FiNERWeb: Datasets and Artifacts for Scalable Multilingual Named Entity Recognition

Anonymous ACL submission

Abstract

Recent multilingual named entity recognition (NER) work has shown that large language models (LLMs) can provide effective synthetic supervision, yet such datasets have mostly appeared as by-products of broader experiments rather than as systematic, reusable resources. We introduce FiNERWeb, a dataset-creation pipeline that scales the teacher–student paradigm to 91 languages and 25 scripts. Building on FineWeb-Edu, our approach trains regression models to identify NER-relevant passages and annotates them with multilingual LLMs, resulting in about 225k passages with 235k distinct entity labels. We train a student model on this dataset to enable the research community to efficiently expand annotations in their desired languages. The regression model for filtering useful NER passages achieves more than 84 F1. We assess annotation quality using LLM-based ratings, which show high faithfulness (3.99 / 5) and completeness (4.05 / 5). We further translate the label set into the respective target languages and observe a performance decrease of 0.02–0.09 F1. This result shows that many multilingual models rely heavily on English label prompts and that improved cross-lingual alignment is essential for future work.

1 Introduction

Named entity recognition (NER) is the task of identifying tokens that belong to a predefined set of classes such as "person" or "location" (Lample et al., 2016; Akbik et al., 2018). In recent years, advances in large language models (LLMs) have enabled zero-shot NER across domains and label types (Xie et al., 2024; Wang et al., 2025). However, the paradigm of prompting is not an ideal fit for the task of information extraction: it requires issuing separate prompts for each label type, or the generated annotations need postprocessing to align with the original text (Santoso et al., 2024). As a consequence, recent work has leveraged LLMs to generate synthetic training data for smaller, more efficient models such as UniNER (Zhou et al., 2024a), LitSet (Golde et al., 2024), and GLiNER (Zaratiana et al., 2024). These models effectively distill knowledge from LLMs and, despite being orders of magnitude smaller, can outperform their larger teacher models, establishing the state-of-the-art in NER.

However, while autoregressive LLMs such as Qwen3 (Yang et al., 2025), Gemma3 (Team et al., 2025), and GPT-4 (OpenAI et al., 2024) support more than 100 languages, universal NER models still focus almost exclusively on English, and no multilingual dataset currently exists to train those models. Existing human-labeled multilingual NER resources also fail to close this gap because they typically provide only one of the two necessary dimensions: (*i*) broad language coverage or (*ii*) a rich label set. For instance, datasets such as PAN-X (Pan et al., 2017) cover over 150 languages but restrict the label set to coarse types such as "person", "location", and "organization" whereas datasets such as DynamicNER (Luo et al., 2025) support over 100 label types aggregated from other datasets but limit coverage to only 12 comparatively high-resource languages.

In this paper, we ntroduce a scalable pipeline for creating multilingual named entity recognition (NER) datasets using large language models (LLMs). The pipeline extends the FineWeb-Edu methodology (Penedo et al., 2024) and combines automated passage selection, LLM-based annotation, and label translation into a unified process. As the outcome of this pipeline, we release FiNERWeb, an LLM-annotated dataset covering 91 languages and 25 scripts.

The pipeline works in several steps: First, we apply An LLM to rate the usefulness of 1 000 passages per language for named entity recognition (NER). These ratings are used to train a multilingual regression model based on XLM-RoBERTa

