OpenNER 1.0: Standardized Open-Access Named Entity Recognition Datasets in 50+ Languages

Chester Palen-Michel and **Maxwell Pickering** and **Maya Kruse**
and **Jonne Sälevä** and **Constantine Lignos**
Michtom School of Computer Science
Brandeis University
{cpalenmichel,pickering,mayakruse,jonnesaleva,lignos}@brandeis.edu

Abstract

We present OpenNER 1.0, a standardized collection of openly-available named entity recognition (NER) datasets. OpenNER contains 36 NER corpora that span 52 languages, human-annotated in varying named entity ontologies. We correct annotation format issues, standardize the original datasets into a uniform representation with consistent entity type names across corpora, and provide the collection in a structure that enables research in multilingual and multi-ontology NER. We provide baseline results using three pretrained multilingual language models and two large language models to compare the performance of recent models and facilitate future research in NER. We find that no single model is best in all languages and that significant work remains to obtain high performance from LLMs on the NER task. OpenNER is released at https://github.com/bltlab/open-ner.

1 Introduction

In the 25+ years following the 7th Message Understanding Conference (MUC-7, Chinchor, 1998), there has been steady development of new datasets for the task of named entity recognition (NER). While the CoNLL 2002–3 shared task datasets (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) and OntoNotes (Hovy et al., 2006) are perhaps the most famous, dozens of corpora have been released since in many languages.

Despite the constant release of new datasets, there is no straightforward way for researchers to work with multiple NER corpora. There is no central repository of NER data, and many of the datasets appearing on lists of NER resources are not readily usable. Many datasets are not consistently formatted and use a variety of chunk encodings (IOB, BIO, etc.), often without documentation.

This paper presents OpenNER 1.0, a first-of-its-kind multilingual, multi-ontology collection of openly-available human-annotated NER datasets to enable painless research into NER beyond the scale of a single corpus. OpenNER is the largest collection of human-annotated NER data created to-date. As new datasets are created and released, we intend for OpenNER to grow in the number of languages covered.

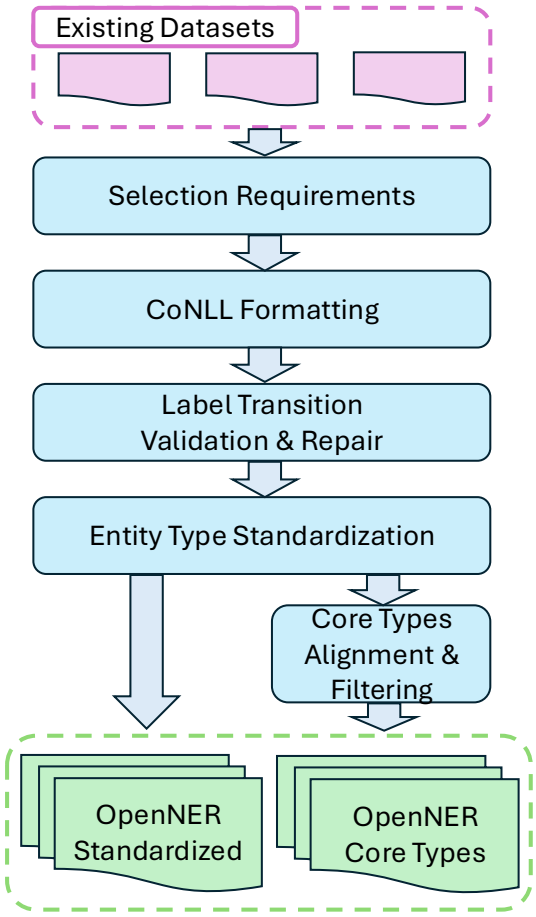

Figure 1: The processing pipeline for OpenNER. Existing datasets (magenta) pass through a series of stages of standardization (blue) to produce two final versions of the dataset (green).

The process of creating OpenNER is shown in

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33638–33663
November 4-9, 2025

Figure 1. We release OpenNER in two versions. The *standardized* version contains all datasets in their original named entity ontologies, with the entity type names mapped to a standard set (e.g. PER is used for "person" in all datasets). The *core types* version contains all datasets but only includes person, location, and organization entity types.

OpenNER is released at https://github.com/bltlab/open-ner. The repository contains all code needed to assemble and preprocess all datasets. The repository README contains links to where any other copies of the data are hosted (Hugging Face, etc.). The OpenNER collection is licensed under the Creative Commons CC BY license; however, all datasets contained in it are licensed under their own licensing terms, some of which prohibit commercial usage.

2 Data Sources

2.1 Selection Requirements

The requirements we set for inclusion of corpora in OpenNER are as follows.

Openly-Accessible First, all datasets must be truly openly-accessible such that they can be easily and legally accessed on the open internet, without requiring the user to request the data or sign an agreement. We do not include datasets that are "available by request" because our goal is to create a benchmark dataset that anyone can automatically run.[1] While all datasets we include are publicly available, some do restrict commercial usage.

Human Annotation Second, the data must have been manually-annotated using explicit annotation guidelines; we do not include any "silver-standard" datasets where all or part of the annotation was automatically generated (e.g. Fetahu et al., 2023; Pan et al., 2017; Zhou et al., 2023).

General Purpose Ontology For the initial release of OpenNER, the annotation must center around traditional named entities, such as persons, locations, organizations, works of art, etc. While we acknowledge their importance, we did not include corpora for chunk extraction in specific domains such as biomedical data or legal cases. Adding these domains presents additional challenges for entity type standardization since they are less likely to have overlap with more generic NER entity types. We leave the incorporation of such datasets to future work as it requires significant additional research.

We do not require any specific entity types to be included in the datasets; we include all types annotated in the original datasets, although we do rename some types to standardize them across corpora, for example renaming all variants of the person type (e.g., PERSON, PERS, PER) to PER. We take a different approach than Universal NER (UNER) (Mayhew et al., 2024) in that our goal is to include as many existing datasets as possible, despite their annotation differences, rather than producing new datasets with uniform annotation.

Sufficient Data We require that there be enough data to create training and test datasets to support experiments. This excludes some small test-only corpora, such as the Europarl annotations (Agerri et al., 2018a), which are significantly smaller than most of the other included datasets. Similarly, UNER (Mayhew et al., 2024) contains a number of test-only datasets that we did not include.

Tokenization and Formatting Finally, the data must be available in a tokenized format; if not already "CoNLL-style," one that can be straightforwardly converted into it. We tried to accept as many corpora as possible, correcting a substantial number of formatting and entity encoding errors. While we are interested in including datasets that do not provide tokenization, doing so would require either performing word segmentation for every corpus and aligning it to the annotation—an error-prone and lossy process—or a new set of tools for preprocessing and training NER models, as most models take pretokenized data as input.

2.2 Datasets Included

We include 36 corpora spanning 52 languages in OpenNER. Most of the datasets use a variant of the CoNLL-02 ontology (Tjong Kim Sang, 2002), and a few are derived from OntoNotes (Hovy et al., 2006) or develop customized ontologies. As seen in Table 5, the datasets span a range of language families and differing numbers of entity types. We categorize the corpora as following either a CoNLL- or OntoNotes-derived ontology in Table 1, which provides names and citations for all included datasets. The CoNLL-02 corpus (Tjong Kim Sang, 2002) consists of Spanish and Dutch newswire data and introduces the LOC/ORG/PER/MISC tagset adapted by many other corpora in this collection. The majority

[1] We have also found that many of datasets that are only available by request have been collected in a way that potentially violates the terms of use or copyright of data sources.

Corpus	Ontology	Source	Corpus	Ontology	Source
AnCora	CoNLL	Taulé et al. (2008)	L3Cube MahaNER	CoNLL	Litake et al. (2022)
AQMAR	CoNLL	Mohit et al. (2012)	MasakhaNER	CoNLL	Adelani et al. (2021)
ArmanPersoNER	CoNLL	Poostchi et al. (2016)	MasakhaNER 2	CoNLL	Adelani et al. (2022)
BarNER	CoNLL	Peng et al. (2024)	NEMO2	CoNLL	Bareket and Tsarfaty (2021)
CONLL-02	CoNLL	Tjong Kim Sang (2002)	NorNE	CoNLL	Jørgensen et al. (2020)
DaNE	CoNLL	Hvingelby et al. (2020)	RONEC	OntoNotes	Dumitrescu and Avram (2020)
EIEC	CoNLL	Alegria et al. (2006)	SLI Galician Corpora	CoNLL	Agerri et al. (2018b)
elNER	OntoNotes	Bartziokas et al. (2020)	ssj500k	CoNLL	Dobrovoljc et al. (2017)
EverestNER	CoNLL	Niraula and Chapagain (2022)	ThaiNNER	OntoNotes	Buaphet et al. (2022)
GermEval	CoNLL	Benikova et al. (2014)	TurkNLP	CoNLL	Luoma et al. (2020)
HiNER	CoNLL	Murthy et al. (2022)	Tweebank	CoNLL	Jiang et al. (2022)
hr500k	CoNLL	Ljubešić et al. (2016)	UNER	CoNLL	Mayhew et al. (2024)
Japanese-GSD	OntoNotes	Asahara et al. (2018)	WikiGoldSK	CoNLL	Suba et al. (2023)
KazNERD	OntoNotes	Yeshpanov et al. (2022)	WNUT17	CoNLL	Derczynski et al. (2017)
KIND	CoNLL	Paccosi and Palmero Aprosio (2022)			

Table 1: Dataset sources and whether the entity type set is more similar to the CoNLL or OntoNotes ontologies.

of corpora in OpenNER follow a type ontology similar to that of CoNLL-02 with `PERSON`, `LOCATION`, `ORGANIZATION`, and `MISC`.

Some CoNLL-inspired corpora leave out `MISC` (e.g. Mayhew et al., 2024), while others replace `MISC` with other types. ArmanPersoNER (Poostchi et al., 2016) adds `EVENT`, `FACILITY`, and `PRODUCT`, while MasakhaNER Adelani et al. (2021) adds `DATE`. Other corpora follow the OntoNotes ontology but collapse types (RONEC, Dumitrescu and Avram, 2020), add types (Japanese-GSD, Asahara et al., 2018), or use a subset (ThaiNNER, Buaphet et al., 2022). More detail about the included datasets, including variations in entity types, are provided in Appendix C.

2.3 Datasets not Included

Unfortunately, some datasets could not be included in our collection for a variety of reasons. Many datasets require the user to request either the annotations or the text. The CoNLL-03 shared task (Tjong Kim Sang and De Meulder, 2003) and OntoNotes (Hovy et al., 2006) datasets use text that cannot be freely distributed; legal use of the data requires that the source text be requested from NIST and the LDC respectively. The data for the EVALITA 2009 Italian NER shared task (Speranza, 2009) was only available by request. The Wojood Arabic NER dataset (Jarrar et al., 2022) only has a sample of data publicly available a at the time this research was performed; the remainder of the dataset is only available upon request. Datasets that require payment, such as the LORELEI language packs for less-resourced languages (Strassel and Tracey, 2016), also could not be included because they are not freely available.

We cannot easily convert datasets to CoNLL format without an authoritative tokenization of the data. This unfortunately excludes some datasets which are otherwise good candidates for inclusion. Datasets which report mentions as character offsets but without tokenization could not be included, such as the MEN corpus of Malaysian English news (Chanthran et al., 2024) and the DANSK corpus of multi-domain Danish (Enevoldsen et al., 2024). Similarly, the multilingual SlavicNER corpus reports a list of mentions with character offsets for each source document, but without tokenization (Piskorski et al., 2024). The ENP-NER corpus of historical Chinese newspapers reports character-level tags (Blouin et al., 2024).

We did not include corpora for specialized domains such as biomedical data (Byun et al., 2024), paper abstracts (Phi et al., 2024; Alkan et al., 2024), and industrial documents (Li et al., 2024).

We only include datasets created using human annotation. Although WikiAnn (Pan et al., 2017) is often used as a multilingual NER benchmark, it is a "silver-standard" dataset and uses automatically-created labels. We did not include MultiCoNER (Malmasi et al., 2022) as it has not been hand-annotated, but rather extracted from text that is linked to articles corresponding to entity types. We do not include NerKor+Cars-OntoNotes++ (Novák and Novák, 2022) because it uses a semi-automatic labeling approach where not all labels are manually checked. As there are many popular and widely-known NER datasets that are not human-annotated, our goal with NER was to provide a complementary dataset that only contains human annotation.

Some candidate datasets were not included because of pervasive dataset quality issues or annotation errors that were too onerous to repair. Details for these datasets are included in Appendix B. Finally, we could not include the datasets for many

older papers because they were no longer available.

3 Standardization

3.1 CoNLL Formatting

We require all included datasets to be converted to the CoNLL format with BIO mention encoding and UTF-8 text encoding. The CoNLL format represents labeled sequences with one token per line, with sentences separated by newlines. The type label and any other metadata pertaining to the token appear on the same line as the token, separated by whitespace. We had to modify the text encoding and file formats used by several corpora; details are provided in Appendix Section A.1.

3.2 Label Transition Validation

We corrected label transition errors—failures to correctly follow the BIO, IOB, etc. encoding schemes—automatically when possible, and manually when required. Repairing invalid label sequences involved validating with SeqScore (Palen-Michel et al., 2021) and manually reviewing the validation errors. If the errors all appeared to be safely repaired with SeqScore's repair functionality, automated repairs were performed. In some cases, manual repairs were conducted, and those repairs are detailed in Appendix Section A.2.

3.3 Entity Type Standardization

Once all datasets were correctly BIO-encoded, we standardized the entity types in order to have a consistent set of entity type labels across all datasets.

We adopted the following conventions for named entity types. Each type is written as capitalized letters, with underscores used to separate words in multi-word names (e.g. `PET_NAME`). Any subtags are written with a hyphen, following the "dash tag" style (e.g. `LOC-DERIV`). When there is a commonly-used short form for a type (e.g. `ORG`), we map longer versions of the name to it. For example, there are six different ways that the "organization" entity type is named across different corpora: `ORG`, `Organization`, `ORGANIZATION`, `ORGANISATION`, `org`, `NEO`. We standardize them all to `ORG`. Similar but non-identical entity types, such as `DATETIME` and `TIME`, are left separate.

This standardization process preserves all annotation in the original datasets. No name mentions are removed, and within each corpus, no types are combined. This process creates the most uniform set of types possible across all of the datasets and allows for better usability; however, users should be aware that annotation guidelines still vary across corpora, leading to the span or entity type for similar mentions across two different corpora to differ. OpenNER provides the first easy way to explore these differences in annotation at scale. Appendix Table 11 gives the full mapping of types used in the standardization process. There are 60 unique entity types across 2,816,304 total mentions.

