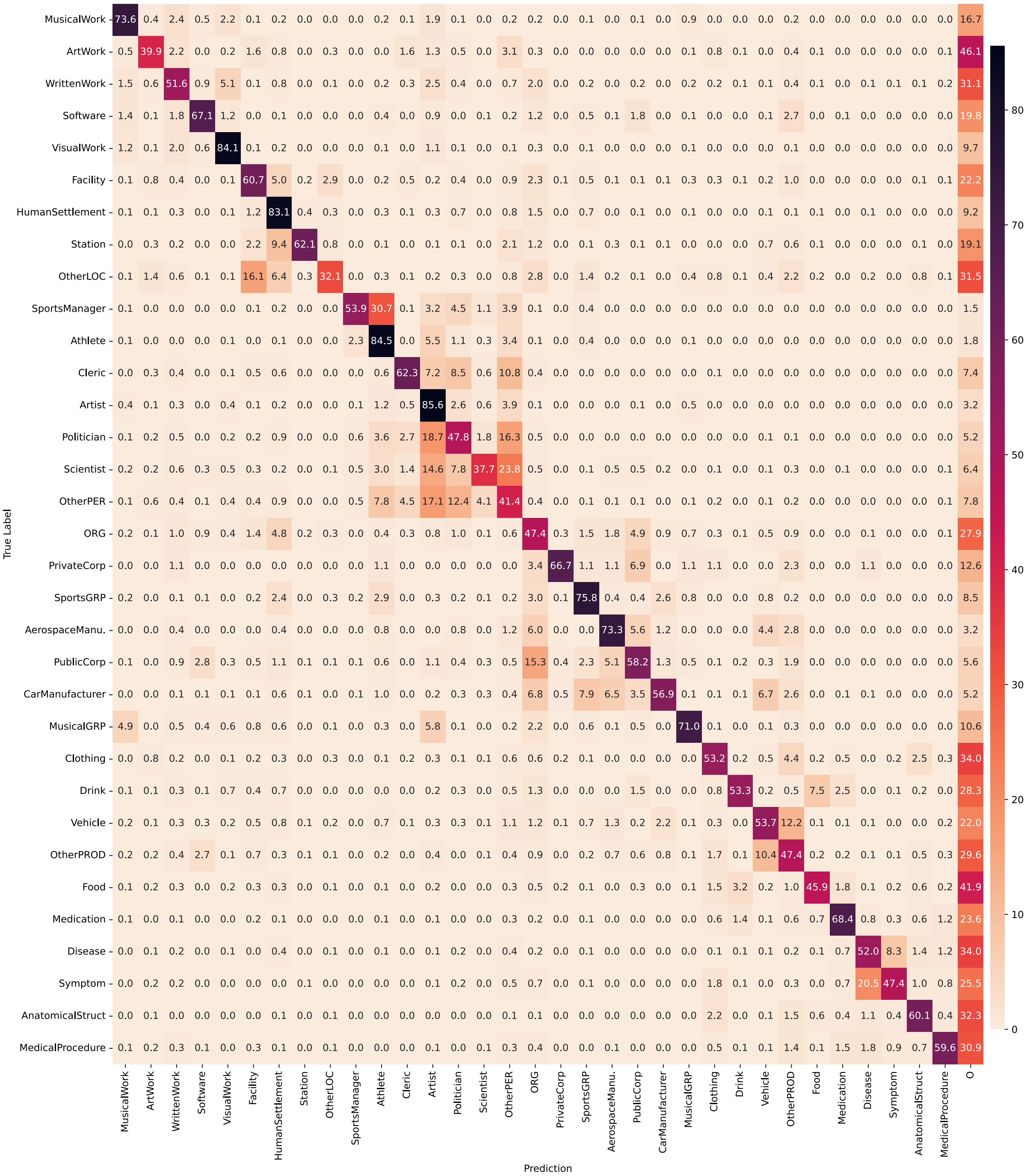


Figure 11: Fine-grained NER predictions for the XLM-RoBERTa baseline on the IT test set.