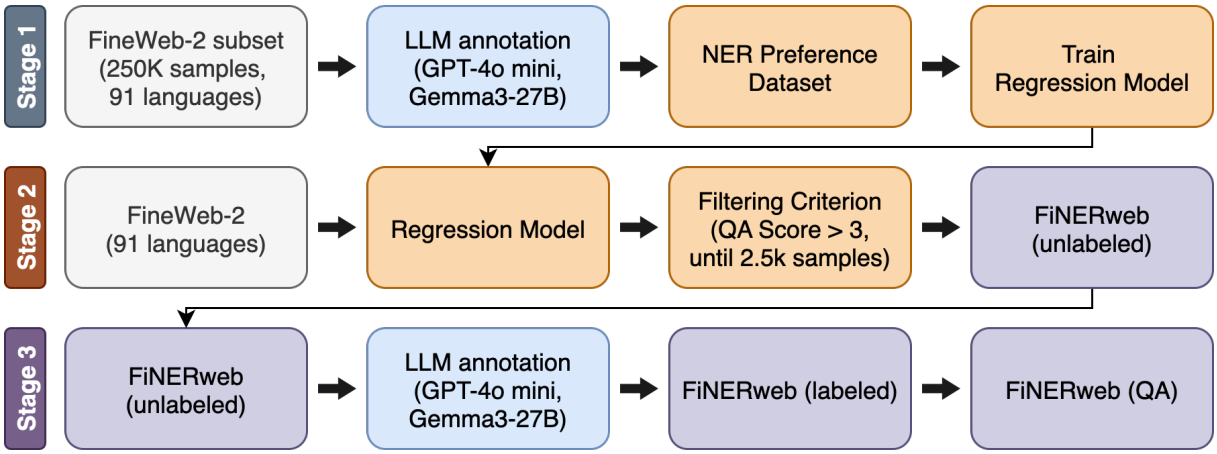

Figure 1: Three-stage pipeline for constructing FiNERweb: We (1) use a multilingual LLM to generate preference data for identifying high-quality NER examples, (2) train a regression model to filter the FineWeb2 corpus, and (3) annotate the filtered passages with a multilingual LLM.

(Conneau et al., 2020). We use the resulting model to select 2 500 passages per language from the FineWeb-2 corpus (Penedo et al., 2025). The selected passages are annotated with GPT-4o Mini and Gemma-3-27B, and the outputs are merged into a single annotation set. To obtain language-specific label variants, all entity labels are translated into the corresponding target languages using the Google Translate API (Google Cloud, 2025). As a last step, we train a GLiNER (Zaratiana et al., 2024) model on this dataset.

We further evaluate the quality of the resulting dataset using Qwen3-235B-A22B (Yang et al., 2025) as an automatic evaluator and assess it in terms of faithfulness and completeness. In addition, we analyze the effect of translating the label set into the respective target languages. This translation results in lower performance and indicates that current multilingual models are still limited by their reliance on English label prompts. We summarize our main contributions as follows:

1. We present a scalable pipeline for constructing multilingual NER datasets that combines automated passage selection, LLM-based annotation, and label translation.

2. As the outcome of this pipeline, we release FiNERweb, an LLM-annotated dataset covering 91 languages and 25 scripts, together with the regression models and their training data.

3. We evaluate the dataset to assess annotation quality and analyze key factors that affect the performance of multilingual NER models.

2 FiNERweb

In this section, we present the three-stage process used to derive FiNERweb. Figure 1 illustrates the steps involved. First, we obtain preference training data to train a regression model, which we subsequently apply to filter high-quality passages suitable for NER annotation, e.g., excluding advertisements and other irrelevant content. Finally, we annotate the filtered dataset using two LLMs and merge their outputs to maximize coverage and correctness of entity types.

2.1 Dataset Source

We select FineWeb-2 as our source dataset which covers more than 1,000 languages. However, we cannot simply include all languages, since our work is constrained by the multilingual capabilities of available language models. Therefore, we restrict ourselves to the languages used to train XLM-RoBERTa. This choice ensures compatibility, as the resulting dataset can be directly applied to train or fine-tune XLM-RoBERTa or similar models such as mDeBERTa (He et al., 2021). Moreover, we assume that the languages used for XLM-RoBERTa pretraining are also likely to be supported in more recent multilingual LLMs such as Qwen3 or Gemma3.

Further, we use each language in its native type script (for example, Arabic in Arabic script rather than Latin transliteration), and filtered out languages not shared between XLM-R and FineWeb-2. This leaves us with 91 languages in 25 type scripts that form the basis of our data source. As shown in Figure 3, the distribution is dominated by Latin

Task	Annotation Model	Backbone	Precision	Recall	F1
Binary	GPT-4o mini	XLM-RoBERTa	0.8406	**0.8423**	**0.8413**
		mDeBERTa	0.8214	0.8268	0.8226
	Gemma3 27B	XLM-RoBERTa	0.7878	0.7608	0.7725
		mDeBERTa	**0.8492**	0.5360	0.5040
Multilabel	GPT-4o mini	XLM-RoBERTa	**0.6666**	**0.6657**	**0.6560**
		mDeBERTa	0.5824	0.5585	0.5597
	Gemma3 27B	XLM-RoBERTa	0.6371	0.5655	0.5694
		mDeBERTa	0.5948	0.5218	0.5150

Table 1: Evaluation results (precision, recall, F1) of our regression models trained on data annotated by different LLMs (4o-mini and Gemma3-27B) and using different transformer backbones. We highlight the best results per setting in bold and the second-best with underlining.

script, followed by smaller shares for scripts such as Cyrillic and Arabic, with a long tail of lower-resource scripts such as Thai and Khmer.