Before Mapping	After Mapping
PER, PER-PART	PER
LOC, GPE, GPE-LOC, LOC-PART, GPE-ORG, FACILITY	LOC
ORG, ORG-PART, CORPORATION, GROUP	ORG

Table 2: Entity types mapped to core types of PER, ORG, and LOC for the core types version of OpenNER.

3.4 Core Types

We additionally provide a secondary version of OpenNER where we map entity types to a set of minimal core types—location (`LOC`), organization (`ORG`), and person (`PER`)—and discard all other types. This minimal ontology is useful for exploring commonalities across datasets and training multi-corpus and multilingual models. Table 2 gives the mapping of types used to create the core entity types version of the data, which was manually created after reviewing the annotation guidelines of each dataset.

While we have reduced all datasets to three common types, that does not mean that in every corpus those entity types are annotated in the same way. For example, some corpora may "tag by usage," such that if a sentence is about the physical location of a corporation it might be tagged as `LOC` instead of `ORG`, while others still use `ORG` in that instance. Corpora may also differ in the extent of the annotated span. OpenNER does not modify the original annotation beyond the mapping of types, so these differences across corpora persist in the core types version of the data.

3.5 Dataset Statistics

OpenNER includes annotations in 52 languages from a diverse set of language families that use 11 different scripts. Table 5 gives the number of entity types and the number of training, validation, and test sentences for each language in each corpus.

Entity Type	Count
LOC	464,426
ORG	247,745
PER	330,094
Total	1,042,265

Table 3: Counts of names of each entity type in the core types version of OpenNER.

Appendix Table 9 gives statistics and general information about the languages included in OpenNER.

Appendix Table 10 gives the counts for every standardized entity type in our resource. The most frequent type is LOC, with 412k mentions out of a total of 2.8M. The rarest types are mostly dash tags (e.g. EVENT-DERIV) and a few very rarely-used types (e.g. NON_HUMAN). We include all entity types present in the original dataset, regardless of their frequency, but users of OpenNER may choose to exclude rare types in future evaluations. Table 3 gives the entity types counts for the 1M mentions in the core types version of the data, where LOC remains the most frequent type.

4 Experiments

In addition to providing a resource, we provide baselines using popular methods for performing NER in many languages. We fine-tuned the XLM-RoBERTa Base (XLM-R, Conneau et al., 2020), mBERT (Devlin et al., 2019), and Glot500 (Imani et al., 2023) models. We selected these models due to their popular usage in NER for less-resourced languages.[2] Many "state of the art" NER papers only evaluate on the CoNLL-03 English data or a few other high-resource language datasets; they also often rely on monolingual or few-language models, making adapting them to the 52 languages of OpenNER non-trivial. We also benchmark two LLMs, Aya-Expanse (Dang et al., 2024) and QwQ-32B-Preview (Qwen Team, 2024) using only the core types to explore LLM performance.

We experimented with three approaches to developing models. The first involves training one model for each language in each dataset (67 language-dataset combinations in total), using the full set of entity types in each dataset. The second involves training one model for each language in each dataset, but using only the core types (LOC, ORG, and PER). The third involves training one multilingual model (or using in-context learning with an LLM) across all datasets and languages using only the core types. For the multilingual model, we cap the training data at 32k sentences per language per dataset to mitigate data imbalances.

These experiments allow us to explore performance on both the original and core types ontologies, demonstrate the feasibility of multilingual NER models, and evaluate the performance of LLMs in a relatively simple NER ontology.

We report micro-averaged mention-level F1 for each model computed with SeqScore using the same method as the conlleval script. We report the mean and standard error of the mean over 10 different training runs, each using a different random seed for initialization. In total, we performed 4,050 fine-tuning runs; this large number is due to the sheer number of languages and datasets in OpenNER and the number of random seeds. Hyperparameters and model details are discussed in Appendix Section A.3.

4.1 Individual Models on Full Ontologies

Results for training individual models using the full ontology for each dataset are shown in Figure 2, with all results in Appendix Table 6. Averaged across all language-dataset combinations, the best performing model was XLM-R (F1 of 81.79), followed by Glot500 (80.96), and then mBERT (74.42). While Glot500 performs best on many of the least-resourced languages, XLM-R performs better on average. While mBERT generally performs worse than the other models, it scores substantially better in both Chinese datasets and does exceptionally well on the Maghrebi Arabic dataset, which is written in NArabizi, a method of writing Arabic in Latin script used in North Africa. However, mBERT does not function at all in Amharic due to not being pretrained on the Ge'ez script. We observe that there may be evidence of catastrophic forgetting in Glot500, as some higher-resourced languages like Spanish, Swedish, and English Tweebank underperform using Glot500 compared with XLM-R, which it is based on.

4.2 Multilingual and Individual Models on Core Types

Results for training individual and multilingual models using only the core types are shown in Figures 3 and 4, summarized in Table 4 and Figure 5, with all results in Appendix Table 7. Glot500

[2]Due to a request from a reviewer, we performed a post-hoc evaluation of XLM-R Large. We found it performs worse than XLM-R Base despite being a larger model, although it is possible performance could be improved with further tuning.

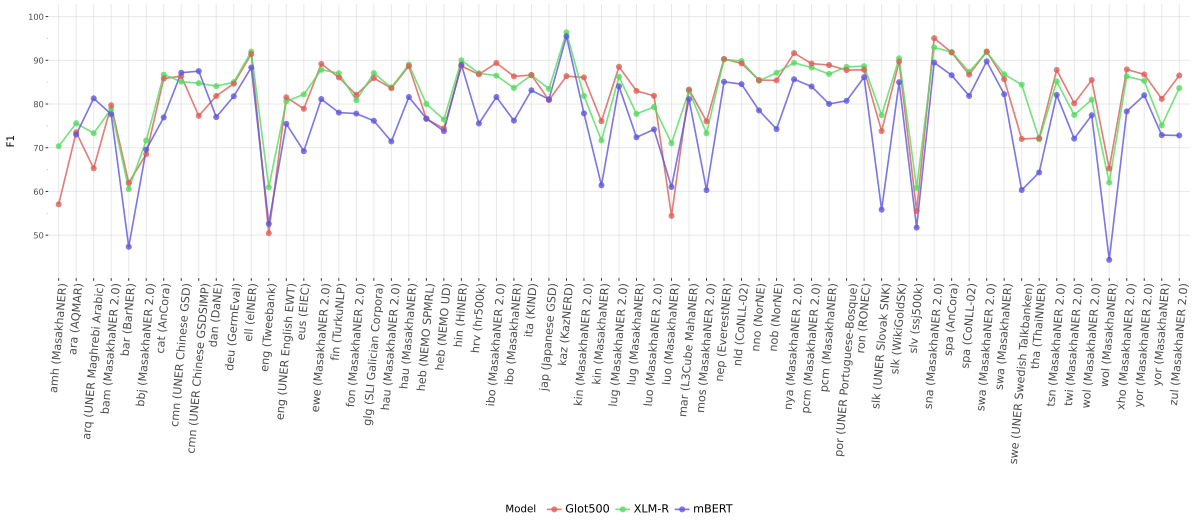

Figure 2: Mean F1 for each dataset-language combination, using all entity types present in each dataset. Models were fine-tuned individually on each dataset-language combination.

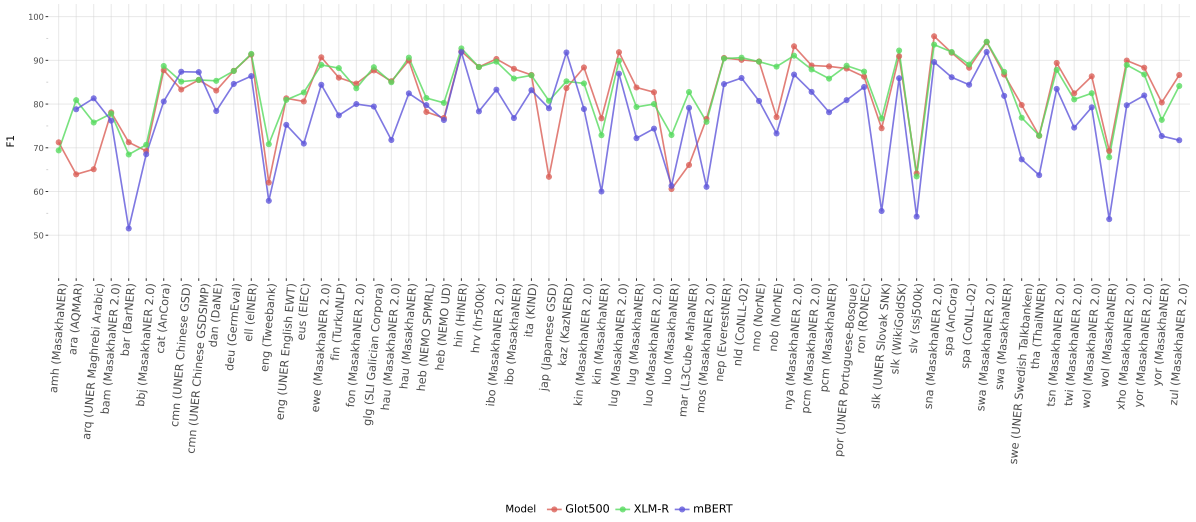

Figure 3: Mean F1 for each dataset-language combination, using only core entity types (location, organization, and person). Models were fine-tuned individually on each dataset-language combination.

again excels on the least-resourced languages and now also achieves the highest average performance. The multilingual models often deliver better performance for less-resourced languages and cases where the exact same ontology is shared across datasets (e.g. MasakhaNER), while for many higher-resourced languages the best performance comes with individual-language models. It is possible differences in annotation guidelines limit the multilingual model's performance in languages which do not have the same guidelines as others, while aiding transfer learning for datasets where guidelines are similar to others.

To assess the statistical significance of our main results, we compared each pair of the fine-tuned models using the non-parametric Wilcoxon signed-rank test with the core types models. Paired comparisons were made for each of the three models using the 67 dataset-language combinations, comparing the mean F1 across random seeds for each model. Each comparison is tested at the conventional 0.05 alpha level. Since we are performing several hypothesis tests, we account for multiple comparisons using a Bonferroni correction and compare all p-values to the corrected confidence threshold of $0.05/3$, equal to the alpha level of the test (0.05) divided by the number of tests (3). This keeps the familywise false positive rate at 0.05.

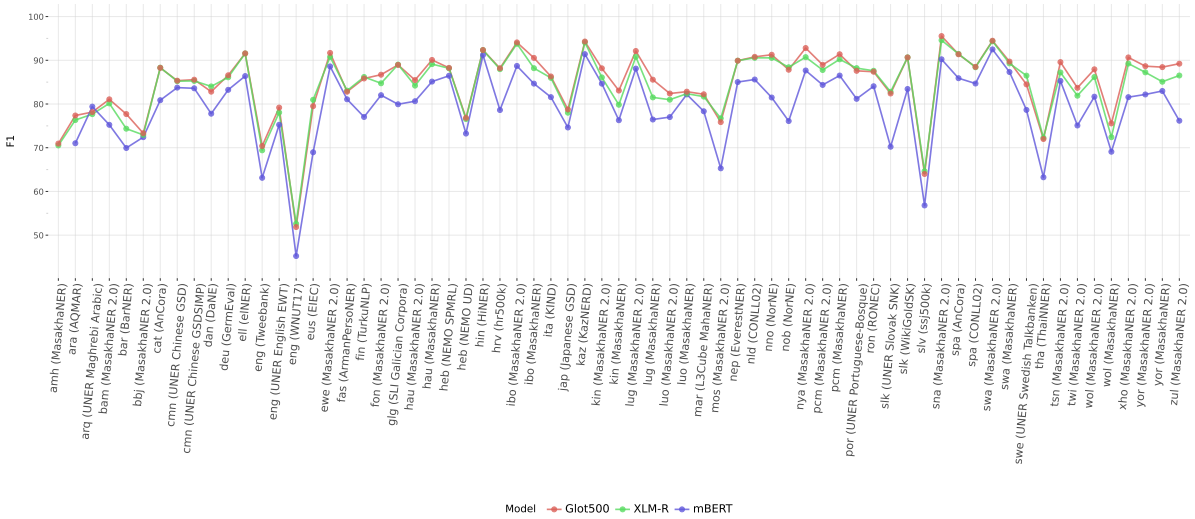

Figure 4: Mean F1 for each dataset-language combination, using only core entity types (location, organization, and person). Multilingual models were fine-tuned using all datasets and languages.

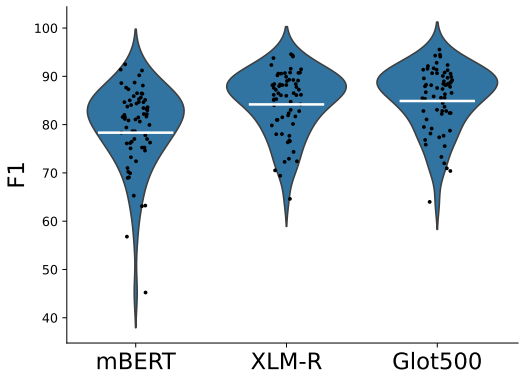

Figure 5: Violin plot of F1 distributions per model, with points depicting mean F1 scores across random seeds for each language-dataset combination. White lines indicate means of all points per-model.

Train Approach	Model	Mean F1
Individual	mBERT	75.69
Individual	XLM-R	83.10
Individual	Glot500	82.10
Multilingual	mBERT	78.33
Multilingual	XLM-R	84.18
Multilingual	Glot500	**84.89**
In-context	Aya-Expanse 32b	60.55
In-context	QwQ 32b Preview	60.66

Table 4: Mean F1 across each language-dataset combination for each approach, using only core types.

When looking at the multilingual models (corresponding to the middle rows of Table 4), on average, Glot500 outperforms XLM-R by 0.71 F1 ($p = 1.64 \cdot 10^{-5} < 0.05/3$), Glot500 outperforms mBERT by 6.56 F1 ($p = 1.34 \cdot 10^{-12} < 0.05/3$), and XLM-R outperforms mBERT by 5.85 F1 ($p = 1.61 \cdot 10^{-12} < 0.05/3$). These statistical tests confirm that the differences are statistically reliable across the language-dataset combinations.

When finetuning is conducted on individual languages rather than multilingually, the pattern changes slightly. On average, XLM-R now outperforms Glot500 by 0.44 F1 ($p = 0.82 > 0.05/3$) though the result is not statistically significant. Glot500 outperforms mBERT by 6.97 F1 ($p = 1.96 \cdot 10^{-7} < 0.05/3$), which is similar to the previous difference of 6.56 F1. XLM-R outperforms mBERT by 7.41 F1 ($p = 4.21 \cdot 10^{-11} < 0.05/3$), which is larger than the 5.85 F1 difference observed earlier.