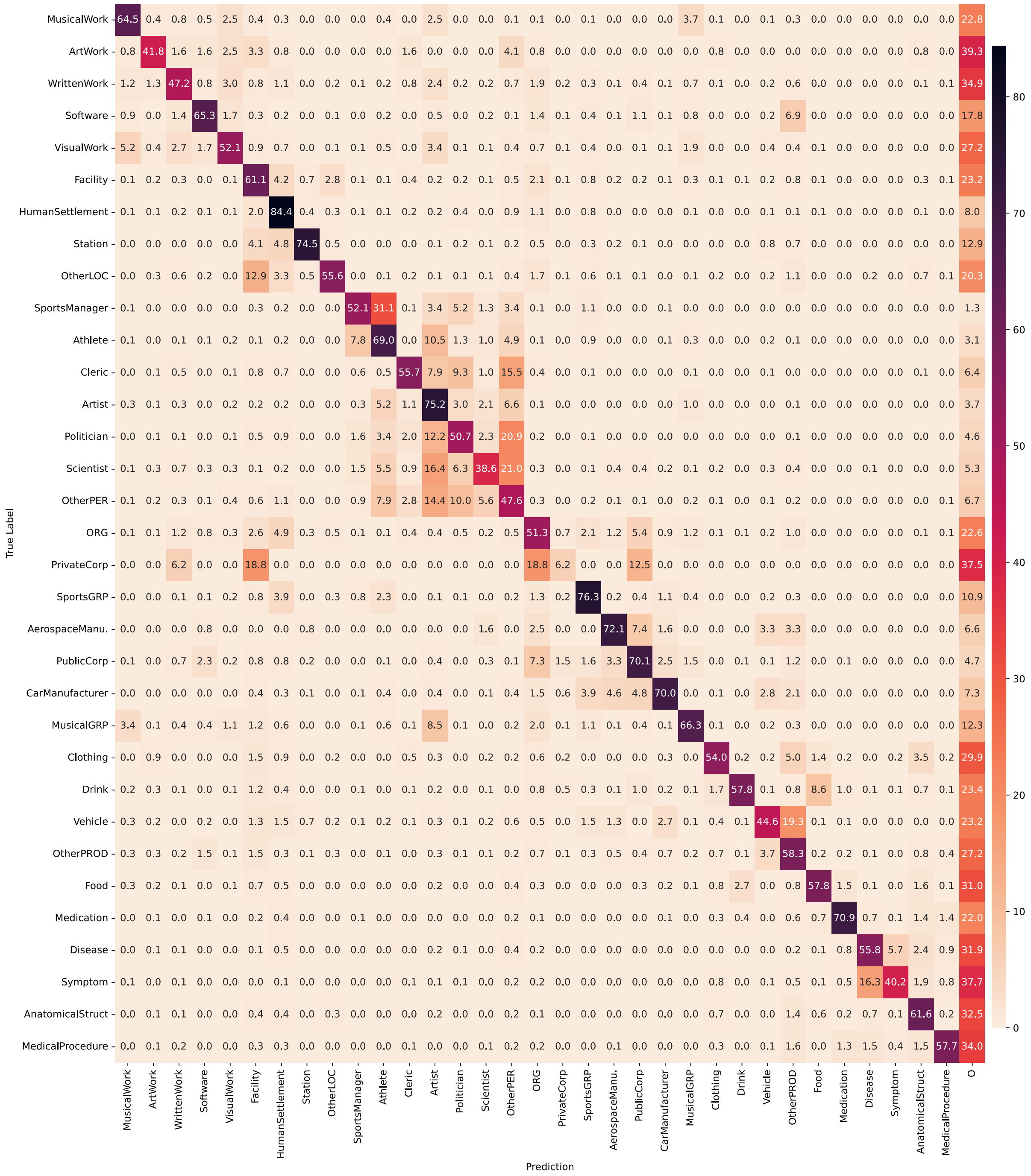


Figure 12: Fine-grained NER predictions for the XLM-RoBERTa baseline on the PT test set.