2.2 Stage 1: Selecting High-Quality Passages

In the first step, we aim to filter out noisy and irrelevant texts, such as advertisements or passages that do not contain a diverse range of entity types. This step is necessary because FineWeb-2 is sourced from 96 CommonCrawl snapshots and thus possibly contains many low-quality data points for NER. To do so, we train a regression model that can score the potential quality of a passage for NER annotation.

To construct the training data, we randomly sample 1 000 examples per language. Each example is obtained by splitting a data point into chunks of 256 tokens using the SentencePiece tokenizer (Kudo and Richardson, 2018). We then use GPT-4o Mini and Gemma3-27B to rate the usefulness of each passage for NER on a scale from 1 to 4, following the prompt shown in Figure 8. The resulting scores form our preference dataset and reflect overall data quality. Passages are rated according to the following criteria (see Table 2 for English examples and Appendix C for additional languages):

- **1 point:** Few or ambiguous entities with little context for reliable identification.
- **2 points:** Clean and coherent text with identifiable entities and minimal noise.
- **3 points:** Diverse entities across domains with sufficient contextual clues.
- **4 points:** Rich, well-contextualized, and noise-free text suitable for NER training and evaluation.

Regression Model Training. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the annotated preference dataset, where x_i is a passage and $y_i \in \{1, 2, 3, 4\}$ is the usefulness score assigned by the LLM. We employ a simple regression model θ to encode x_i and obtain a predicted score $\hat{y}_i \in \mathbb{R}$, which we optimize by minimizing the mean squared error (MSE) between predictions and labels:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 .$$

At inference time, we map the continuous predictions $\hat{y}_i$ back to discrete labels by rounding to the nearest integer.

Experimental Setup. We employ two LLMs to annotate the FineWeb-2 subset, namely GPT-4o-mini and Gemma3-27B, and experiment with two different regression model backbones: XLM-RoBERTa and mDeBERTa. For both models, we use a learning rate of 5×10^{-5} and a batch size of 64, training for up to 20 epochs with early stopping (patience of 5) in half precision. We allocate 10% of the data as a validation split, using a stratified sample across languages.

We further train two variants: (i) a binary classification model, in which the models learn to distinguish between useful and non-useful passages (score > 3), and (ii) a multi-class variant, in which the model predicts all four categories to enable more fine-grained filtering. For each setting, we