4.3 LLMs

We conducted baseline experiments with two LLMs, Aya-Expanse 32b and QwQ-32B-Preview using 5-shot demonstrations. We chose these models because QwQ supports 29 languages and Aya-Expanse supports 23, and the models had similar numbers of parameters. The vLLM engine (Kwon et al., 2023) was used for inference. We evaluated using only the core types so that a standard prompt could be used across all datasets.

To score LLM output using traditional NER evaluation methods, one must map back to the original text, which poses many challenges. Hallucinated tokens detected as names must either be discarded

or penalized as false positives. There is also no guarantee that the generated labels will even be part of the ontology. We conducted preliminary experiments with a handful of approaches for using LLMs to conduct NER, discussed further in Appendix A.4. Our preliminary experiments generally align with the findings of Villena et al. (2024): few-shot demonstrations perform better than zero-shot, and inline entity labeling generally outperforms JSON output.

For our LLM experiments, we used inline responses with prompts including 5 example demonstrations. The prompt template, inspired by prompting techniques from Wang et al. (2025), is as follows: `Find names of persons, organizations or locations. Label the following sentence with labels where the name is enclosed with the entity type PER, ORG or LOC and @@ ##. For example PER @@ John Smith ##, or ORG @@ Springfield University ## or LOC @@ United Kingdom ## . Find named entities in the following sentence: «SENTTEXT».`

Full examples of the prompt with demonstrations are included in Appendix A.4. We conducted runs with 3 random seeds which were used for selecting the 5-shot demonstrations. The demonstrations were the same for all sentences for each dataset-language combination.

The results, given in Table 4 and Appendix Table 8, show that LLM performance is substantially worse than the other methods we evaluate, consistent with other work showing performance lags behind encoder-only models with a classification head (Villena et al., 2024; Wang et al., 2025; Xie et al., 2023). Curiously, the two models often scored identically—even though their predictions were not identical—suggesting that the in-context examples may be the limiting factor.

While we only explore comparatively simple prompting experiments in this work to establish LLM baselines, further methods with LLMs such as synthetic data generation (Santoso et al., 2024) and few-shot demonstration retrieval (Wang et al., 2025) are promising lines of research which OpenNER could facilitate. Whether and how to best leverage LLMs for NER, and in particular multilingual NER, remains an open problem, which we leave for future work.

4.4 Discussion

Overall, the results show that mBERT tends to perform worse than XLM-R or Glot500, but there are still cases where mBERT outperforms other models despite its age. XLM-R and Glot500 perform similarly, with the former performing better when trained on individual languages, and the latter performing better when trained multilingually. There does not currently appear to be a single best, one-size-fits-all model for these datasets. While LLMs may eventually outperform sequence-labeling methods, further research is required to improve their performance.

5 Future Work and Conclusion

We believe OpenNER will facilitate future research in multilingual NER by drastically reducing the barrier to entry for researchers working with multiple NER datasets. We have shown the potential for future experimentation with transfer learning and that there are challenges for training NER models that can handle multiple languages.

While OpenNER does not cover as many languages as "silver standard" (automatically annotated) datasets, it provides high-quality data in a smaller set of languages, many of them less-resourced. We welcome the inclusion of additional datasets that we may have missed along with new datasets when they are created and released publicly. We plan to release regular updates to include additional datasets as they are released and update our benchmarks with new methods.

Corpus	Language	Code	Types	Sentences			
				Train	Dev	Test	Total
AnCora	Catalan	cat	4	10,629	1,428	1,527	13,584
AnCora	Spanish	spa	6	11,374	2,992	2,983	17,349
ArmanPersoNER	Persian	fas	6	5,121	0	2,560	7,681
AQMAR	Arabic	ara	4	1,328	710	605	2,643
BarNER	Bavarian German	bar	23	2,869	338	370	3,577
CoNLL-02	Dutch	nld	4	15,806	2,895	5,195	23,896
CoNLL-02	Spanish	spa	4	8,323	1,915	1,517	11,755
DaNE	Danish	dan	4	4,383	564	565	5,512
EIEC	Basque	eus	4	2,552	0	842	3,394
elNER	Greek	ell	18	17,132	1,904	2,116	21,152
EverestNER	Nepali	nep	5	13,848	0	1,950	15,798
GermEval	German	deu	12	24,000	2,200	5,100	31,300
HiNER	Hindi	hin	11	75,827	10,851	21,657	108,335
hr500k	Croatian	hrv	5	17,869	2,499	4,341	24,709
Japanese-GSD	Japanese	jpn	22	7,050	507	543	8,100
KazNERD	Kazakh	kaz	25	90,228	11,167	11,307	112,702
KIND	Italian	ita	3	37,765	0	7,385	45,150
L3Cube MahaNER	Marathi	mar	7	21,493	1,499	1,998	24,990
MasakhaNER	Amharic	amh	4	1,750	250	500	2,500
MasakhaNER	Hausa	hau	4	1,903	272	545	2,720
MasakhaNER	Igbo	ibo	4	2,233	319	638	3,190
MasakhaNER	Kinyarwanda	kin	4	2,110	301	604	3,015
MasakhaNER	Luganda	lug	4	1,402	200	401	2,003
MasakhaNER	Luo	luo	4	644	92	185	921
MasakhaNER	Naija	pcm	4	2,100	300	600	3,000
MasakhaNER	Kiswahili	swa	4	2,104	300	602	3,006
MasakhaNER	Wolof	wol	4	1,871	267	536	2,674
MasakhaNER	Yoruba	yor	4	2,124	303	608	3,035
MasakhaNER 2.0	Bambara	bam	4	4,462	638	1,274	6,374
MasakhaNER 2.0	Ghomálá'	bbj	4	3,384	483	966	4,833
MasakhaNER 2.0	Éwé	ewe	4	3,505	501	1,001	5,007
MasakhaNER 2.0	Fon	fon	4	4,343	623	1,228	6,194
MasakhaNER 2.0	Hausa	hau	4	5,716	816	1,633	8,165
MasakhaNER 2.0	Igbo	ibo	4	7,634	1,090	2,181	10,905
MasakhaNER 2.0	Kinyarwanda	kin	4	7,825	1,118	2,235	11,178
MasakhaNER 2.0	Luganda	lug	4	4,942	706	1,412	7,060
MasakhaNER 2.0	Luo	luo	4	5,161	737	1,474	7,372
MasakhaNER 2.0	Mossi	mos	4	4,532	648	1,294	6,474
MasakhaNER 2.0	Chichewa	nya	4	6,250	893	1,785	8,928
MasakhaNER 2.0	Naija	pcm	4	5,646	806	1,613	8,065
MasakhaNER 2.0	chiShona	sna	4	6,207	887	1,773	8,867
MasakhaNER 2.0	Kiswahili	swa	4	6,593	942	1,883	9,418
MasakhaNER 2.0	Setswana	tsn	4	3,489	499	996	4,984
MasakhaNER 2.0	Akan/Twi	twi	4	4,240	605	1,211	6,056
MasakhaNER 2.0	Wolof	wol	4	4,593	656	1,312	6,561
MasakhaNER 2.0	isiXhosa	xho	4	5,718	817	1,633	8,168
MasakhaNER 2.0	Yorùbá	yor	4	6,876	983	1,964	9,823
MasakhaNER 2.0	Zulu	zul	4	5,848	836	1,670	8,354
NEMO SPMRL	Hebrew	heb	9	4,937	500	706	6,143
NEMO UD	Hebrew	heb	9	5,168	484	491	6,143
NorNE	Norwegian (Nynorsk)	nno	9	14,174	1,890	1,511	17,575
NorNE	Norwegian (Bokmål)	nob	9	15,696	2,410	1,939	20,045
RONEC	Romanian	ron	15	9,000	1,330	2,000	12,330
SLI Galician Corpora	Galician	glg	4	6,483	0	1,655	8,138
ssj500k	Slovenian	slv	5	9,077	1,147	1,134	11,358
ThaiNNER	Thai	tha	10	3,914	0	980	4,894
TurkuNLP	Finnish	fin	6	12,217	1,364	1,555	15,136
Tweebank	English	eng	4	1,639	710	1,201	3,550
UNER Chinese GSD	Mandarin Chinese	cmn	3	3,997	500	500	4,997
UNER Chinese GSDSIMP	Mandarin Chinese	cmn	3	3,997	500	500	4,997
UNER English EWT	English	eng	3	12,543	2,001	2,077	16,621
UNER Maghrebi French-Arabic	Maghrebi Arabic	arq	3	1,003	139	145	1,287
UNER Portuguese-Bosque	Portuguese	por	3	7,018	1,172	1,167	9,357
UNER Slovak SNK	Slovak	slk	3	8,483	1,060	1,061	10,604
UNER Swedish Talkbanken	Swedish	swe	3	4,303	504	1,219	6,026
WikiGoldSK	Slovak	slk	4	4,687	669	1,340	6,696
WNUT17	English	eng	6	3,394	1,009	1,287	5,690
Total				624,532	76,746	130,786	832,064

Table 5: Statistics for corpora included in OpenNER. Language codes are given using the ISO 639-3 standard.

Limitations

Despite tremendous efforts to include every eligible dataset, there may be datasets that met our criteria that we missed due to the difficulty of trying to find every single hand-annotated NER dataset in existence. We hope that our commitment to regular releases will mitigate this limitation.

OpenNER is a collection of existing corpora, and thus it faithfully represents the biases in both the construction of such corpora (i.e. which languages they are created in) and their contents. Due to the recent release of the MasakhaNER datasets, African languages are overrepresented in OpenNER compared to those of the rest of the world. OpenNER has substantial coverage of European and African languages but has little coverage of languages spoken in Asia and South America beyond the majority languages spoken in the larger countries. Unfortunately, part of the cause of underrepresentation of Asian languages is due to many of the NER datasets existing in those languages not being in usable condition (see Appendix B). Outside of Africa, there is little coverage of indigenous or minority languages due to the limited corpora in existence.

The majority of the languages included in OpenNER are written in Latin script, but broader script coverage is likely to be a key element of building multilingual NER models.

When mapping to core types, users should be aware that although datasets might be using the same entity type names, these were not annotated using common guidelines and are hence expected to be noisy. While this is a limitation, it also presents an opportunity for future work in exploring better mappings from the standardized version of OpenNER to alternative unified types or other approaches to train models to learn from datasets in disparate ontologies.

Ethical Considerations

We believe OpenNER will have a positive impact on future multilingual NER research. Compared with "silver-standard" datasets, OpenNER is high-quality human-annotated data and will bring attention to multilingual gold standard NER datasets that had previously received less attention than some of the larger "silver-standard" datasets.

We have undertaken significant efforts to confirm that all datasets we include allow redistribution and were not derived from more restrictive sources that would not allow it. However, there is always risk that the authors of datasets have misrepresented the restrictions on the data that was annotated, causing us to accidentally redistribute data against the original data owner's wishes.

Acknowledgments

This work was primarily supported by the grant *Improving Relevance and Recovery by Extracting Latent Query Structure* by eBay to Brandeis University. This work was also supported by Brandeis University through internal research funds.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Q. Hacheme, Idris Abdulmumin, Odunayo Ogundepo,

Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka, and German Rigau. 2018a. Building named entity recognition taggers via parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Rodrigo Agerri, Xavier Gómez Guinovart, German Rigau, and Miguel Anxo Solla Portela. 2018b. Developing new linguistic resources and tools for the Galician language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Inaki Alegria, Olatz Arregi, Nerea Ezeiza, and Izaskun Fernández. 2006. Lessons from the development of a named entity recognizer for Basque.

Atilla Kaan Alkan, Felix Grezes, Cyril Grouin, Fabian Schussler, and Pierre Zweigenbaum. 2024. Enriching a time-domain astrophysics corpus with named entity, coreference and astrophysical relationship annotations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6177–6188, Torino, Italia. ELRA and ICCL.

Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura, and Yugo Murawaki. 2018. Universal Dependencies version 2 for Japanese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sankalp Bahad, Pruthwik Mishra, Parameswari Krishnamurthy, and Dipti Sharma. 2024. Fine-tuning pretrained named entity recognition models for Indian languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 75–82, Mexico City, Mexico. Association for Computational Linguistics.

Dan Bareket and Reut Tsarfaty. 2021. Neural modeling for named entities and morphology (NEMO2). *Transactions of the Association for Computational Linguistics*, 9:909–928.

Nikos Bartziokas, Thanassis Mavropoulos, and Constantine Kotropoulos. 2020. Datasets and performance metrics for greek named entity recognition. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 160–167, New York, NY, USA. Association for Computing Machinery.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Pado. 2014. GermEval 2014 named entity recognition shared task: companion paper.

Baptiste Blouin, Cécile Armand, and Christian Henriot. 2024. A dataset for named entity recognition and entity linking in Chinese historical newspapers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 385–394, Torino, Italia. ELRA and ICCL.

Weerayut Buaphet, Can Udomcharoenchaikit, Peerat Limkonchotiwat, Attapol Rutherford, and Sarana Nutanong. 2022. Thai nested named entity recognition corpus. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1473–1486, Dublin, Ireland. Association for Computational Linguistics.

Sungjoo Byun, Jiseung Hong, Sumin Park, Dongjun Jang, Jean Seo, Minseok Kim, Chaeyoung Oh, and Hyopil Shin. 2024. Korean bio-medical corpus (KBMC) for medical named entity recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9941–9947, Torino, Italia. ELRA and ICCL.

MohanRaj Chanthran, Lay-Ki Soon, Huey Fang Ong, and Bhawani Selvaretnam. 2024. Malaysian English news decoded: A linguistic resource for named entity and relation extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10999–11022, Torino, Italia. ELRA and ICCL.

Nancy A. Chinchor. 1998. Overview of MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi,

Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaja Dobrovoljc, Tomaž Erjavec, and Simon Krek. 2017. The Universal Dependencies treebank for Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 33–38, Valencia, Spain. Association for Computational Linguistics.

Stefan Daniel Dumitrescu and Andrei-Marius Avram. 2020. Introducing RONEC - the Romanian named entity corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4436–4443, Marseille, France. European Language Resources Association.

Kenneth Enevoldsen, Emil Jessen, and Rebekah Baglini. 2024. Dansk: Domain generalization of danish named entity recognition. *Northern European Journal of Language Technology*, 10.

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)*, pages 2247–2265, Toronto, Canada. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. DaNE: A named entity resource for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4597–4604, Marseille, France. European Language Resources Association.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. Wojood: Nested Arabic named entity corpus and recognition using BERT. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3626–3636, Marseille, France. European Language Resources Association.