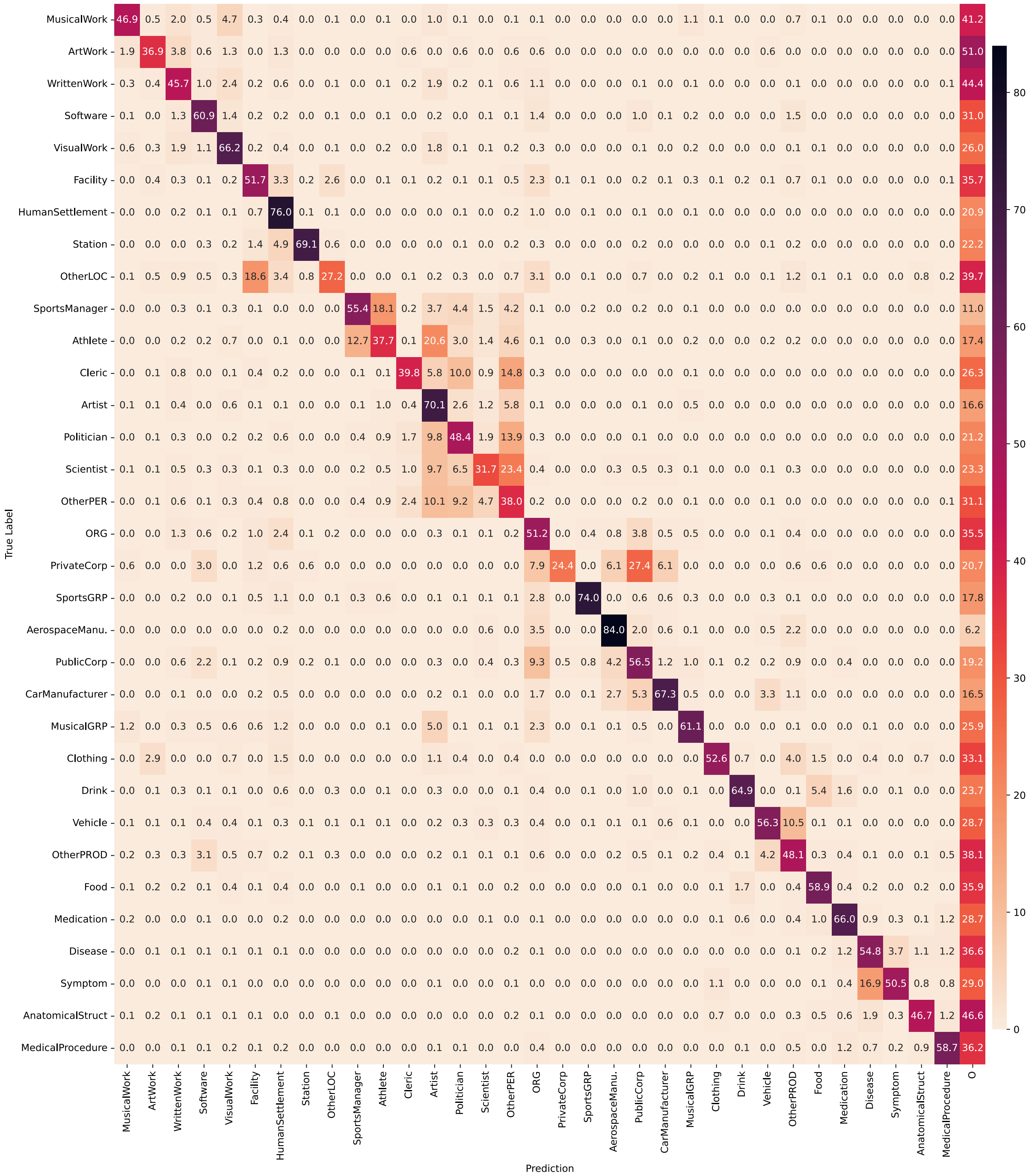
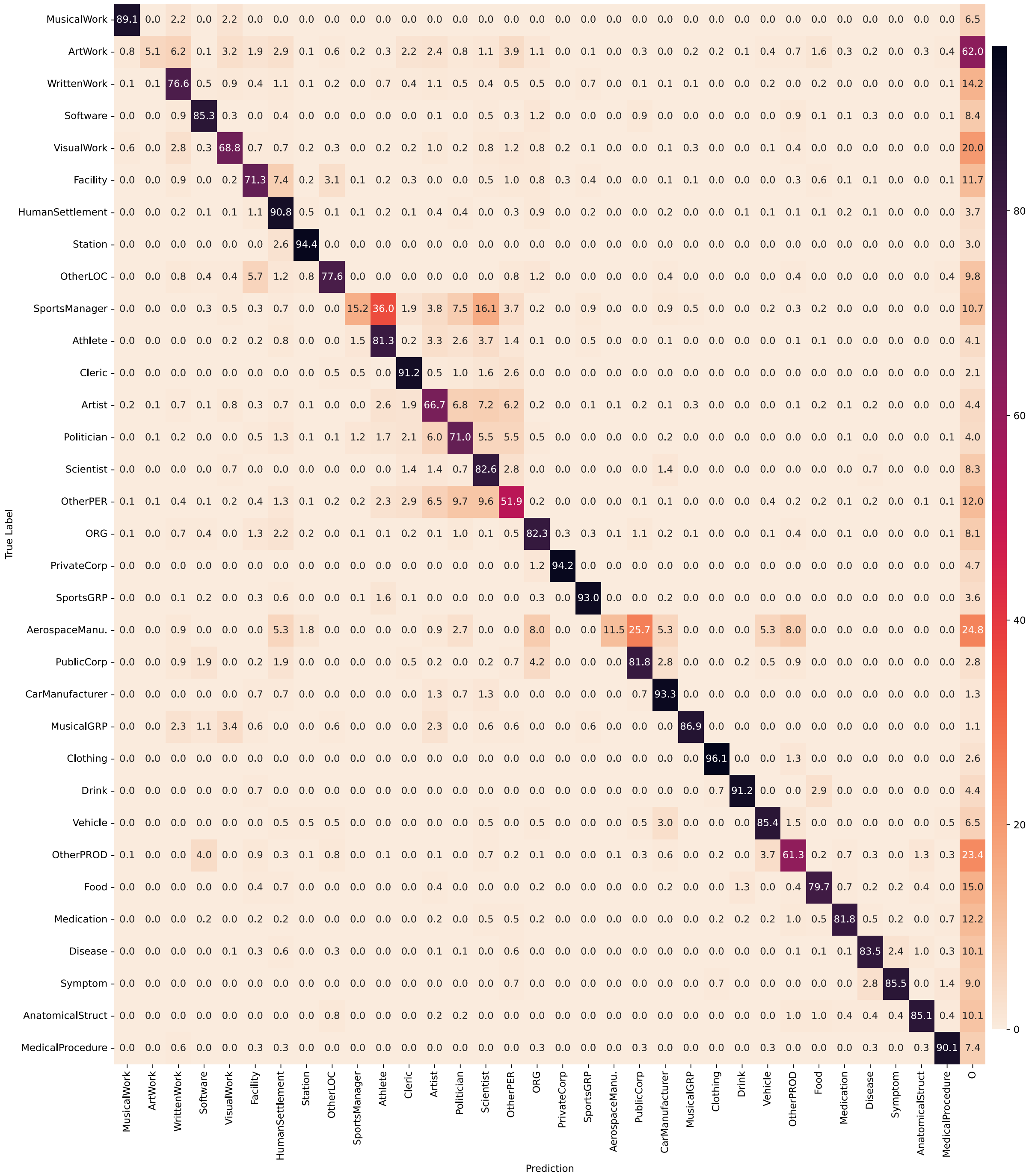