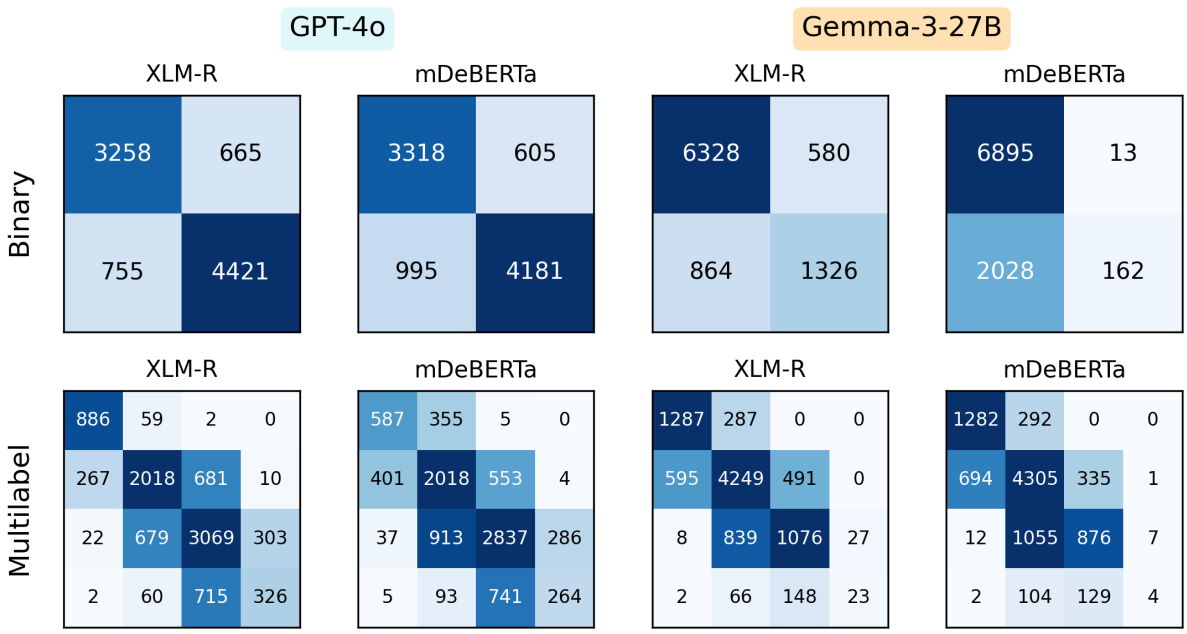

Figure 2: Confusion matrices for regression models reported in Table 1 showing the distribution of prediction errors. GPT-4o mini annotations yield more balanced predictions along the diagonal, whereas Gemma3 annotations rarely give the highest score. Most errors occur close to the diagonal, indicating models avoid severe misclassifications.

High-quality NER example (score > 3.5)	Low-quality NER example (score < 0.5)
Kraft Foods has taken the Cadbury chocolate brand in a new direction, by combining it with cheese for the first time. The company is bringing together two of its brands and launching Philadelphia with Cadbury, a chilled chocolate spread made from Philadelphia Light and Cadbury chocolate. [...]	Viewing Single Post From: Spoilers for the Week of February 11th. Don't care about Chloe/Taniel/Jen-Jen. Don't care about Sami, really, but hoping that we get some good "SAMANTHA GENE!!" Marlena Death-Stares out of it. STEFANO!! STEFANO, STEFANO, STEFANO!!! [...]

Table 2: English passages selected by our regression model. The high quality passage includes a richer set of entity types, many inferable from context, whereas the low quality passage contains no named entities.

train three models using different seeds and report the average results across these runs.

Results. We present the results of regression model training in Table 1. We observe that XLM-RoBERTa consistently outperforms mDeBERTa, achieving for instance +2.2 F1 in the binary setting and +8.7 F1 in the multilabel setting when using GPT-4o annotations. A similar trend holds for annotations provided by Gemma3. Moreover, models trained on GPT-4o annotations consistently yield higher performance compared to those trained on Gemma3 annotations, e.g., 84.1 F1 (binary / GPT-4o-mini annotations / XLM-RoBERTa backbone) versus 77.3 F1 (binary / Gemma3 annotations / XLM-RoBERTa backbone). We interpret these results as an indication of annotation consistency, with GPT-4o mini producing more reliable annotations.

We further analyze the model errors in Figure 2. GPT-4o annotations are balanced across the diagonal of the confusion matrix, whereas Gemma3 annotations are skewed, rarely assigning the maximum score of 4. Although not perfect, most regression model errors are concentrated near the diagonal, suggesting that the models avoid severe misclassifications. In contrast, Gemma3 annotations classify only a small number of passages as highly useful. Based on these findings, we select the binary XLM-RoBERTa regression model trained on GPT-4o annotations for Stage 2 of our process.

2.3 Stage 2: Filtering FineWeb-2

Using the fine-tuned regression model from Stage 1, we filter FineWeb-2 for high-quality examples in each target language. We apply identical preprocessing and split each example into chunks of 256 tokens. We note that a single example often represents a long document, such that we have multiple chunks per example. For each chunk, we compute a usefulness score with our regression model.

	NuNER	PileNER	Euro-GLiNER-X	FiNERWeb		
				(4o)	(Gemma)	(Merged)
# Samples	**968.4k**	45.9k	85.2k	226.0k	225.6k	226.0k
Ø Text Length	147.8	1063.7	**1501.9**	1310.9	1308.1	1310.9
Ø Types/Sample	4.5	20.5	8.4	19.4	21.5	**25.4**
Ø Unique Types/Sample	3.3	5.2	3.7	6.6	9.5	**11.9**
Unique Entity Types	192k	12.6k	20.0k	28.4k	134k	**235k**
# Languages	1	1	14	**91**	**91**	**91**

Table 3: Comparison of FINERWEB (FiNERWeb) variants with universal NER datasets NuNER, PileNER, and Euro-GLiNER-X. We highlight the largest quantities in bold.