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the Tweebank corpus on named entity recognition and building NLP models for social media analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7199–7208, Marseille, France. European Language Resources Association.

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. 2020. NorNE: Annotating named entities for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4547–4556, Marseille, France. European Language Resources Association.

Redwanul Karim, MA Muhiminul Islam, Sazid Rahman Simanto, Saif Ahmed Chowdhury, Kalyan Roy, Adnan Al Neon, Md Sajid Hasan, Adnan Firoze, and Rashedur M Rahman. 2019. A step towards information extraction: Named entity recognition in bangla using deep learning. *Journal of Intelligent & Fuzzy Systems*, 37(6):7401–7413.

Britt Keson. 1998. Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Ruiting Li, Peiyan Wang, Libang Wang, Danqingxin Yang, and Dongfeng Cai. 2024. A corpus and method for Chinese named entity recognition in manufacturing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 264–272, Torino, Italia. ELRA and ICCL.

Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade, and Raviraj Joshi. 2022. L3Cube-MahaNER: A Marathi named entity recognition dataset and BERT models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 29–34, Marseille, France. European Language Resources Association.

Liyuan Liu, Jingbo Shang, and Jiawei Han. 2019. Arabic named entity recognition: What works and what's next. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 60–67, Florence, Italy. Association for Computational Linguistics.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4264–4270, Portorož, Slovenia. European Language Resources Association (ELRA).

Jouni Luoma, Miika Oinonen, Maria Pyykönen, Veronika Laippala, and Sampo Pyysalo. 2020. A broad-coverage corpus for Finnish named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4615–4624, Marseille, France. European Language Resources Association.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. 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, Terra Blevins, Shuheng Liu, Marek Suppa, Hila Gonen, Joseph Marvin Imperial, Börje Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A. Smith. 2012. Recall-oriented learning of named entities in Arabic Wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173, Avignon, France. Association for Computational Linguistics.

Rudra Murthy, Pallab Bhattacharjee, Rahul Sharnagat, Jyotsana Khatri, Diptesh Kanojia, and Pushpak Bhattacharyya. 2022. HiNER: A large Hindi named entity recognition dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4467–4476, Marseille, France. European Language Resources Association.

Nobal Niraula and Jeevan Chapagain. 2022. Named entity recognition for nepali: Data sets and algorithms.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Attila Novák and Borbála Novák. 2022. NerKor+Cars-OntoNotes++. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1907–1916, Marseille, France. European Language Resources Association.

Teresa Paccosi and Alessio Palmero Aprosio. 2022. KIND: an Italian multi-domain dataset for named entity recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 501–507, Marseille, France. European Language Resources Association.

Chester Palen-Michel, Nolan Holley, and Constantine Lignos. 2021. SeqScore: Addressing barriers to reproducible named entity recognition evaluation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 40–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Siyao Peng, Zihang Sun, Huangyan Shan, Marie Kolm, Verena Blaschke, Ekaterina Artemova, and Barbara Plank. 2024. Sebastian, Basti, Wastl?! recognizing named entities in Bavarian dialectal data. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14478–14493, Torino, Italia. ELRA and ICCL.

Van-Thuy Phi, Hiroki Teranishi, Yuji Matsumoto, Hiroyuki Oka, and Masashi Ishii. 2024. PolyNERE: A novel ontology and corpus for named entity recognition and relation extraction in polymer science domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12856–12866, Torino, Italia. ELRA and ICCL.

Jakub Piskorski, Michał Marcińczuk, and Roman Yangarber. 2024. Cross-lingual named entity corpus for Slavic languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4143–4157, Torino, Italia. ELRA and ICCL.

Hanieh Poostchi, Ehsan Zare Borzeshi, Mohammad Abdous, and Massimo Piccardi. 2016. PersoNER: Persian named-entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3381–3389, Osaka, Japan. The COLING 2016 Organizing Committee.

Peng Qi and Koichi Yasuoka. 2019. UD_Chinese-GSDSimp.

Qwen Team. 2024. QwQ: Reflect deeply on the boundaries of the unknown.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206, Pisa, Italy. Linköping University Electronic Press.

Shoval Sade, Amit Seker, and Reut Tsarfaty. 2018. The Hebrew Universal Dependency treebank: Past present and future. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 133–143, Brussels, Belgium. Association for Computational Linguistics.

Joan Santoso, Patrick Sutanto, Billy Cahyadi, and Esther Setiawan. 2024. Pushing the limits of low-resource NER using LLM artificial data generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9652–9667, Bangkok, Thailand. Association for Computational Linguistics.

Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content North-African Arabizi treebank: Tackling hell. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139–1150, Online. Association for Computational Linguistics.

Mo Shen, Ryan McDonald, Daniel Zeman, and Peng Qi. 2016. UD_Chinese-GSD.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Khalil Sima'an, Alon Itai, Yoad Winter, Alon Altman, and Noa Nativ. 2001. Building a treebank of modern hebrew text. *Traitement Automatique des Langues*, 42(2):347–380.

Anil Kumar Singh. 2008. Named entity recognition for south and south East Asian languages: Taking stock. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*.

Manuela Speranza. 2009. The named entity recognition task at EVALITA 2009.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

David Suba, Marek Suppa, Jozef Kubik, Endre Hamerlik, and Martin Takac. 2023. WikiGoldSK: Annotated dataset, baselines and few-shot learning experiments for Slovak named entity recognition. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 138–145, Dubrovnik, Croatia. Association for Computational Linguistics.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Kaung Lwin Thant, Kwankamol Nongpong, Ye Kyaw Thu, Thura Aung, Khaing Hsu Wai, and Thazin Myint Oo. 2025. myner: Contextualized burmese named entity recognition with bidirectional lstm and fasttext embeddings via joint training with pos tagging. *Preprint*, arXiv:2504.04038.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. 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.

Fabián Villena, Luis Miranda, and Claudio Aracena. 2024. llmner: (zerolfew)-shot named entity recognition, exploiting the power of large language models. *Preprint*, arXiv:2406.04528.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. KazNERD: Kazakh named entity recognition dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 417–426, Marseille, France. European Language Resources Association.

Daniel Zeman. 2017. Slovak dependency treebank in universal dependencies. *Journal of Linguistics/Jazykovedný casopis*, 68:385 – 395.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. UniversalNER: Targeted distillation from large language models for open named entity recognition. *Preprint*, arXiv:2308.03279.

A Appendix

A.1 Data Formatting Corrections

hr500k, ssj500k, and NorNE are represented in CoNLL-U Plus format, which does not explicitly include `O` tags. We converted these datasets to CoNLL format using SeqScore.

In the KIND corpus, each token is annotated with just the type name (e.g. `LOC`). We converted to BIO encoding by prepending all type labels with I-, and then using SeqScore to convert from IO to BIO encoding.

The L3Cube MahaNER dataset delineates sentence breaks with sentence IDs. We added appropriate newlines. We also standardized the encoding prefixes to be separate from the type name with a dash (e.g. `BNEO` to `B-NEO`, `BLOC` to `B-LOC`).

The data for MasakhaNER is taken from commit `9745180390b3507858ea57f7b1e4f8a944d280fc` due to later commits splitting sentences at an arbitrary maximum length, causing some entity mentions to be split across two sentences. Two lines in MasakhaNER 2.0 contain only O labels, with no corresponding tokens. We removed these two lines.

RONEC is distributed in JSON format with BIO-encoded labels and tokens as fields. We converted it to CoNLL format.

The ThaiNNER dataset uses BIOES encoding, and uses a nested ontology. We used the top layer of the nested annotation and converted the encoding to BIO. ThaiNNER comes with two levels of entity types, coarse-grained and fine-grained. We use the coarse-grained types and the top layer of annotation here since the coarse-grained types directly align with most OntoNotes entity types which aligns best with the majority of other corpora we collected.

CoNLL-02 is distributed using ISO-8859-1 text encoding. We converted the files to UTF-8.

Unfortunately, the majority of the NER corpora included in OpenNER do not contain document boundary information. Sometimes this is due to copyright limitations in the original dataset, but often it was due to poor data preparation practices and many old NER data preparation scripts removing document boundaries. OpenNER retains any document boundaries if it was included in the source data. CoNLL-02 Dutch, the UNER corpora (except Maghrebi Arabic), TurkuNLP, hr500k, and ssj500k retain document boundaries. OpenNER uses the convention introduced by CoNLL-02 of marking document boundaries with the sentence `-DOCSTART- O`.

A.2 Label Repair

Invalid label sequences in datasets need to be reviewed manually to ensure the problem is just an issue with invalid transitions and not an annotation error. Once reviewed manually, many errors can be

repaired automatically. For example, when SeqScore encounters the label sequence `O I-PER I-PER O` in a dataset that is supposed to be BIO encoded, it is repaired to `O B-PER I-PER O` using the same approach taken in the original `conlleval` script.

In most cases, automatic repair using SeqScore is possible after brief manual review. This approach corrected 108 errors across the included datasets. In 32 cases for SLI Galician, manual repairs were performed. While most of these could be repaired using the `conlleval` approach, there were 14 which would have been incorrectly labeled using an automatic repair.

A.3 Hyperparameters and Computational Resources

Models are trained using HuggingFace's transformers package (Wolf et al., 2020). We use an encoder model with a TokenClassification head (linear layer). No CRF is used in our experiments. The first subtoken of each word is used for the label. Hyperparameters for fine-tuning were set to a learning rate of 5.0e-5, 10 epochs of fine-tuning, weight decay of 0.05, a batch size of 16, and a warm-up ratio of 0.1. Model sizes in terms of number of parameters are as follows: XLM-R: 279 million, mBERT: 179 million, Glot500: 395 million.

Experiments were run on a SLURM cluster with 32 NVIDIA GPUs (16 RTX A5000, 8 A40, 8 RTX 6000 Ada Generation). Approximately 400 GPU hours were used to train and evaluate the mBERT, XLM-R, and Glot500 models, and approximately 384 GPU hours were used to evaluate the LLM models.

A.4 LLM Few-Shot Prompt Example

We conducted a small set of trials with Aya-Expanse 32b and CoNLL-03 English to determine which approach to prompting the LLMs appeared to perform best.

We explored prompting the model to output each token and its BIO tag, but this performed poorly and it was challenging to map the output back to the original text.

We also experimented with prompting the model to return a JSON object with the three entity types and lists of entities. For example, we expect an output from the LLM such as the following:

```
{
  "Person": [],
  "Location": ["AL-AIN",
               "United Arab Emirates"],
  "Organization": []
}
```

Using JSON as the response had an overall F1 of 61.46.

We then prompted the model to return the original sentence "inline" with special markers for the start and end of the entity text. This inline approach scored 67.42, notably better than the response as JSON. Since the inline approach worked best, we further attempted this with 5-shot demonstrations. Inline with demonstrations was our best performing approach at 74.86 F1, so we continued with this approach for our experiments with LLMs across all datasets in OpenNER.

Demonstrations were randomly selected from the training dataset on a per-language per-dataset basis. Example sentences were filtered by length to include only examples with more than 8 tokens. Examples were chosen to show one demonstration with no named entities and the rest of the examples were randomly selected but required to have at least one named entity.

Hyperparameters for LLMs were chosen from the generation configurations from the models themselves. Temperature was set at 0.3 following the generation configuration in Aya-Expanse.

To conduct 5-shot demonstrations, we provided the examples as demonstrations in a conversation template. A conversation history with 5 turns between the user and the chatbot demonstrates how the chatbot should respond to the provided input. Then the user input is provided with the start token for the chatbot to complete its response. An example of the conversation with examples is shown in Table 12. The target sentence to be labeled is the final example "It is a place in Argentina lol".

B Datasets not Included Due to Quality Concerns

Singh (2008) created a dataset for Southeast Asian languages available at `http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=5`. Named entity tags should all be of the form `<ne=X>` where X is one of the types in the ontology. However, many of the types do not match the ontology, and some entity types are blank. For example, correct annotations might be of the form `<ne=NEP>`, `<ne=NEL>`, and `<ne=NEM>`, but there are examples such as `<ne=k1>` and `<ne=>`.

Bahad et al. (2024) created an NER dataset

with 4 Indic languages. The dataset is available at https://github.com/ltrc/IL-NER/tree/main/Datasets and https://huggingface.co/datasets/Sankalp-Bahad/NER-Dataset. We were unable to include this dataset due to errors in the formatting of the labels. Each named entity in BIO or BIOES encoding must have a state indicating whether it is Begin, Inside, Single, etc. but there are 1,662 examples of tags where the state is missing such as `-NEO` or `-NEL` across all four languages. It is nontrivial to correct these errors, since adjacent entities may need different states and corrections cannot be made automatically without having to examine each example. For Hindi, the test set has 30 invalid labels. The training data has at least one missing token and 150 invalid labels. The dev set has 16 invalid labels. For Odia, the training set has 28 invalid labels, the dev set has 1 invalid label, and test set has 1 invalid label. For Telugu, the train set has 1,073 invalid labels, the dev has 181 invalid labels, and the test has 166 invalid labels. For Urdu, the train set has 12 invalid labels, the dev set has 3 invalid labels, and the test set has 1 invalid label.

The myNER dataset (Thant et al., 2025) consists of NER data in Burmese. myNER uses the BIOES tagging scheme and has validation errors, the majority of which appear to occur around parentheses and hyphens as part of dates. Unfortunately, these invalid transitions are too many and not consistent enough to fix automatically. The train set contains 2,388 invalid transitions, the dev set 312, and the test set 280.

Karim et al. (2019) created a dataset for NER in Bangla. Unfortunately, this dataset has 311 invalid label transitions. While it is sometimes possible to repair invalid transitions automatically (for example, `O` to `I` can be automatically converted to `O` to `B`), the training dataset has 26 transitions which have invalid transitions with differing adjacent types. The dev set has 3 of these invalid transitions with adjacent types, and the test set has 4. These invalid transitions can require knowledge of the language in order to ensure that the correct interpretation is chosen.

C Ontologies

C.1 CoNLL-Derived Ontologies

The CoNLL-02 corpus (Tjong Kim Sang, 2002) consists of Spanish and Dutch newswire data and introduces the LOC/ORG/PER/MISC tagset adapted by many other corpora in this collection.

The AnCora corpus (Taulé et al., 2008) performed multi-level linguistic annotation on news data in both Catalan and Spanish. The NER component of the corpus uses the CoNLL-02 ontology, but with OTHER instead of MISC, and with the addition of NUMBER and DATE.

The AQMAR corpus (Mohit et al., 2012) contains NER data sourced from Wikipedia articles in Arabic. We use a version[3] with fixes of invalid label sequences by Liu et al. (2019).