Figure 13: Fine-grained NER predictions for the XLM-RoBERTa baseline on the FA test set.



	MusicalWork	ArtWork	WrittenWork	Software	VisualWork	Facility	HumanSettlement	Station	OtherLOC	SportsManager	Athlete	Cleric	Artist	Politician	Scientist	OtherPER	ORG	PrivateCorp	SportsGRP	AerospaceManu.	PublicCorp	CarManufacturer	MusicalGRP	Clothing	Drink	Vehicle	OtherPROD	Food	Medication	Disease	Symptom	AnatomicalStruct	MedicalProcedure	O
MusicalWork	89.1	0.0	2.2	0.0	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.5
ArtWork	0.8	5.1	6.2	0.1	3.2	1.9	2.9	0.1	0.6	0.2	0.3	2.2	2.4	0.8	1.1	3.9	1.1	0.0	0.1	0.0	0.3	0.0	0.2	0.2	0.1	0.4	0.7	1.6	0.3	0.2	0.0	0.3	0.4	62.0
WrittenWork	0.1	0.1	76.6	0.5	0.9	0.4	1.1	0.1	0.2	0.0	0.7	0.4	1.1	0.5	0.4	0.5	0.5	0.0	0.7	0.0	0.1	0.1	0.1	0.0	0.0	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.1	14.2
Software	0.0	0.0	0.9	85.3	0.3	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.5	0.3	1.2	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.9	0.1	0.1	0.3	0.0	0.0	0.1	8.4
VisualWork	0.6	0.0	2.8	0.3	68.8	0.7	0.7	0.2	0.3	0.0	0.2	0.2	1.0	0.2	0.8	1.2	0.8	0.2	0.1	0.0	0.0	0.1	0.3	0.0	0.0	0.1	0.4	0.0	0.0	0.0	0.0	0.0	0.0	20.0
Facility	0.0	0.0	0.9	0.0	0.2	71.3	7.4	0.2	3.1	0.1	0.2	0.3	0.0	0.0	0.5	1.0	0.8	0.3	0.4	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.3	0.6	0.1	0.1	0.0	0.0	0.1	11.7
HumanSettlement	0.0	0.0	0.2	0.1	0.1	1.1	90.8	0.5	0.1	0.1	0.2	0.1	0.4	0.4	0.0	0.3	0.9	0.0	0.2	0.0	0.0	0.2	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.1	0.0	0.0	0.0	3.7
Station	0.0	0.0	0.0	0.0	0.0	0.0	2.6	94.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0
OtherLOC	0.0	0.0	0.8	0.4	0.4	5.7	1.2	0.8	77.6	0.0	0.0	0.0	0.0	0.0	0.0	0.8	1.2	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.4	9.8
SportsManager	0.0	0.0	0.0	0.3	0.5	0.3	0.7	0.0	0.0	15.2	36.0	1.9	3.8	7.5	16.1	3.7	0.2	0.0	0.9	0.0	0.0	0.9	0.5	0.0	0.0	0.2	0.3	0.2	0.0	0.0	0.0	0.0	0.0	10.7
Athlete	0.0	0.0	0.0	0.0	0.2	0.2	0.8	0.0	0.0	1.5	81.3	0.2	3.3	2.6	3.7	1.4	0.1	0.0	0.5	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	4.1
Cleric	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.5	0.0	91.2	0.5	1.0	1.6	2.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.1
Artist	0.2	0.1	0.7	0.1	0.8	0.3	0.7	0.1	0.0	0.0	2.6	1.9	66.7	6.8	7.2	6.2	0.2	0.0	0.1	0.1	0.2	0.1	0.3	0.0	0.0	0.0	0.1	0.2	0.1	0.2	0.0	0.0	0.0	4.4
Politician	0.0	0.1	0.2	0.0	0.0	0.5	1.3	0.1	0.1	1.2	1.7	2.1	6.0	71.0	5.5	5.5	0.5	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	4.0
Scientist	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	1.4	1.4	0.7	82.6	2.8	0.0	0.0	0.0	0.0	0.0	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	8.3
OtherPER	0.1	0.1	0.4	0.1	0.2	0.4	1.3	0.1	0.2	0.2	2.3	2.9	6.5	9.7	9.6	51.9	0.2	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.4	0.2	0.2	0.1	0.2	0.0	0.1	0.1	12.0
ORG	0.1	0.0	0.7	0.4	0.0	1.3	2.2	0.2	0.0	0.1	0.1	0.2	0.1	1.0	0.1	0.5	82.3	0.3	0.3	0.1	1.1	0.2	0.1	0.0	0.0	0.1	0.4	0.0	0.1	0.0	0.0	0.0	0.1	8.1
PrivateCorp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	94.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.7
SportsGRP	0.0	0.0	0.1	0.2	0.0	0.3	0.6	0.0	0.0	0.1	1.6	0.1	0.0	0.0	0.0	0.0	0.3	0.0	93.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.6
AerospaceManu.	0.0	0.0	0.9	0.0	0.0	0.0	5.3	1.8	0.0	0.0	0.0	0.0	0.9	2.7	0.0	0.0	8.0	0.0	0.0	11.5	25.7	5.3	0.0	0.0	0.0	5.3	8.0	0.0	0.0	0.0	0.0	0.0	0.0	24.8
PublicCorp	0.0	0.0	0.9	1.9	0.0	0.2	1.9	0.0	0.0	0.0	0.0	0.5	0.2	0.0	0.2	0.7	4.2	0.0	0.0	0.0	81.8	2.8	0.0	0.0	0.2	0.5	0.9	0.0	0.0	0.0	0.0	0.0	0.0	2.8
CarManufacturer	0.0	0.0	0.0	0.0	0.0	0.7	0.7	0.0	0.0	0.0	0.0	0.0	1.3	0.7	1.3	0.0	0.0	0.0	0.0	0.0	0.7	93.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.3
MusicalGRP	0.0	0.0	2.3	1.1	3.4	0.6	0.0	0.0	0.6	0.0	0.0	0.0	2.3	0.0	0.6	0.6	0.0	0.0	0.6	0.0	0.0	0.0	86.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1
Clothing	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	96.1	0.0	0.0	1.3	0.0	0.0	0.0	0.0	0.0	0.0	2.6
Drink	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	91.2	0.0	0.0	2.9	0.0	0.0	0.0	0.0	0.0	4.4
Vehicle	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.5	0.0	0.0	0.0	0.5	3.0	0.0	0.0	0.0	85.4	1.5	0.0	0.0	0.0	0.0	0.0	0.5	6.5
OtherPROD	0.1	0.0	0.0	4.0	0.0	0.9	0.3	0.1	0.8	0.0	0.1	0.0	0.1	0.0	0.7	0.2	0.1	0.0	0.0	0.1	0.3	0.6	0.0	0.2	0.0	3.7	61.3	0.2	0.7	0.3	0.0	1.3	0.3	23.4
Food	0.0	0.0	0.0	0.0	0.0	0.4	0.7	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.2	0.0	0.0	1.3	0.0	0.4	79.7	0.7	0.2	0.2	0.4	0.0	15.0
Medication	0.0	0.0	0.0	0.2	0.0	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.5	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.2	1.0	0.5	81.8	0.5	0.2	0.0	0.7	12.2
Disease	0.0	0.0	0.0	0.0	0.1	0.3	0.6	0.0	0.3	0.0	0.0	0.0	0.1	0.1	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	83.5	2.4	1.0	0.3	10.1
Symptom	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	2.8	85.5	0.0	1.4	9.0
AnatomicalStruct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.4	0.4	0.4	85.1	0.4	10.1
MedicalProcedure	0.0	0.0	0.6	0.0	0.0	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.3	0.0	0.3	90.1	7.4