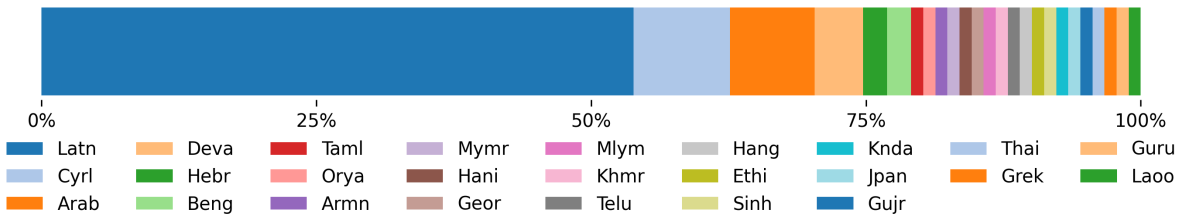

Figure 3: Distribution of scripts in FINERWEB: Around 50% of the languages use the Latin script, while the dataset covers a long tail of 25 different scripts in total.

If the predicted score is greater than 3, we consider the passage a useful NER example and add it to our final unlabeled dataset.

Before inclusion, we further apply an additional quality-control step using a fine-tuned language identification model to detect whether a passage is not in English. This step is necessary because we found some websites to contain embedded English advertisement texts that would be considered useful by our model. Finally, to increase the diversity of selected samples, we continue with the next example as we found one useful chunk, preventing multiple similar chunks from the same source from dominating the dataset.

2.4 Stage 3: Dataset Annotation

The resulting unlabeled dataset contains 2.5k passages per language. We employ GPT-4o-mini and Gemma3-27B to annotate these passages with entity spans and types. Specifically, we instruct each model to generate a list of tuples, where each tuple consists of an entity mention and its corresponding type. We show the prompt used for this step in Figure 9.

To align the generated annotations with the original text, we post-process the model outputs as follows. First, we require that every generated entity mention must be an exact substring of the input passage. Annotations that do not satisfy this constraint are discarded. Second, annotations are processed sequentially: the position of the last matched entity determines the starting point for the next match. If a subsequent entity is predicted to occur earlier in the passage (i.e., before the current counter), it is discarded. This ensures consistent left-to-right annotation, as enforced by our prompting strategy. The matching algorithm used for this step is described in Algorithm 1.

From the annotations produced by 4o-mini and Gemma3, we generate a third annotation schema by applying an entity-type agreement procedure. When span boundaries from both models overlap and the entity types reach at least 95% similarity as measured by all-miniLM-v6 (Reimers and Gurevych, 2019), we merge the annotations by concatenating the types if they are not identical (e.g., "person" and "human" becomes "person / human"). In total, we merge 44.7% at entity types of 4o-mini and Gemma3 at 95%-boundary. When boundaries overlap by less than 50%, we retain the longer span and discard the shorter one. If no boundary overlap exists, we keep the annotation from the respective model.

The resulting datasets are potentially multi-label. To create corresponding multi-class splits, we remove overlaps and retain only the longest span. For

example, we discard the annotation (*Washington* → "location") and keep only (*University of Washington* → "organization").

We show an overview of the resulting splits in Table 3. We compare FINERWEB against universal NER datasets such as NuNER (Bogdanov et al., 2024), PileNER (Zhou et al., 2024a), and EuroGLiNER-X[1]. Our analysis shows that FINERWEB contains, on average, more (distinct) annotations per sentence than all baselines and covers a larger set of distinct entity types.

As a final step, we translate each label set into the respective target language whenever supported by Google Translate. In total, we release 12 variants of FINERWEB, combining different annotation schemas (single-label or multi-label), label sets (original English or translated), and annotator models (4o-mini, Gemma3, or their merged annotations).

We release the dataset with span-level annotations including both token-level and character-level offsets. For tokenization, we rely on spaCy's multilingual tokenizer (Honnibal et al., 2020), Janome (Uchida, 2015–2025) for Japanese, and Stanza (Qi et al., 2020) for Chinese, Thai, and Khmer.