The BarNER corpus (Peng et al., 2024) consists of NER annotation on Bavarian Wikipedia and Twitter data, using CoNLL core types in addition to LANG (language), RELIGION, EVENT, and WOA (work of art). Each entity type can also appear with suffix -part or -deriv, to capture the nominal derivation and compounding common in the language.

The DaNE corpus (Hvingelby et al., 2020) is named entity annotation as an extension of Universal Dependencies. The underlying data is the PAROLE corpus (Keson, 1998), which was built from paragraphs from a Danish Dictionary.

EIEC (Alegria et al., 2006)[4] is a corpus of Basque newswire.

EverestNER is an NER corpus of news articles (Niraula and Chapagain, 2022). It uses the CoNLL-02 ontology without MISC but with EVENT and DATE types.

The GermEval2014 corpus (Benikova et al., 2014) contains data from the 2014 GermEval NER shared task which includes newswire and German Wikipedia data. The tagset used to annotate this corpus is very similar to the CoNLL-02 one, however the MISC type is renamed OTH (other) and subtypes are introduced. These subtypes occur in the form of TYPEderiv and TYPEpart, with deriv signifying a derivation of the original type and part a named entity that is part of a larger entity.

HiNER (Murthy et al., 2022) is a Hindi dataset that is made up of newswire and data from the tourism domain. The tagset used corpus is based on the CoNLL-02 ontology, with additional custom tags added to further specify categories encompassed by the MISC type (FESTIVAL, GAME, LANGUAGE, LITERATURE, RELIGION).

[3] https://github.com/LiyuanLucasLiu/ArabicNER/tree/master

[4] http://www.ixa.eus/node/4486?language=en

The KIND corpus (Paccosi and Palmero Aprosio, 2022) is a multi-domain Italian corpus which uses the CoNLL-02 types without MISC. The domains included are literature, political discourse, and Wikinews. During preprocessing the train and test sets across all domains were concatenated. The dataset did not contain a development set.

hr500k is corpus of morpho-syntactic annotation on Croatian web and news data (Ljubešić et al., 2016). L3Cube-MahaNER (Litake et al., 2022) is a Marathi news dataset for named entity recognition.

The MasakhaNER version 1.0 dataset (Adelani et al., 2021) is a multilingual dataset that contains local news data in 10 different African languages. It uses the CoNLL-02 types without MISC and with the addition of DATE. We also include MasakhaNER 2.0 (Adelani et al., 2022), which uses the same ontology but covers additional languages.

NEMO^2 (Bareket and Tsarfaty, 2021) consists of both morpheme and token based NER annotation on the Hebrew Treebank (Sima'an et al., 2001), based on the CoNLL-02 guidelines but also adopting GPE, facility, work of art, language, product, and event types. Parallel annotations are provided for the UD version of the Hebrew Treebank (Sade et al., 2018).

NorNE (Jørgensen et al., 2020) is an NER corpus containing both Norwegian Bokmål (nob) and Nynorsk (nno) standards. The corpus is mainly news data, but also contains government reports, parliamentary transcripts and blog posts. The ontology is CoNLL-02-like but includes GPE_LOC and GPE_ORG. EVT and PROD are also included.

A subset of the SLI Galician CTG corpora (Agerri et al., 2018b), from the news and environmental sciences domains, has been annotated for NER, following the CoNLL guidelines.

ssj500k (Dobrovoljc et al., 2017) uses the CoNLL-02 ontology. It contains data from fiction, non-fiction, periodical and Wikipedia texts. Since canonical splits did not appear to exist we created splits in a 80/10/10 manner following the approach used in the GitHub repository.[5]

WikiGoldSK (Suba et al., 2023) is Slovak NER on Wikipedia data with the CoNLL-02 ontology.

The Turku NER corpus (Luoma et al., 2020) is a Finnish corpus that builds on the original Universal Dependencies Finnish corpus (Nivre et al., 2016), which is made up of multi-domain data including news, web, legal, fiction and political data. It uses the CoNLL-02 tags LOC, PER and ORG, but not MISC. The types PRO (Product), DATE and EVENT are also included.

The Tweebank NER dataset (Jiang et al., 2022) is an English dataset developed by annotating the Tweebank V2 (Liu et al., 2018), the main universal dependency treebank for English Twitter NLP tasks. Tweebank uses standard CoNLL-02 tags.

WNUT17 (Derczynski et al., 2017) annotates web text with emerging entities of 6 entity types which further subdivide ORG into group and corporation.

The ArmanPersoNERCorpus (Poostchi et al., 2016) in Persian extends CoNLL ontology with facility, event, and product.

We also included several UNER datasets: Chinese GSD (Shen et al., 2016; Qi and Yasuoka, 2019), English EWT (Silveira et al., 2014), Maghrebi (Seddah et al., 2020), Portuguese Bosque (Rademaker et al., 2017), SNK (Zeman, 2017), and Swedish Talkbanken (McDonald et al., 2013).

C.2 OntoNotes-Derived Ontologies

elNER (Bartziokas et al., 2020) performs NER annotations on Greek news data based on the OntoNotes ontology, but also provides a CoNLL-derived version by merging and filtering types. We use the OntoNotes version of their data.

NER labels were added to Japanese-GSD-UD (Asahara et al., 2018) by Meganon labs.[6] The ontology has 21 entity types largely following OntoNotes with the addition of TITLE_AFFIX, MOVEMENT, PHONE, and PET_NAME, and the corpus is made up of Wikipedia data.

The KazNERD corpus (Yeshpanov et al., 2022) uses the OntoNotes ontology for annotation of Kazakh.

RONEC (Dumitrescu and Avram, 2020) uses an OntoNotes-like ontology but with some types collapsed (i.e. DATETIME, NAT_REL_POL) and some missing (PROD, LAW). The data included in this dataset is collected from news texts.

Thai NNER (Buaphet et al., 2022) uses a fine-grained NER ontology and 10 coarse-grained top-level types. The coarse-grained types are from the OntoNotes ontology. Because Thai NNER is nested NER, we make use of only the top level entity span. The data is made up of news articles

[5] https://github.com/TajaKuzman/NER-recognition/blob/master/create_NER_task_files.py

[6] https://github.com/megagonlabs/UD_Japanese-GSD

and restaurant reviews. The dataset is syllable and document segmented, but not sentence segmented. This segmentation is why Thai appears to have a comparatively small number of sentences.

D Additional Tables

Additional tables are included on the following pages.

Dataset	Lang.	mBERT	XLM-R	Glot500
AnCora	cat	76.93 $^{\pm 0.20}$	**86.71** $^{\pm 0.12}$	85.82 $^{\pm 0.13}$
AnCora	spa	86.58 $^{\pm 0.07}$	**91.89** $^{\pm 0.05}$	91.78 $^{\pm 0.08}$
AQMAR	ara	73.00 $^{\pm 0.15}$	**75.61** $^{\pm 0.27}$	73.53 $^{\pm 0.51}$
ArmanPersoNER	fas	80.44 $^{\pm 0.09}$	**82.25** $^{\pm 0.12}$	81.32 $^{\pm 0.12}$
BarNER	bar	47.35 $^{\pm 0.42}$	60.53 $^{\pm 0.32}$	**61.97** $^{\pm 0.62}$
CoNLL-02	nld	84.54 $^{\pm 0.13}$	**89.83** $^{\pm 0.09}$	89.29 $^{\pm 0.13}$
CoNLL-02	spa	81.85 $^{\pm 0.13}$	**87.36** $^{\pm 0.13}$	86.72 $^{\pm 0.12}$
DaNE	dan	77.01 $^{\pm 0.31}$	**84.09** $^{\pm 0.26}$	81.84 $^{\pm 0.28}$
EIEC	eus	69.21 $^{\pm 0.24}$	**82.20** $^{\pm 0.27}$	78.91 $^{\pm 0.33}$
elNER	ell	88.35 $^{\pm 0.09}$	**91.99** $^{\pm 0.08}$	91.45 $^{\pm 0.08}$
EverestNER	nep	85.09 $^{\pm 0.12}$	90.18 $^{\pm 0.11}$	**90.30** $^{\pm 0.08}$
GermEval	deu	81.77 $^{\pm 0.12}$	**84.98** $^{\pm 0.11}$	84.65 $^{\pm 0.1}$
HiNER	hin	88.99 $^{\pm 0.01}$	**90.00** $^{\pm 0.02}$	88.69 $^{\pm 1.11}$
hr500k	hrv	75.54 $^{\pm 0.16}$	**87.03** $^{\pm 0.13}$	86.78 $^{\pm 0.09}$
Japanese GSD	jpn	81.08 $^{\pm 0.21}$	**83.46** $^{\pm 0.24}$	80.87 $^{\pm 0.23}$
KazNERD	kaz	95.48 $^{\pm 0.04}$	**96.38** $^{\pm 0.04}$	86.37 $^{\pm 9.6}$
KIND	ita	83.13 $^{\pm 0.06}$	**86.62** $^{\pm 0.12}$	**86.62** $^{\pm 0.07}$
L3Cube MahaNER	mar	81.10 $^{\pm 0.14}$	83.05 $^{\pm 0.15}$	**83.31** $^{\pm 0.16}$
MasakhaNER	amh	00.00 $^{\pm 0.00}$	**70.33** $^{\pm 0.45}$	57.05 $^{\pm 9.52}$
MasakhaNER	hau	81.60 $^{\pm 0.37}$	**89.03** $^{\pm 0.21}$	88.66 $^{\pm 0.23}$
MasakhaNER	ibo	76.20 $^{\pm 0.31}$	83.66 $^{\pm 0.33}$	**86.33** $^{\pm 0.28}$
MasakhaNER	kin	61.40 $^{\pm 0.64}$	71.65 $^{\pm 0.47}$	**76.07** $^{\pm 0.18}$
MasakhaNER	lug	72.38 $^{\pm 0.24}$	77.74 $^{\pm 0.41}$	**83.00** $^{\pm 0.27}$
MasakhaNER	luo	61.03 $^{\pm 1.10}$	**71.01** $^{\pm 0.65}$	54.42 $^{\pm 2.77}$
MasakhaNER	pcm	80.02 $^{\pm 0.26}$	86.87 $^{\pm 0.25}$	**88.90** $^{\pm 0.19}$
MasakhaNER	swa	82.24 $^{\pm 0.25}$	**86.80** $^{\pm 0.22}$	85.67 $^{\pm 0.26}$
MasakhaNER	wol	44.36 $^{\pm 4.94}$	62.03 $^{\pm 0.49}$	**65.24** $^{\pm 0.83}$
MasakhaNER	yor	72.88 $^{\pm 0.25}$	75.10 $^{\pm 0.44}$	**81.17** $^{\pm 0.37}$
MasakhaNER 2.0	bam	77.65 $^{\pm 0.24}$	78.86 $^{\pm 0.35}$	**79.70** $^{\pm 0.22}$
MasakhaNER 2.0	bbj	69.57 $^{\pm 0.36}$	**71.66** $^{\pm 0.51}$	68.52 $^{\pm 0.44}$
MasakhaNER 2.0	ewe	81.12 $^{\pm 0.14}$	87.83 $^{\pm 0.19}$	**89.17** $^{\pm 0.18}$
MasakhaNER 2.0	fon	77.81 $^{\pm 0.27}$	80.85 $^{\pm 0.26}$	**82.10** $^{\pm 0.18}$
MasakhaNER 2.0	hau	71.44 $^{\pm 0.28}$	**83.77** $^{\pm 0.22}$	83.62 $^{\pm 0.10}$
MasakhaNER 2.0	ibo	81.60 $^{\pm 0.20}$	86.49 $^{\pm 0.33}$	**89.39** $^{\pm 0.36}$
MasakhaNER 2.0	kin	77.86 $^{\pm 0.22}$	81.85 $^{\pm 0.26}$	**86.08** $^{\pm 0.12}$
MasakhaNER 2.0	lug	84.06 $^{\pm 0.17}$	86.23 $^{\pm 0.19}$	**88.51** $^{\pm 0.13}$
MasakhaNER 2.0	luo	74.19 $^{\pm 0.21}$	79.31 $^{\pm 0.19}$	**81.87** $^{\pm 0.2}$
MasakhaNER 2.0	mos	60.30 $^{\pm 0.28}$	73.29 $^{\pm 0.35}$	**76.03** $^{\pm 0.28}$
MasakhaNER 2.0	nya	85.66 $^{\pm 0.22}$	89.42 $^{\pm 0.09}$	**91.64** $^{\pm 0.09}$
MasakhaNER 2.0	pcm	84.00 $^{\pm 0.14}$	88.34 $^{\pm 0.15}$	**89.24** $^{\pm 0.10}$
MasakhaNER 2.0	sna	89.49 $^{\pm 0.11}$	92.93 $^{\pm 0.20}$	**95.05** $^{\pm 0.09}$
MasakhaNER 2.0	swa	89.75 $^{\pm 0.07}$	91.86 $^{\pm 0.07}$	**92.02** $^{\pm 0.06}$
MasakhaNER 2.0	tsn	82.08 $^{\pm 0.21}$	85.15 $^{\pm 0.29}$	**87.79** $^{\pm 0.17}$
MasakhaNER 2.0	twi	72.08 $^{\pm 0.23}$	77.49 $^{\pm 0.39}$	**80.16** $^{\pm 0.36}$
MasakhaNER 2.0	wol	77.43 $^{\pm 0.16}$	81.00 $^{\pm 0.55}$	**85.48** $^{\pm 0.22}$
MasakhaNER 2.0	xho	78.31 $^{\pm 0.18}$	86.33 $^{\pm 0.07}$	**87.91** $^{\pm 0.17}$
MasakhaNER 2.0	yor	82.00 $^{\pm 0.18}$	85.30 $^{\pm 0.22}$	**86.76** $^{\pm 0.36}$
MasakhaNER 2.0	zul	72.80 $^{\pm 0.19}$	83.65 $^{\pm 0.33}$	**86.53** $^{\pm 0.23}$
NEMO SPMRL	heb	76.68 $^{\pm 0.35}$	**80.02** $^{\pm 0.43}$	76.60 $^{\pm 0.56}$
NEMO UD	heb	73.80 $^{\pm 0.28}$	**76.44** $^{\pm 0.49}$	74.40 $^{\pm 0.38}$
NorNE	nno	78.53 $^{\pm 0.38}$	85.30 $^{\pm 0.32}$	**85.48** $^{\pm 0.25}$
NorNE	nob	74.26 $^{\pm 0.27}$	**87.14** $^{\pm 0.21}$	85.40 $^{\pm 0.23}$
RONEC	ron	86.14 $^{\pm 0.04}$	**88.65** $^{\pm 0.05}$	87.82 $^{\pm 0.07}$
SLI Galician Corpora	glg	76.16 $^{\pm 0.16}$	**87.08** $^{\pm 0.24}$	85.94 $^{\pm 0.25}$
ssj500k	slv	51.73 $^{\pm 0.65}$	**60.79** $^{\pm 0.42}$	55.51 $^{\pm 6.17}$
ThaiNNER	tha	64.34 $^{\pm 0.03}$	71.94 $^{\pm 0.18}$	**72.19** $^{\pm 0.05}$
TurkuNLP	fin	78.04 $^{\pm 0.22}$	**87.04** $^{\pm 0.18}$	86.08 $^{\pm 0.25}$
Tweebank	eng	52.59 $^{\pm 0.32}$	**60.93** $^{\pm 2.06}$	50.45 $^{\pm 2.79}$
UNER Chinese GSD	cmn	**87.15** $^{\pm 0.21}$	85.10 $^{\pm 0.29}$	86.37 $^{\pm 0.21}$
UNER Chinese GSDSIMP	cmn	**87.52** $^{\pm 0.21}$	84.75 $^{\pm 0.39}$	77.29 $^{\pm 8.59}$
UNER English EWT	eng	75.46 $^{\pm 0.22}$	80.61 $^{\pm 0.30}$	**81.53** $^{\pm 0.24}$
UNER Maghrebi Arabic	arq	**81.30** $^{\pm 0.35}$	73.30 $^{\pm 1.66}$	65.28 $^{\pm 1.30}$
UNER Portuguese-Bosque	por	80.73 $^{\pm 0.23}$	**88.48** $^{\pm 0.22}$	87.74 $^{\pm 0.20}$
UNER Slovak SNK	slk	55.83 $^{\pm 0.94}$	**77.44** $^{\pm 0.71}$	73.82 $^{\pm 0.47}$
UNER Swedish Talkbanken	swe	60.33 $^{\pm 10.08}$	**84.42** $^{\pm 0.76}$	72.01 $^{\pm 8.28}$
WikiGoldSK	slk	84.98 $^{\pm 0.18}$	**90.45** $^{\pm 0.31}$	89.70 $^{\pm 0.18}$
WNUT17	eng	33.07 $^{\pm 0.50}$	**50.15** $^{\pm 0.39}$	46.45 $^{\pm 0.45}$
Mean		74.42 $^{\pm 1.78}$	81.79 $^{\pm 1.08}$	80.96 $^{\pm 1.30}$