Figure 14: Fine-grained NER predictions for the XLM-RoBERTa baseline on the HI test set.

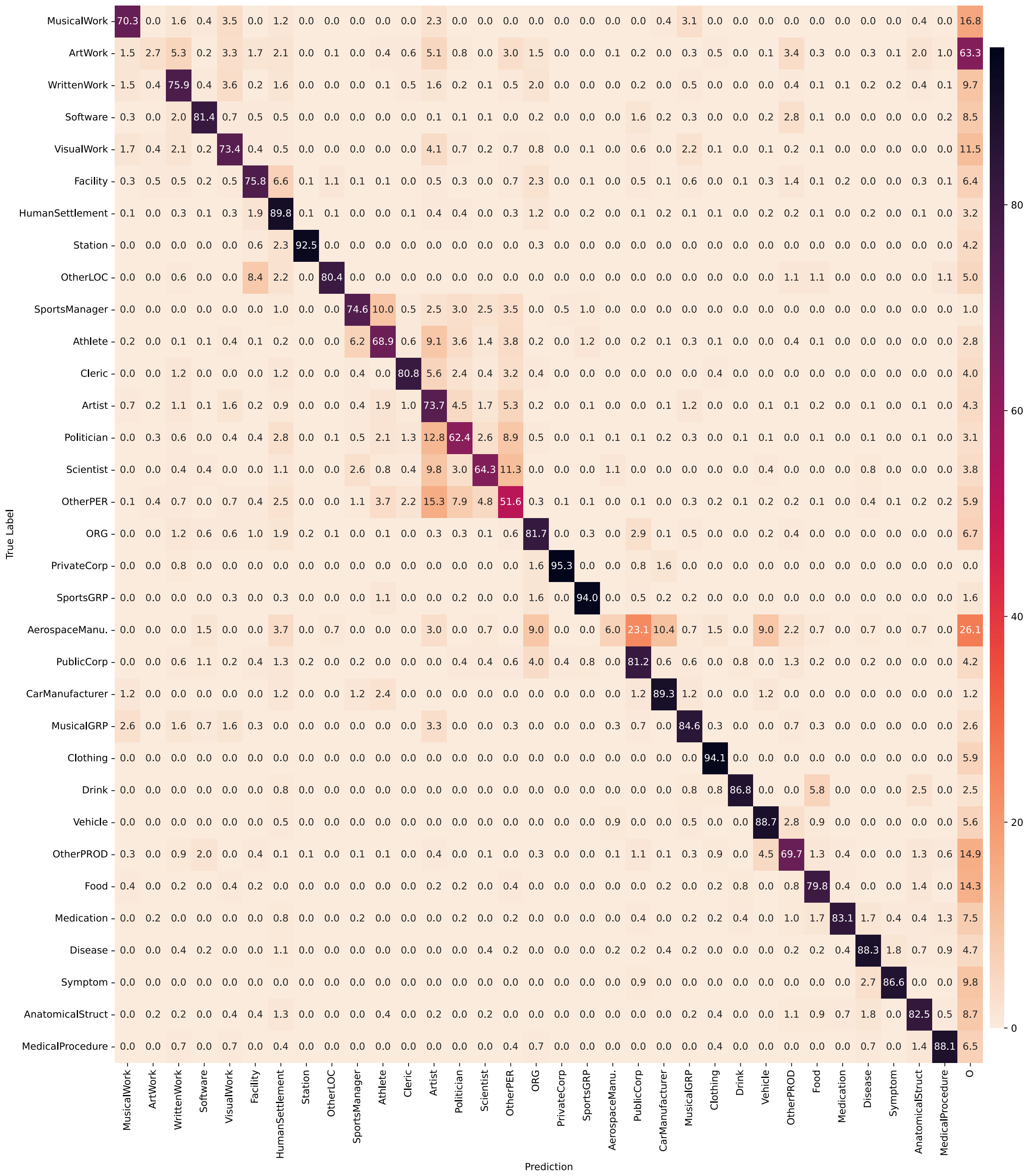


Figure 15: Fine-grained NER predictions for the XLM-RoBERTa baseline on the BN test set.