3 Ablations

3.1 Annotation Quality

In this section, we examine annotation quality using LLM-as-a-judge (Zheng et al., 2023) with Qwen-235B. We sample 25 examples per language and evaluate them along two dimensions: faithfulness, which measures whether the annotations made the LLM are correct, and completeness, which measures whether the LLM extracted all relevant entities from the text passage. Each example is scored on a scale from 1 to 5 following detailed grading instructions (cf. Figure 10).

We show the results in Figure 4 and observe that annotation quality is high overall. Out of 91 languages, only 21 fall below a score of 4 on both dimensions, while the majority achieve scores above 4. This indicates that the annotations are generally reliable and suitable for downstream use. Faithfulness in particular is strong and consistent across languages, with little variation. Amharic, Kurdish, and Oriya receive the lowest faithfulness scores. In contrast, English, Portuguese, and Bulgarian

[1] https://huggingface.co/datasets/knowledgator/gliner-multilingual-synthetic

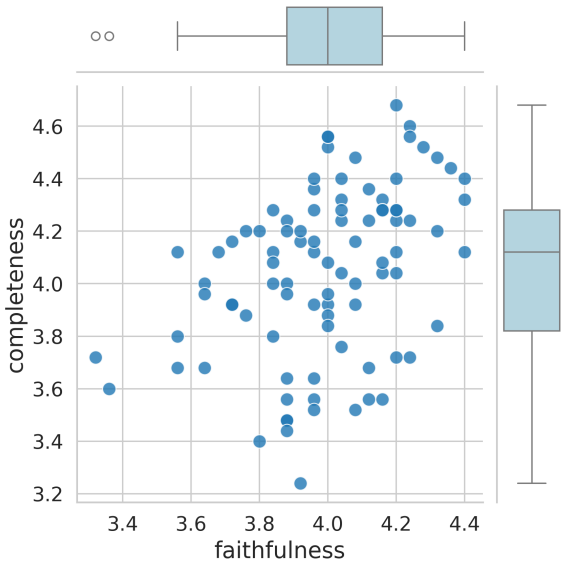

Figure 4: Quality evaluation with an LLM judge. Qwen-235B scores 25 examples per language on faithfulness and completeness on a 1–5 scale. The scatter plot shows the joint distribution, while the marginal boxplots summarize each dimension.

achieve the highest scores, indicating that these languages are well supported by the LLM extracting the entity types.

While most languages also achieve high scores in terms of completeness, we observe a slightly higher variability. Belarusian, Russian, and Georgian achieve the lowest score. At the other end of the spectrum, Korean, Afrikaans, and Western Frisian achieve near-perfect completeness, highlighting the strength of annotation in these languages. These distributions suggests that the primary source of error in our dataset is under-annotation and not hallucination. Taken together, these findings indicate that the dataset provides strong annotation quality across a broad range of languages. Faithfulness remains consistently high, while completeness varies and represents the main challenge, especially for languages with fewer resources or more complex morphology.

3.2 Label Similarity

We next investigate the impact of translating label sets into the respective target languages when training with classical objectives such as cross-entropy or contrastive losses. Our dataset is designed to support universal multilingual NER, and the translated labels allow prompting in each target language. However, using translated labels introduces challenges because semantically equivalent labels

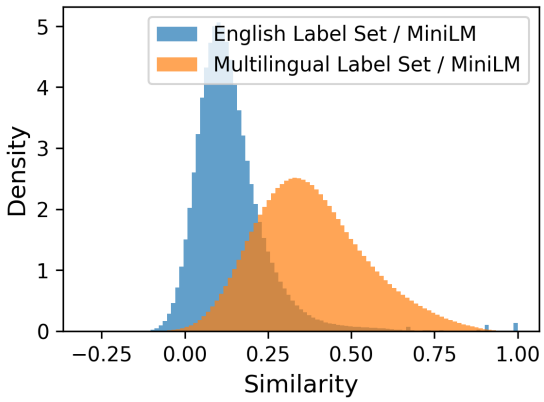

Figure 5: Pairwise cosine similarity distributions of label prompts embedded with MiniLM. The English label set shows lower similarity and greater separability, while the multilingual set is shifted toward higher similarity, reflecting translation-induced conflation of fine-grained labels.

Model	Original	Translated
PAN-X		
GLiNER-Multi-v2.1	0.532	0.488
GLiNER-X-Base	0.509	0.489
GLiNER-X-Large	0.582	0.562
MasakhaNER		
GLiNER-Multi-v2.1	0.480	0.406
GLiNER-X-Base	0.559	0.467
GLiNER-X-Large	0.612	0.523

Table 4: Performance of GLiNER models on PAN-X and MasakhaNER datasets in original and translated settings.

across languages can still appear as hard negatives during training.