Table 6: Mean F1 $\pm$ standard error for individual (per language-dataset) models.

		Individual			Multilingual		
Dataset	Lang.	mBERT	XLM-R	Glot500	mBERT	XLM-R	Glot500
AnCora	cat	80.59 $^{\pm 0.13}$	88.71 $^{\pm 0.12}$	87.74 $^{\pm 0.22}$	80.87 $^{\pm 0.15}$	**88.31** $^{\pm 0.15}$	88.27 $^{\pm 0.11}$
AnCora	spa	86.13 $^{\pm 0.08}$	**91.92** $^{\pm 0.07}$	91.74 $^{\pm 0.05}$	85.90 $^{\pm 0.16}$	91.46 $^{\pm 0.11}$	91.37 $^{\pm 0.08}$
AQMAR	ara	78.81 $^{\pm 0.25}$	**80.89** $^{\pm 0.39}$	63.93 $^{\pm 10.66}$	71.02 $^{\pm 0.17}$	76.33 $^{\pm 0.24}$	77.37 $^{\pm 0.25}$
ArmanPersoNER	fas	82.68 $^{\pm 0.11}$	**84.66** $^{\pm 0.06}$	83.62 $^{\pm 0.06}$	81.08 $^{\pm 0.07}$	83.06 $^{\pm 0.07}$	82.81 $^{\pm 0.12}$
BarNER	bar	51.54 $^{\pm 0.48}$	68.47 $^{\pm 0.65}$	71.26 $^{\pm 0.86}$	69.94 $^{\pm 0.65}$	74.35 $^{\pm 0.53}$	**77.70** $^{\pm 0.54}$
CONLL02	nld	85.96 $^{\pm 0.13}$	90.61 $^{\pm 0.12}$	90.11 $^{\pm 0.14}$	85.60 $^{\pm 0.08}$	90.56 $^{\pm 0.12}$	**90.79** $^{\pm 0.10}$
CONLL02	spa	84.42 $^{\pm 0.07}$	**89.05** $^{\pm 0.13}$	88.31 $^{\pm 0.14}$	84.68 $^{\pm 0.18}$	88.40 $^{\pm 0.10}$	88.50 $^{\pm 0.11}$
DaNE	dan	78.41 $^{\pm 0.33}$	**85.32** $^{\pm 0.22}$	83.09 $^{\pm 0.52}$	77.79 $^{\pm 0.40}$	84.02 $^{\pm 0.21}$	82.82 $^{\pm 0.29}$
EIEC	eus	70.96 $^{\pm 0.27}$	**82.66** $^{\pm 0.29}$	80.60 $^{\pm 0.32}$	68.96 $^{\pm 0.24}$	80.96 $^{\pm 0.21}$	79.51 $^{\pm 0.29}$
elNER	ell	86.40 $^{\pm 0.12}$	91.48 $^{\pm 0.04}$	91.32 $^{\pm 0.08}$	86.39 $^{\pm 0.07}$	**91.57** $^{\pm 0.06}$	**91.57** $^{\pm 0.11}$
EverestNER	nep	84.57 $^{\pm 0.12}$	90.41 $^{\pm 0.12}$	**90.55** $^{\pm 0.10}$	85.02 $^{\pm 0.14}$	89.91 $^{\pm 0.11}$	89.90 $^{\pm 0.16}$
GermEval	deu	84.61 $^{\pm 0.10}$	87.57 $^{\pm 0.09}$	**87.61** $^{\pm 0.06}$	83.24 $^{\pm 0.11}$	86.11 $^{\pm 0.09}$	86.57 $^{\pm 0.09}$
HiNER	hin	91.87 $^{\pm 0.02}$	**92.73** $^{\pm 0.01}$	92.06 $^{\pm 0.66}$	91.21 $^{\pm 0.02}$	92.35 $^{\pm 0.01}$	92.34 $^{\pm 0.02}$
hr500k	hrv	78.32 $^{\pm 0.12}$	88.47 $^{\pm 0.13}$	**88.49** $^{\pm 0.09}$	78.62 $^{\pm 0.10}$	87.97 $^{\pm 0.13}$	88.18 $^{\pm 0.13}$
Japanese GSD	jpn	79.04 $^{\pm 0.52}$	**80.75** $^{\pm 0.62}$	63.36 $^{\pm 10.57}$	74.63 $^{\pm 0.50}$	78.01 $^{\pm 0.28}$	78.75 $^{\pm 0.34}$
KazNERD	kaz	**91.80** $^{\pm 0.13}$	85.20 $^{\pm 8.89}$	83.65 $^{\pm 9.32}$	91.40 $^{\pm 0.10}$	94.11 $^{\pm 0.09}$	94.29 $^{\pm 0.07}$
KIND	ita	83.16 $^{\pm 0.09}$	86.56 $^{\pm 0.07}$	**86.65** $^{\pm 0.09}$	81.57 $^{\pm 0.12}$	85.98 $^{\pm 0.08}$	86.30 $^{\pm 0.13}$
L3Cube MahaNER	mar	79.14 $^{\pm 0.18}$	**82.75** $^{\pm 0.16}$	66.07 $^{\pm 11.01}$	78.34 $^{\pm 0.16}$	81.70 $^{\pm 0.19}$	82.23 $^{\pm 0.21}$
MasakhaNER	amh	00.00 $^{\pm 0.00}$	69.40 $^{\pm 0.78}$	**71.25** $^{\pm 0.61}$	00.00 $^{\pm 0.00}$	70.52 $^{\pm 0.51}$	70.95 $^{\pm 0.31}$
MasakhaNER	hau	82.45 $^{\pm 0.18}$	**90.62** $^{\pm 0.20}$	89.95 $^{\pm 0.14}$	85.13 $^{\pm 0.24}$	89.13 $^{\pm 0.18}$	90.08 $^{\pm 0.17}$
MasakhaNER	ibo	76.80 $^{\pm 0.26}$	85.85 $^{\pm 0.31}$	88.06 $^{\pm 0.46}$	84.66 $^{\pm 0.23}$	88.19 $^{\pm 0.20}$	**90.55** $^{\pm 0.21}$
MasakhaNER	kin	60.00 $^{\pm 0.67}$	72.87 $^{\pm 0.41}$	76.73 $^{\pm 0.30}$	76.29 $^{\pm 0.22}$	79.83 $^{\pm 0.20}$	**83.07** $^{\pm 0.11}$
MasakhaNER	lug	72.20 $^{\pm 0.40}$	79.33 $^{\pm 0.43}$	83.79 $^{\pm 0.57}$	76.45 $^{\pm 0.24}$	81.51 $^{\pm 0.24}$	**85.56** $^{\pm 0.18}$
MasakhaNER	luo	61.29 $^{\pm 0.70}$	72.92 $^{\pm 0.70}$	60.58 $^{\pm 4.33}$	82.20 $^{\pm 0.54}$	82.35 $^{\pm 0.35}$	**82.82** $^{\pm 0.26}$
MasakhaNER	pcm	78.13 $^{\pm 0.32}$	85.82 $^{\pm 0.57}$	88.62 $^{\pm 0.27}$	86.51 $^{\pm 0.24}$	90.21 $^{\pm 0.27}$	**91.38** $^{\pm 0.26}$
MasakhaNER	swa	81.88 $^{\pm 0.23}$	87.37 $^{\pm 0.16}$	86.74 $^{\pm 0.16}$	87.32 $^{\pm 0.17}$	89.29 $^{\pm 0.22}$	**89.70** $^{\pm 0.12}$
MasakhaNER	wol	53.68 $^{\pm 0.41}$	67.84 $^{\pm 0.48}$	69.29 $^{\pm 1.31}$	69.09 $^{\pm 0.61}$	72.40 $^{\pm 0.42}$	**75.55** $^{\pm 0.32}$
MasakhaNER	yor	72.69 $^{\pm 0.20}$	76.39 $^{\pm 0.36}$	80.36 $^{\pm 0.94}$	82.97 $^{\pm 0.16}$	85.10 $^{\pm 0.15}$	**88.43** $^{\pm 0.18}$
MasakhaNER 2.0	bam	76.21 $^{\pm 0.27}$	77.82 $^{\pm 0.26}$	78.10 $^{\pm 0.47}$	75.24 $^{\pm 0.27}$	80.14 $^{\pm 0.18}$	**81.07** $^{\pm 0.25}$
MasakhaNER 2.0	bbj	68.52 $^{\pm 0.38}$	70.69 $^{\pm 0.29}$	69.26 $^{\pm 0.34}$	72.42 $^{\pm 0.33}$	72.93 $^{\pm 0.31}$	**73.32** $^{\pm 0.22}$
MasakhaNER 2.0	ewe	84.41 $^{\pm 0.14}$	88.96 $^{\pm 0.15}$	90.69 $^{\pm 0.15}$	88.60 $^{\pm 0.10}$	90.78 $^{\pm 0.12}$	**91.68** $^{\pm 0.12}$
MasakhaNER 2.0	fon	80.00 $^{\pm 0.35}$	83.60 $^{\pm 0.16}$	84.68 $^{\pm 0.23}$	82.02 $^{\pm 0.25}$	84.74 $^{\pm 0.30}$	**86.70** $^{\pm 0.37}$
MasakhaNER 2.0	hau	71.78 $^{\pm 0.37}$	84.99 $^{\pm 0.20}$	85.22 $^{\pm 0.23}$	80.64 $^{\pm 0.13}$	84.20 $^{\pm 0.23}$	**85.47** $^{\pm 0.21}$
MasakhaNER 2.0	ibo	83.30 $^{\pm 0.19}$	89.70 $^{\pm 0.28}$	90.31 $^{\pm 0.30}$	88.71 $^{\pm 0.24}$	93.78 $^{\pm 0.06}$	**94.09** $^{\pm 0.54}$
MasakhaNER 2.0	kin	78.87 $^{\pm 0.17}$	84.72 $^{\pm 0.24}$	**88.38** $^{\pm 0.16}$	84.65 $^{\pm 0.16}$	86.08 $^{\pm 0.12}$	88.16 $^{\pm 0.10}$
MasakhaNER 2.0	lug	86.94 $^{\pm 0.21}$	89.92 $^{\pm 0.10}$	91.87 $^{\pm 0.08}$	88.10 $^{\pm 0.08}$	90.79 $^{\pm 0.11}$	**92.12** $^{\pm 0.08}$
MasakhaNER 2.0	luo	74.38 $^{\pm 0.18}$	80.00 $^{\pm 0.20}$	**82.71** $^{\pm 0.20}$	77.01 $^{\pm 0.21}$	80.99 $^{\pm 0.24}$	82.40 $^{\pm 0.15}$
MasakhaNER 2.0	mos	61.07 $^{\pm 0.42}$	75.94 $^{\pm 0.42}$	76.62 $^{\pm 0.44}$	65.29 $^{\pm 0.33}$	**76.72** $^{\pm 0.21}$	75.86 $^{\pm 0.38}$
MasakhaNER 2.0	nya	86.75 $^{\pm 0.22}$	91.07 $^{\pm 0.11}$	**93.22** $^{\pm 0.08}$	87.67 $^{\pm 0.09}$	90.72 $^{\pm 0.05}$	92.80 $^{\pm 0.09}$
MasakhaNER 2.0	pcm	82.79 $^{\pm 0.26}$	87.92 $^{\pm 0.10}$	88.82 $^{\pm 0.17}$	84.37 $^{\pm 0.18}$	87.78 $^{\pm 0.12}$	**88.96** $^{\pm 0.09}$
MasakhaNER 2.0	sna	89.61 $^{\pm 0.15}$	93.60 $^{\pm 0.15}$	95.51 $^{\pm 0.08}$	90.23 $^{\pm 0.12}$	94.58 $^{\pm 0.07}$	**95.54** $^{\pm 0.07}$
MasakhaNER 2.0	swa	91.93 $^{\pm 0.12}$	94.27 $^{\pm 0.07}$	94.17 $^{\pm 0.06}$	92.49 $^{\pm 0.08}$	94.35 $^{\pm 0.08}$	**94.48** $^{\pm 0.07}$
MasakhaNER 2.0	tsn	83.47 $^{\pm 0.42}$	87.83 $^{\pm 0.28}$	89.41 $^{\pm 0.25}$	85.28 $^{\pm 0.17}$	87.22 $^{\pm 0.23}$	**89.58** $^{\pm 0.15}$
MasakhaNER 2.0	twi	74.61 $^{\pm 0.24}$	81.06 $^{\pm 0.37}$	82.46 $^{\pm 0.19}$	75.09 $^{\pm 0.34}$	81.90 $^{\pm 0.30}$	**83.70** $^{\pm 0.33}$
MasakhaNER 2.0	wol	79.26 $^{\pm 0.20}$	82.47 $^{\pm 0.37}$	86.35 $^{\pm 0.20}$	81.68 $^{\pm 0.19}$	86.17 $^{\pm 0.24}$	**87.93** $^{\pm 0.15}$
MasakhaNER 2.0	xho	79.73 $^{\pm 0.14}$	88.94 $^{\pm 0.21}$	89.95 $^{\pm 0.11}$	81.57 $^{\pm 0.13}$	89.25 $^{\pm 0.11}$	**90.66** $^{\pm 0.08}$
MasakhaNER 2.0	yor	81.97 $^{\pm 0.19}$	86.77 $^{\pm 0.18}$	88.32 $^{\pm 0.29}$	82.18 $^{\pm 0.14}$	87.24 $^{\pm 0.14}$	**88.65** $^{\pm 0.15}$
MasakhaNER 2.0	zul	71.73 $^{\pm 0.29}$	84.13 $^{\pm 0.24}$	86.64 $^{\pm 0.22}$	76.18 $^{\pm 0.12}$	86.52 $^{\pm 0.16}$	**89.22** $^{\pm 0.19}$
NEMO SPMRL	heb	79.76 $^{\pm 0.55}$	81.38 $^{\pm 0.23}$	78.20 $^{\pm 0.42}$	86.45 $^{\pm 0.15}$	88.16 $^{\pm 0.14}$	**88.26** $^{\pm 0.14}$
NEMO UD	heb	76.36 $^{\pm 0.48}$	80.28 $^{\pm 0.30}$	**76.82** $^{\pm 0.51}$	73.24 $^{\pm 0.30}$	76.56 $^{\pm 0.24}$	**76.82** $^{\pm 0.26}$
NorNE	nno	80.70 $^{\pm 0.23}$	89.70 $^{\pm 0.26}$	89.74 $^{\pm 0.17}$	81.48 $^{\pm 0.34}$	90.55 $^{\pm 0.24}$	**91.27** $^{\pm 0.23}$
NorNE	nob	73.27 $^{\pm 0.32}$	**88.57** $^{\pm 0.23}$	77.01 $^{\pm 8.56}$	76.10 $^{\pm 0.45}$	88.40 $^{\pm 0.17}$	87.84 $^{\pm 0.18}$
RONEC	ron	83.90 $^{\pm 0.09}$	87.43 $^{\pm 0.07}$	86.27 $^{\pm 0.11}$	84.06 $^{\pm 0.09}$	**87.61** $^{\pm 0.06}$	87.38 $^{\pm 0.06}$
SLI Galician Corpora	glg	79.43 $^{\pm 0.21}$	88.42 $^{\pm 0.14}$	87.70 $^{\pm 0.18}$	79.94 $^{\pm 0.26}$	**89.03** $^{\pm 0.16}$	88.94 $^{\pm 0.16}$
ssj500k	slv	54.26 $^{\pm 0.59}$	63.45 $^{\pm 0.53}$	64.14 $^{\pm 0.37}$	56.80 $^{\pm 0.42}$	**64.63** $^{\pm 0.18}$	64.00 $^{\pm 0.32}$
ThaiNNER	tha	63.75 $^{\pm 0.08}$	**72.79** $^{\pm 0.08}$	72.75 $^{\pm 0.08}$	63.23 $^{\pm 0.06}$	72.29 $^{\pm 0.09}$	72.00 $^{\pm 0.07}$
TurkuNLP	fin	77.42 $^{\pm 0.27}$	**88.22** $^{\pm 0.28}$	86.04 $^{\pm 0.37}$	77.03 $^{\pm 0.33}$	86.18 $^{\pm 0.20}$	85.84 $^{\pm 0.23}$
Tweebank	eng	57.88 $^{\pm 0.59}$	**70.82** $^{\pm 0.27}$	62.04 $^{\pm 2.47}$	63.12 $^{\pm 0.26}$	69.41 $^{\pm 0.36}$	70.40 $^{\pm 0.59}$
UNER Chinese GSD	cmn	**87.40** $^{\pm 0.15}$	85.13 $^{\pm 0.17}$	83.33 $^{\pm 2.69}$	83.75 $^{\pm 0.40}$	85.22 $^{\pm 0.31}$	85.34 $^{\pm 0.24}$
UNER Chinese GSDSIMP	cmn	**87.31** $^{\pm 0.16}$	85.52 $^{\pm 0.20}$	85.52 $^{\pm 0.20}$	83.59 $^{\pm 0.35}$	85.28 $^{\pm 0.30}$	85.56 $^{\pm 0.22}$
UNER English EWT	eng	75.25 $^{\pm 0.20}$	80.96 $^{\pm 0.23}$	**81.31** $^{\pm 0.10}$	75.25 $^{\pm 0.28}$	78.05 $^{\pm 0.30}$	79.19 $^{\pm 0.26}$
UNER Maghrebi Arabic	arq	**81.32** $^{\pm 0.39}$	75.78 $^{\pm 0.32}$	65.09 $^{\pm 1.67}$	79.40 $^{\pm 0.48}$	77.68 $^{\pm 0.56}$	78.16 $^{\pm 0.45}$
UNER Portuguese-Bosque	por	80.90 $^{\pm 0.28}$	**88.77** $^{\pm 0.20}$	88.12 $^{\pm 0.22}$	81.18 $^{\pm 0.23}$	88.21 $^{\pm 0.15}$	87.59 $^{\pm 0.15}$
UNER Slovak SNK	slk	55.54 $^{\pm 0.56}$	76.70 $^{\pm 0.40}$	74.46 $^{\pm 0.54}$	70.23 $^{\pm 0.23}$	**82.80** $^{\pm 0.16}$	82.40 $^{\pm 0.27}$
UNER Swedish Talkbanken	swe	67.34 $^{\pm 6.44}$	76.88 $^{\pm 8.56}$	79.79 $^{\pm 3.20}$	78.65 $^{\pm 0.43}$	**86.49** $^{\pm 0.54}$	84.49 $^{\pm 0.67}$
WikiGoldSK	slk	85.91 $^{\pm 0.19}$	**92.25** $^{\pm 0.15}$	90.94 $^{\pm 0.13}$	83.43 $^{\pm 0.13}$	90.68 $^{\pm 0.18}$	90.65 $^{\pm 0.10}$
WNUT17	eng	38.70 $^{\pm 0.50}$	**53.85** $^{\pm 0.51}$	53.02 $^{\pm 0.62}$	45.23 $^{\pm 0.22}$	52.61 $^{\pm 0.44}$	51.86 $^{\pm 0.28}$
Mean		75.69$^{\pm 1.78}$	83.10$^{\pm 1.08}$	82.10 $^{\pm 1.29}$	78.33$^{\pm 1.57}$	84.18$^{\pm 0.93}$	**84.89**$^{\pm 0.94}$