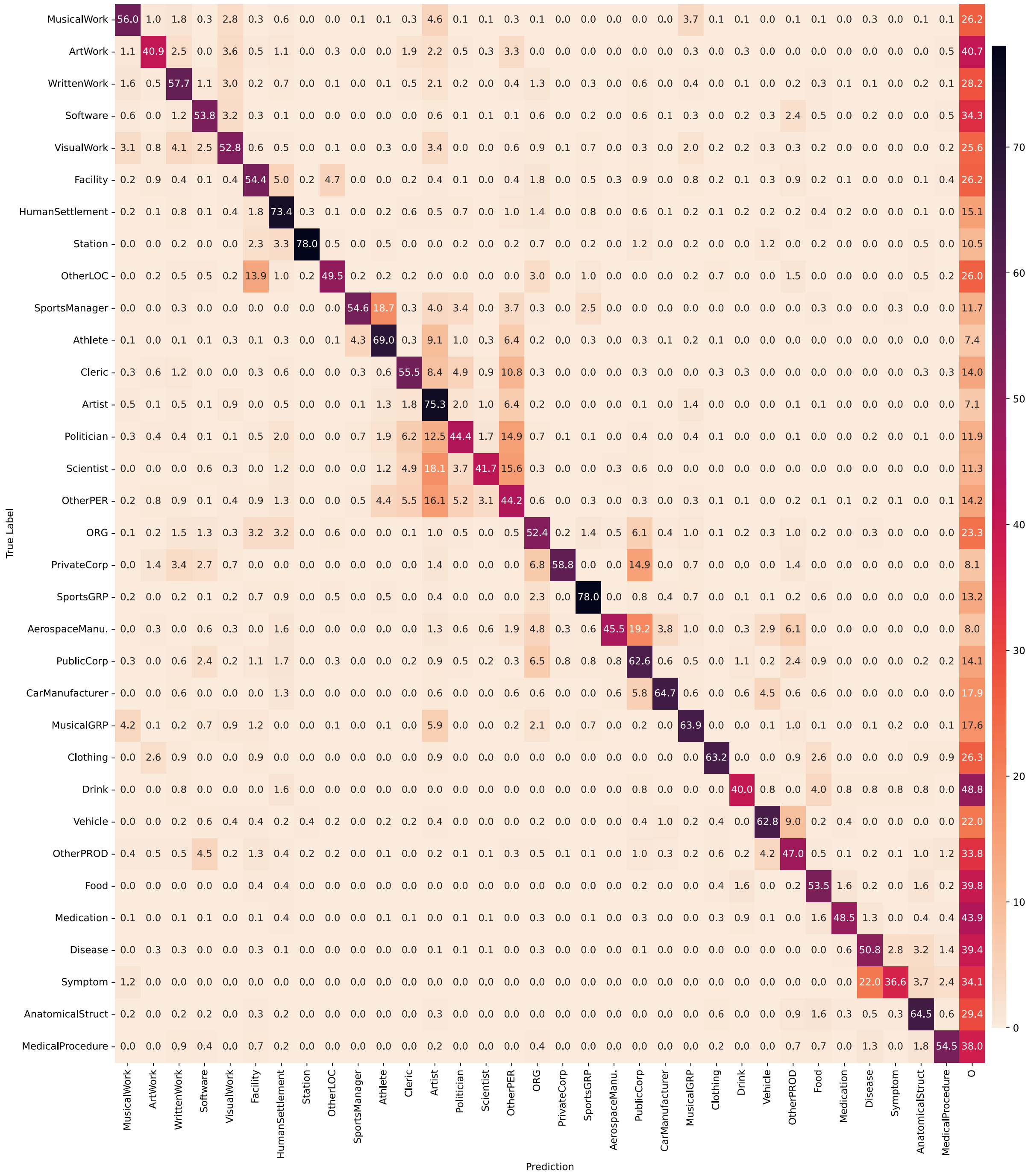


Figure 16: Fine-grained NER predictions for the XLM-RoBERTa baseline on the ZH test set.