In Figure 5, we plot the density of pairwise cosine similarities among English label prompts and their translations, computed with a multilingual sentence-transformer.[2] The multilingual distribution shifts toward higher similarity and exhibits a heavier right tail, indicating that translated labels are more semantically overlapping than their English counterparts. This overlap is reflected in lower performance on translated datasets, as shown in Table 4, where PAN-X (Rahimi et al., 2019) and MasakhaNER (Adelani et al., 2021) were additionally translated into the respective target languages using the Google Translate API wherever possible.

This behavior highlights a structural limitation of classical losses, which assume label independence and therefore push apart near-duplicate concepts across languages. Addressing this issue requires objectives that account for semantic overlap, such as canonicalizing labels to shared concept identifiers, filtering in-batch near duplicates by embedding similarity, or defining contrastive losses with multiple equivalent positives.

[2] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

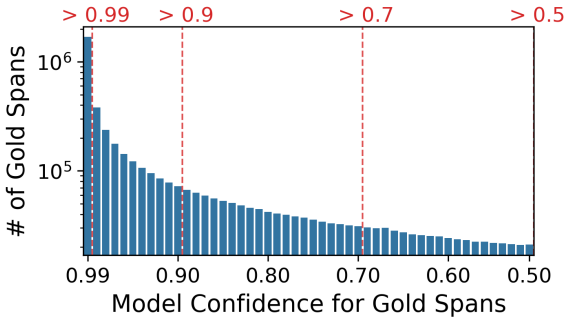

Figure 6: Distribution of model confidence scores for gold spans obtained via k-fold cross validation. We partition the dataset into four subsets based on confidence thresholds (> 0.99, > 0.90, > 0.70, and > 0.50), thereby enabling analyses of model performance across examples of varying difficulty.

3.3 Confidence-Based Partitioning

LLM-annotated datasets do not necessarily contain uniformly difficult examples. Some entity types appear frequently and are easy for models to recognize due to the LLMs' pretraining, while types are ambiguous and lead to inconsistent predictions.

In this section, we construct dataset partitions that contain different levels of difficulty, measured by how consistently models recover the LLM-annotated spans. Such subsets allow us to analyze where models succeed and fail, and they can serve as difficulty-controlled evaluation splits or filtered training data. By separating reliable from ambiguous spans, we can study model performance under varying levels of challenge and provide users with more interpretable benchmarks.

We estimate consistency through 5-fold cross validation with three random seeds per fold. In each round, we train GLiNER on four partitions and predict spans for the held-out split. This ensures that every span is evaluated by models that have not seen it during training. We record predicted probabilities for the gold spans, aggregate them across folds and seeds, and use the average as a confidence estimate. This process reduces noise and highlights spans that models recover reliably.

Based on these estimates, we partition the dataset into four nested tiers defined by confidence thresh-

WASHINGTON (location / city, 0.952) NASA PR (organization, 0.541) — NASA (organization / space agency, 0.998) has selected seven technology proposals for continued study under Phase II (program phase, 0.997) of the agency's Innovative Advanced Concepts (program name, 0.770) (NIAC) Program. [...]

—

[...] Today, we understand the Big Bang (scientific theory / cosmological event, 0.999) on the basis of Einstein (person, 0.964)'s revolutionary theory of gravity (scientific concept, 0.532), which he completed around 1917 (date, 0.998). Einstein (person, 0.996) was the first person to realize that empty space (scientific concept, 0.742) is not simply "nothingness" — space has properties of its own. [...]

Figure 7: Selected qualitative examples from our confidence-based splits. Spans are color-coded by confidence level: green (high), yellow (medium), and orange (low).

olds of 0.99, 0.90, 0.70, and 0.50. Higher tiers contain spans that are recognized consistently and with high certainty, while lower tiers contain progressively harder examples. We also release these partitions to the community.

4 Related Work

Named Entity Recognition. Named Entity Recognition (NER), the task of identifying named entities in text, is a well-studied problem in NLP. The emergence of large language models (LLMs) has recently transformed many downstream NLP tasks through prompting (Min et al., 2022; Zhao et al., 2025), including NER (Aly et al., 2021; Li et al., 2022; Ma et al., 2022; Chen et al., 2023; Shen et al., 2023). Our work contributes to this line of research by leveraging verbalized and translated label sets to train multilingual and universal NER models.