Table 7: Mean F1 $\pm$ standard error for individual and multilingual models on core types (PER, LOC, ORG).

Dataset	Language	Aya-Expanse 32b	QwQ 32b Preview
AnCora	cat	**74.47** $^{\pm 0.61}$	74.19 $^{\pm 0.68}$
AnCora	spa	**71.28** $^{\pm 1.18}$	71.23 $^{\pm 1.13}$
AQMAR	ara	44.80 $^{\pm 1.94}$	**45.22** $^{\pm 2.06}$
ArmanPersoNER	fas	50.94 $^{\pm 0.44}$	**53.23** $^{\pm 1.35}$
BarNER	bar	**57.04** $^{\pm 2.45}$	**57.04** $^{\pm 2.45}$
CoNLL-02	nld	**72.36** $^{\pm 1.18}$	71.77 $^{\pm 0.66}$
CoNLL-02	spa	74.08 $^{\pm 1.29}$	**74.45** $^{\pm 1.63}$
DaNE	dan	**73.02** $^{\pm 1.96}$	**73.02** $^{\pm 1.96}$
EIEC	eus	**65.03** $^{\pm 0.65}$	**65.03** $^{\pm 0.65}$
elNER	ell	58.95 $^{\pm 2.88}$	**62.11** $^{\pm 0.31}$
EverestNER	nep	**57.53** $^{\pm 0.50}$	**57.53** $^{\pm 0.50}$
GermEval	deu	**69.40** $^{\pm 0.84}$	**69.40** $^{\pm 0.84}$
HiNER	hin	**61.27** $^{\pm 5.71}$	**61.27** $^{\pm 5.71}$
hr500k	hrv	**69.32** $^{\pm 0.99}$	**69.32** $^{\pm 0.99}$
Japanese GSD	jpn	**43.45** $^{\pm 1.23}$	**43.45** $^{\pm 1.23}$
KazNERD	kaz	**41.13** $^{\pm 2.40}$	**41.13** $^{\pm 2.40}$
KIND	ita	69.22 $^{\pm 0.66}$	**69.29** $^{\pm 0.62}$
L3Cube MahaNER	mar	**43.51** $^{\pm 3.38}$	**43.51** $^{\pm 3.38}$
MasakhaNER	amh	**9.67** $^{\pm 0.78}$	**9.67** $^{\pm 0.78}$
MasakhaNER	hau	**76.61** $^{\pm 0.52}$	76.53 $^{\pm 0.60}$
MasakhaNER	ibo	**72.46** $^{\pm 1.78}$	**72.46** $^{\pm 1.78}$
MasakhaNER	kin	**62.13** $^{\pm 1.23}$	**62.13** $^{\pm 1.23}$
MasakhaNER	lug	**68.28** $^{\pm 0.44}$	67.87 $^{\pm 0.18}$
MasakhaNER	luo	**58.52** $^{\pm 1.48}$	**58.52** $^{\pm 1.48}$
MasakhaNER	pcm	62.65 $^{\pm 1.94}$	**64.11** $^{\pm 1.53}$
MasakhaNER	swa	**76.13** $^{\pm 1.21}$	**76.13** $^{\pm 1.21}$
MasakhaNER	wol	**63.35** $^{\pm 0.54}$	**63.35** $^{\pm 0.54}$
MasakhaNER	yor	**70.32** $^{\pm 1.11}$	**70.32** $^{\pm 1.11}$
MasakhaNER 2.0	bam	**59.08** $^{\pm 2.21}$	**59.08** $^{\pm 2.21}$
MasakhaNER 2.0	bbj	**52.15** $^{\pm 1.25}$	**52.15** $^{\pm 1.25}$
MasakhaNER 2.0	ewe	**78.02** $^{\pm 1.15}$	**78.02** $^{\pm 1.15}$
MasakhaNER 2.0	fon	**69.55** $^{\pm 1.84}$	**69.55** $^{\pm 1.84}$
MasakhaNER 2.0	hau	**59.32** $^{\pm 1.47}$	**59.32** $^{\pm 1.47}$
MasakhaNER 2.0	ibo	**58.07** $^{\pm 3.61}$	**58.07** $^{\pm 3.61}$
MasakhaNER 2.0	kin	**59.59** $^{\pm 0.30}$	**59.59** $^{\pm 0.30}$
MasakhaNER 2.0	lug	**76.31** $^{\pm 0.78}$	**76.31** $^{\pm 0.78}$
MasakhaNER 2.0	luo	**61.82** $^{\pm 1.39}$	**61.82** $^{\pm 1.39}$
MasakhaNER 2.0	mos	**65.61** $^{\pm 0.76}$	**65.61** $^{\pm 0.76}$
MasakhaNER 2.0	nya	**66.76** $^{\pm 2.46}$	**66.76** $^{\pm 2.46}$
MasakhaNER 2.0	pcm	**65.51** $^{\pm 0.74}$	**65.51** $^{\pm 0.74}$
MasakhaNER 2.0	sna	**50.21** $^{\pm 2.37}$	**50.21** $^{\pm 2.37}$
MasakhaNER 2.0	swa	**80.07** $^{\pm 1.25}$	**80.07** $^{\pm 1.25}$
MasakhaNER 2.0	tsn	**67.80** $^{\pm 3.45}$	**67.80** $^{\pm 3.45}$
MasakhaNER 2.0	twi	**58.51** $^{\pm 1.07}$	**58.51** $^{\pm 1.07}$
MasakhaNER 2.0	wol	**68.62** $^{\pm 0.34}$	**68.62** $^{\pm 0.34}$
MasakhaNER 2.0	xho	**49.82** $^{\pm 4.46}$	**49.82** $^{\pm 4.46}$
MasakhaNER 2.0	yor	**28.73** $^{\pm 0.18}$	**28.73** $^{\pm 0.18}$
MasakhaNER 2.0	zul	**47.37** $^{\pm 1.49}$	**47.37** $^{\pm 1.49}$
NEMO SPMRL	heb	**48.49** $^{\pm 7.21}$	**48.49** $^{\pm 7.21}$
NEMO UD	heb	**44.67** $^{\pm 0.89}$	**44.67** $^{\pm 0.89}$
NorNE	nno	**65.32** $^{\pm 0.55}$	**65.32** $^{\pm 0.55}$
NorNE	nob	**60.78** $^{\pm 0.98}$	**60.78** $^{\pm 0.98}$
RONEC	ron	**40.90** $^{\pm 0.21}$	40.35 $^{\pm 0.76}$
SLI Galician Corpora	glg	**64.53** $^{\pm 0.52}$	**64.53** $^{\pm 0.52}$
ssj500k	slv	**49.36** $^{\pm 1.46}$	48.43 $^{\pm 1.66}$
ThaiNNER	tha	**32.10** $^{\pm 1.04}$	**32.10** $^{\pm 1.04}$
TurkuNLP	fin	**65.82** $^{\pm 0.79}$	**65.82** $^{\pm 0.79}$
Tweebank	eng	**61.39** $^{\pm 0.94}$	**61.39** $^{\pm 0.94}$
UNER Chinese GSD	cmn	**63.29** $^{\pm 1.31}$	**63.29** $^{\pm 1.31}$
UNER Chinese GSDSIMP	cmn	**65.91** $^{\pm 1.27}$	**65.91** $^{\pm 1.27}$
UNER English EWT	eng	**69.44** $^{\pm 1.70}$	69.26 $^{\pm 1.54}$
UNER Maghrebi Arabic	arq	**50.66** $^{\pm 4.64}$	**50.66** $^{\pm 4.64}$
UNER Portuguese-Bosque	por	**76.63** $^{\pm 0.27}$	**76.63** $^{\pm 0.27}$
UNER Slovak SNK	slk	**67.30** $^{\pm 0.76}$	**67.30** $^{\pm 0.76}$
UNER Swedish Talkbanken	swe	**56.92** $^{\pm 2.09}$	**56.92** $^{\pm 2.09}$
WikiGoldSK	slk	**71.22** $^{\pm 0.91}$	71.02 $^{\pm 1.06}$
WNUT17	eng	**53.97** $^{\pm 0.45}$	52.28 $^{\pm 0.95}$
Mean		60.55$^{\pm 1.64}$	60.66$^{\pm 1.63}$

Table 8: Results for 5-shot demonstrations with LLMs. Each configuration was run with three random seeds.