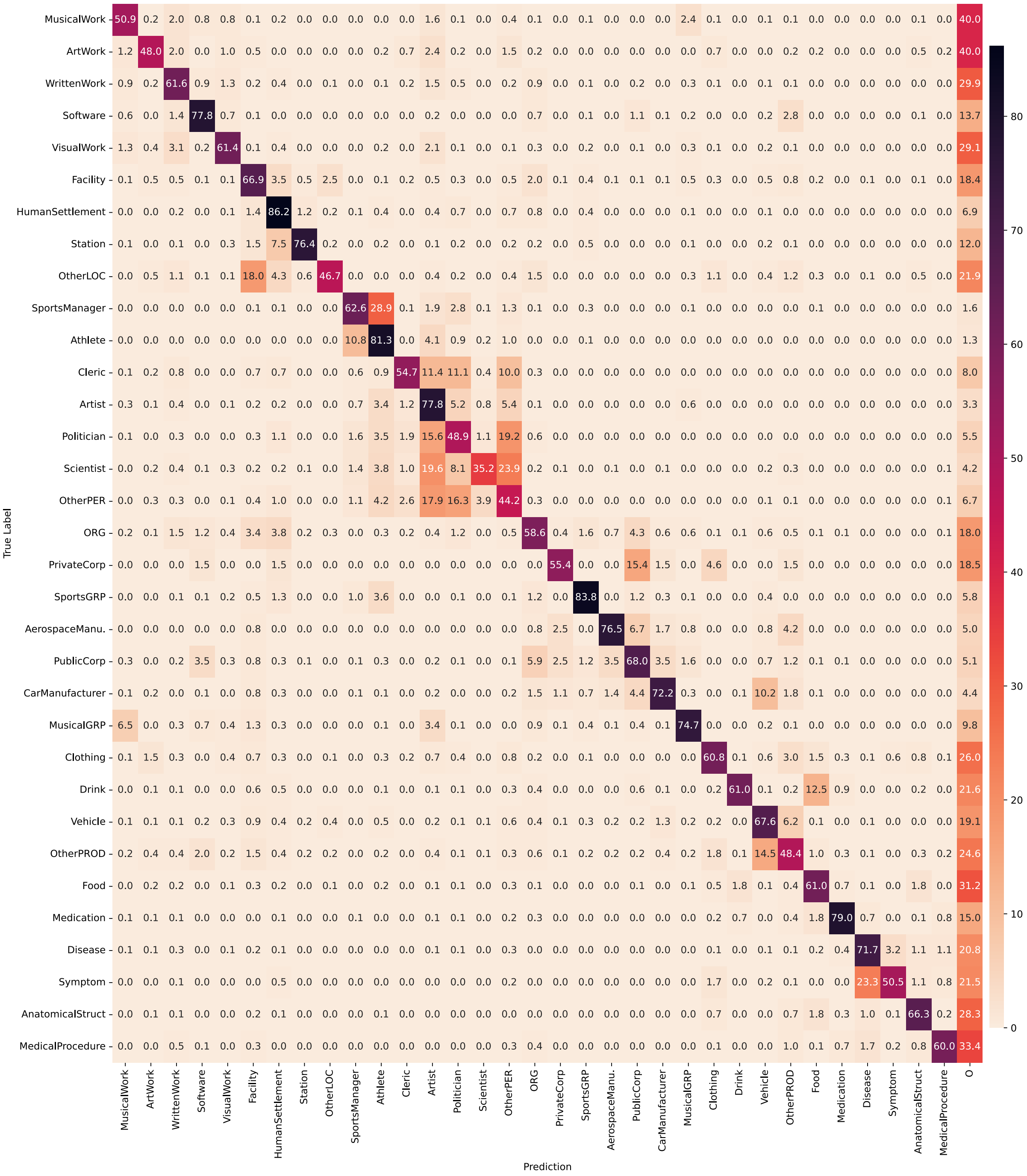


Figure 17: Fine-grained NER predictions for the XLM-RoBERTa baseline on the `UK` test set.