Language	Code	Family	Branch	Script (in Data)	Spkrs. (10^6)	Wikipedia Articles	XLM-R Train	mBERT Train	Glot500 Train
Amharic	amh	Indo-European	Semitic	Ge'ez	35	15,370	✓		✓
Arabic	ara	Afro-Asiatic	Semitic	Arabic	380	1,242,904	✓	✓	✓
Bambara	bam	Niger-Congo	Mande	Latin	4.2	840			✓
Bavarian German	bar	Indo-European	Germanic	Latin	15	27,169		✓	✓
Ghomálá'	bbj	Niger-Congo	Bantoid	Latin	0.4	0			
Catalan	cat	Indo-European	Romance	Latin	4.1	761,156	✓	✓	✓
Mandarin Chinese	cmn	Sino-Tibetan	Sinitic	Chi. Trad./Simp.	940	1,446,573	✓	✓	✓
Danish	dan	Indo-European	Germanic	Latin	6	302,658	✓	✓	✓
German	deu	Indo-European	Germanic	Latin	95	2,950,458	✓	✓	✓
Greek	ell	Indo-European	Hellenic	Greek	13.5	240,894	✓	✓	✓
English	eng	Indo-European	Germanic	Latin	380	6,895,998	✓	✓	✓
Basque	eus	Isolate	Isolate	Latin	0.8	445,654	✓	✓	✓
Éwé	ewe	Niger-Congo	Volta-Niger	Latin	5	951			✓
Persian Farsi	fas	Indo-European	Indo-Iranian	Perso-Arabic	91	1,038,033	✓	✓	✓
Finnish	fin	Uralic	Finnic	Latin	5	581,741	✓	✓	✓
Fon	fon	Niger-Congo	Volta-Niger	Latin	2.3	2,059			✓
Galician	glg	Indo-European	Romance	Latin	2.4	214,945	✓	✓	✓
Hausa	hau	Afro-Asiatic	Chadic	Latin	54	50,534	✓		✓
Hebrew	heb	Afro-Asiatic	Semitic	Hebrew	5	363,721	✓	✓	✓
Hindi	hin	Indo-European	Indo-Aryan	Devanagari	345	163,371	✓	✓	✓
Croatian	hrv	Indo-European	Slavic	Latin	5.1	222,728	✓	✓	✓
Igbo	ibo	Niger-Congo	Volta-Niger	Latin	31	36,914			✓
Italian	ita	Indo-European	Romance	Latin	65	1,886,223	✓	✓	✓
Japanese	jpn	Japonic	Japanese	Kana/Kanji	123	1,433,365	✓	✓	✓
Kazakh	kaz	Turkic	Kipchak	Cyrillic	16.7	237,780	✓	✓	✓
Kinyarwanda	kin	Niger-Congo	Bantu	Latin	15	7,821			✓
Luganda	lug	Niger-Congo	Bantu	Latin	5.6	3,337			✓
Luo	luo	Nilo-Saharan	Nilotic	Latin	4.2	0			✓
Marathi	mar	Indo-European	Indo-Aryan	Devanagari	83	98,164	✓	✓	✓
Mossi	mos	Niger-Congo	Gur	Latin	6.5	0			✓
Nepali	nep	Indo-European	Indo-Aryan	Devanagari	19	31,357	✓	✓	✓
Dutch	nld	Indo-European	Germanic	Latin	25	2,169,462	✓	✓	✓
Norwegian (Nynorsk)	nno	Indo-European	Germanic	Latin	4.3	171,312	✓	✓	✓
Norwegian (Bokmål)	nob	Indo-European	Germanic	Latin	4.3	636,583	✓	✓	✓
Chichewa	nya	Niger-Congo	Bantu	Latin	7	1,035			✓
Naija	pcm	English Creole	English Creole	Latin	4.7	1,243			✓
Portuguese	por	Indo-European	Romance	Latin	260	1,134,982	✓	✓	✓
Algerian Arabic	arq	Afro-Asiatic	Semitic	Latin	88	0			
Romanian	ron	Indo-European	Romance	Latin	25	493,880	✓	✓	✓
Slovak	slk	Indo-European	Slavic	Latin	5	250,676	✓	✓	✓
Slovenian	slv	Indo-European	Slavic	Latin	2.5	187,001	✓	✓	✓
chiShona	sna	Niger-Congo	Bantu	Latin	6.5	11,448			✓
Spanish	spa	Indo-European	Romance	Latin	500	1,983,918	✓	✓	✓
Kiswahili	swa	Niger-Congo	Bantu	Latin	5.3	84,161	✓	✓	✓
Swedish	swe	Indo-European	Germanic	Latin	10	2,596,219	✓	✓	✓
Thai	tha	Kra-Dai	Tai	Thai	21	167,460	✓		✓
Setswana	tsn	Niger-Congo	Bantu	Latin	5.2	1,889			✓
Akan/Twi	twi	Niger-Congo	Kwa	Latin	8.9	0			✓
Wolof	wol	Niger-Congo	Senegambian	Latin	7.1	1,704			✓
isiXhosa	xho	Niger-Congo	Bantu	Latin	8	2,107	✓		✓
Yoruba	yor	Niger-Congo	Volta-Niger	Latin	45	34,397		✓	✓
Zulu	zul	Niger-Congo	Bantu	Latin	13	11,539			✓

Table 9: Language information for the included datasets.

Entity Type	Count
ADAGE	197
ART	6,547
ART-DERIV	2
ART-PART	9
CARDINAL	38,290
CONTACT	202
CORPORATION	321
CREATIVE_WORK	387
DATE	104,510
DATETIME	9,614
DERIV	1,176
DESIGNATION	980
DISEASE	1,273
EVENT	7,785
EVENT-DERIV	6
EVENT-PART	9
FACILITY	8,825
FESTIVAL	266
GAME	1,762
GPE	41,915
GPE-LOC	5,104
GPE-ORG	938
GROUP	468
LANG-DERIV	64
LANG-PART	6
LANGUAGE	7,127
LAW	857
LITERATURE	847
LOC	412,987
LOC-DERIV	3,871
LOC-PART	699
MEASURE	6,752
MISC	40,901
MISC-DERIV	292
MISC-PART	253
MONEY	7,371
MOVEMENT	65
NON_HUMAN	8
NORP	15,495
NUM	57,371
ORDINAL	8,061
ORG	245,855
ORG-DERIV	85
ORG-PART	1,101
PER	329,843
PER-DERIV	614
PER-PART	251
PERCENT	1,907
PERCENTAGE	4,284
PERIOD	1,188
PET_NAME	18
PHONE	2
POSITION	6,142
PRODUCT	6,133
PROJECT	2,111
QUANTITY	7,496
RELIGION	1,168
RELIGION-DERIV	5
TIME	22,874
TITLE_AFFIX	322
Total	2,816,304

Table 10: Counts of names of each entity type in the standardized version of OpenNER.

Before Mapping	After Mapping
PER, Person, PERSON, per, NEP, pers, person	PER
ORG, Organization, ORGANIZATION, ORGANISATION, org, NEO	ORG
LOC, Location, LOCATION, loc, NEL, location	LOC
MISC, MIS, OTH, MISCELLANEOUS, misc	MISC
DATE, Date, NED	DATE
TIMEX, TIME, NETI	TIME
DATETIME	DATETIME
EVENT, Event, EVT, EVE, event	EVENT
PERIOD	PERIOD
NUMEX, NUMERIC, NUM	NUM
CARDINAL	CARDINAL
ORDINAL	ORDINAL
PERCENTAGE	PERCENTAGE
GPE	GPE
GPE_LOC	GPE-LOC
GPE_ORG	GPE-ORG
FACILITY, FAC, fac	FACILITY
LANGUAGE, LANG, ANG	LANGUAGE
MONEY	MONEY
NORP, NAT_REL_POL	NORP
PRODUCT, PROD, PRO, DUC, pro, product	PRODUCT
QUANTITY	QUANTITY
PROJECT	PROJECT
ART, WORK_OF_ART, WOA	ART
CONTACT	CONTACT
DISEASE	DISEASE
FESTIVAL	FESTIVAL
GAME	GAME
LAW	LAW
LITERATURE	LITERATURE
NON_HUMAN	NON_HUMAN
RELIGION	RELIGION
ADAGE	ADAGE
POSITION	POSITION
PERderiv, deriv-per	PER-DERIV
ORGderiv	ORG-DERIV
LOCderiv	LOC-DERIV
OTHderiv, MISCderiv	MISC-DERIV
PERpart	PER-PART
ORGpart	ORG-PART
LOCpart	LOC-PART
OTHpart, MISCpart	MISC-PART
EVENTderiv	EVENT-DERIV
EVENTpart	EVENT-PART
LANGderiv	LANG-DERIV
LANGpart	LANG-PART
WOAderiv	ART-DERIV
WOApart	ART-PART
RELIGIONderiv	RELIGION-DERIV
DRV	DERIV
ED	DESIGNATION
NEM	MEASURE
group	GROUP
corporation	CORPORATION

Table 11: Mapping from original entity types to standardized entity types.

<|START_OF_TURN_TOKEN|><|SYSTEM_TOKEN|>
You are a powerful conversational AI trained by Cohere to help people. You are augmented by a number of tools, and your job is to use and consume the output of these tools to best help the user. You will see a conversation history between yourself and a user, ending with an utterance from the user. You will then see a specific instruction instructing you what kind of response to generate. When you answer the user's requests, you cite your sources in your answers, according to those instructions.
<|END_OF_TURN_TOKEN|>
<|START_OF_TURN_TOKEN|><|USER_TOKEN|>
Find names of persons, organizations or locations. Label the following sentence with labels where the name is enclosed with the entity type PER, ORG or LOC and @@ ##. For example PER @@ John Smith ##, or ORG @@ Springfield University ## or LOC @@ United Kingdom ## . Find named entities in the following sentence: Check out the kids ' space as soon as you get onboard .
<|END_OF_TURN_TOKEN|>
<|START_OF_TURN_TOKEN|><|CHATBOT_TOKEN|>
Check out the kids ' space as soon as you get onboard .
<|END_OF_TURN_TOKEN|>
<|START_OF_TURN_TOKEN|><|USER_TOKEN|>
Find names of persons, organizations or locations. Label the following sentence with labels where the name is enclosed with the entity type PER, ORG or LOC and @@ ##. For example PER @@ John Smith ##, or ORG @@ Springfield University ## or LOC @@ United Kingdom ## .
Find named entities in the following sentence:
A pair of military pilots embarked on China 's second manned space flight early today , hurtling into orbit on a test mission that could span five days .
<|END_OF_TURN_TOKEN|>
<|START_OF_TURN_TOKEN|><|CHATBOT_TOKEN|>
A pair of military pilots embarked on LOC @@ China ## 's second manned space flight early today , hurtling into orbit on a test mission that could span five days .
<|END_OF_TURN_TOKEN|>
<|START_OF_TURN_TOKEN|><|USER_TOKEN|>
Find names of persons, organizations or locations. Label the following sentence with labels where the name is enclosed with the entity type PER, ORG or LOC and @@ ##. For example PER @@ John Smith ##, or ORG @@ Springfield University ## or LOC @@ United Kingdom ## .
Find named entities in the following sentence:
Very recently , US forces in Al - Anbar made a terrible blunder by accidentally killing Abdul - Razaq Inad Al - Gu'ud , Sheikh of the Al - Bu Nimr clan from the powerful Dulaym tribe .
<|END_OF_TURN_TOKEN|>
<|START_OF_TURN_TOKEN|><|CHATBOT_TOKEN|>Very recently , LOC @@ US ## forces in LOC @@ Al - Anbar ## made a terrible blunder by accidentally killing PER @@ Abdul - Razaq Inad Al - Gu'ud ## , Sheikh of the PER @@ Al - Bu Nimr ## clan from the powerful Dulaym tribe. <|END_OF_TURN_TOKEN|>
<|START_OF_TURN_TOKEN|><|USER_TOKEN|>Find names of persons, organizations or locations. Label the following sentence with labels where the name is enclosed with the entity type PER, ORG or LOC and @@ ##. For example PER @@ John Smith ##, or ORG @@ Springfield University ## or LOC @@ United Kingdom ## .
Find named entities in the following sentence:
Ordinarily , George would come in around noon ; he would ordinarily leave around 5:30 or 6:00 in the evening .<|END_OF_TURN_TOKEN|>
<|START_OF_TURN_TOKEN|><|CHATBOT_TOKEN|>
Ordinarily , PER @@ George ## would come in around noon ; he would ordinarily leave around 5:30 or 6:00 in the evening.
<|END_OF_TURN_TOKEN|>
<|START_OF_TURN_TOKEN|><|USER_TOKEN|>Find names of persons, organizations or locations. Label the following sentence with labels where the name is enclosed with the entity type PER, ORG or LOC and @@ ##. For example PER @@ John Smith ##, or ORG @@ Springfield University ## or LOC @@ United Kingdom ## .
Find named entities in the following sentence:
I received this draft from Niagara Mohawk Marketing , Inc. for our review .<|END_OF_TURN_TOKEN|>
<|START_OF_TURN_TOKEN|><|CHATBOT_TOKEN|>
I received this draft from ORG @@ Niagara Mohawk Marketing , Inc. ## for our review.
<|END_OF_TURN_TOKEN|>
<|START_OF_TURN_TOKEN|><|USER_TOKEN|>
Find names of persons, organizations or locations. Label the following sentence with labels where the name is enclosed with the entity type PER, ORG or LOC and @@ ##. For example PER @@ John Smith ##, or ORG @@ Springfield University ## or LOC @@ United Kingdom ## .
Find named entities in the following sentence:
It is a place in Argentina lol
<|END_OF_TURN_TOKEN|>
<|START_OF_TURN_TOKEN|><|CHATBOT_TOKEN|>

Table 12: Full example of inline filled prompt template with 5–shot demonstrations.