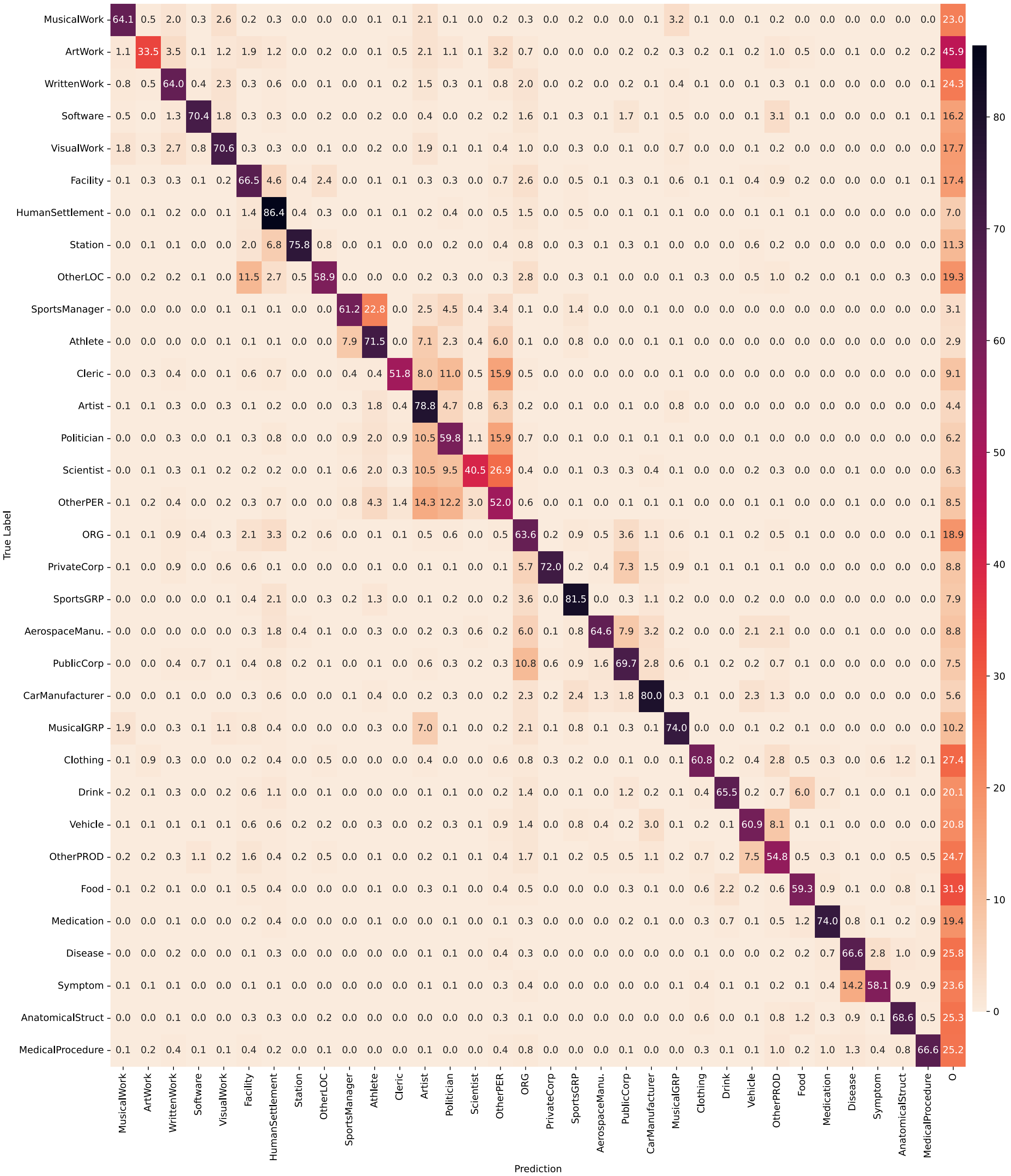


Figure 18: Fine-grained NER predictions for the XLM-RoBERTa baseline on the `MULTI` test set.