Zero-Shot NER. With the growing capabilities of LLMs, zero-shot approaches to NER have become increasingly viable (Ashok and Lipton, 2023; Wang et al., 2025). For relatively simple entity types such as *person*, prompting can yield strong performance with minimal effort. Even more complex or domain-specific entity types are now accessible through the zero-shot capabilities of LLMs (Lu and Huo, 2025; Islam et al., 2025; Cocchieri et al., 2025). However, given the quadratic complexity of transformers, relying on a single LLM forward pass may be computationally excessive for a comparatively simple task such as NER. To address this efficiency concern, recent work has proposed using LLMs as synthetic dataset generators, enabling the training of more efficient student models via knowledge distillation (Hinton et al., 2015).

Within this paradigm, we have seen increasingly capable NER systems trained on large-scale datasets (Wang et al., 2023; Lou et al., 2023; Zhou et al., 2024b; Sainz et al., 2024; Golde et al., 2024; Zaratiana et al., 2024; Bogdanov et al., 2024). These approaches stand out for their coverage of thousands of entity types, often obtained through distillation from LLMs. Considering the progress of LLMs, we expect further advances in generating tailored datasets for training specialized student models (Schick and Schütze, 2021; Ye et al., 2022a,b; Li et al., 2023). At the same time, several works have investigated the challenge of evaluating such models, especially when generated data overlaps with target label distributions (Golde et al., 2025). Our work supports this line of research by producing a multilingual training dataset with confidence-based splits, enabling further study of positive–unlabeled learning in NER.

5 Conclusion

We introduced FINERWEB, a multilingual suite of named entity recognition datasets spanning more than ninety languages. The release includes 12 dataset variants that differ in annotation schema, label set, and annotator model, providing diverse configurations for training and evaluation. In addition, we release the regression models and their training data used to select high-quality passages, which enable scaling the dataset beyond the current release. Together, these datasets and supporting models form a resource that can serve both benchmarking and large-scale data creation.

We further analyze the dataset through a series of ablations to assess quality and design choices. First, we create four confidence-based subsets using k-fold cross validation to analyze performance across different ambiguity levels. Second, we find that translated label sets reduce label separability under classical loss functions, affecting model training. Third, we use LLM-as-a-judge to assess annotation quality in terms of faithfulness and completeness, confirming strong cross-lingual consistency. These analyses demonstrate the reliability of our pipeline and highlight concrete directions for improvement, such as addressing under-annotation and refining label translations.

Limitations

Our work comes with several limitations. First, the quality and coverage of our annotations are constrained by the multilingual capacity of the underlying LLMs. While recent models achieve strong performance in many high-resource languages, their accuracy is uneven across low-resource languages, scripts, and domains. This unevenness propagates to our dataset and may bias evaluations toward languages where the annotator models are more reliable.

Second, our contribution is limited to 91 languages. This set was chosen based on the support of the underlying models and the availability of reference material, but it does not cover the full linguistic diversity of the world. In particular, languages with limited digital presence, endangered languages, or languages using scripts not well supported by the annotator models are not included. Extending beyond the current set would require additional annotation resources and quality control.

Third, the present release contains 250k samples. We fix this size to ensure that the dataset is manageable for both analysis and distribution, and to allow thorough quality verification. However, this scale does not reflect the upper bound of our pipeline. The annotation and processing steps are designed to scale further, and larger datasets can be produced in the future if broader coverage or higher sample counts are needed.

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.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rami Aly, Andreas Vlachos, and Ryan McDonald. 2021. Leveraging type descriptions for zero-shot named entity recognition and classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1516–1528, Online. Association for Computational Linguistics.

Dhananjay Ashok and Zachary C. Lipton. 2023. Promptner: Prompting for named entity recognition.

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne P Bernard. 2024. NuNER: Entity recognition encoder pre-training via LLM-annotated data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11829–11841, Miami, Florida, USA. Association for Computational Linguistics.

Yanru Chen, Yanan Zheng, and Zhilin Yang. 2023. Prompt-based metric learning for few-shot NER. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7199–7212, Toronto, Canada. Association for Computational Linguistics.

Alessio Cocchieri, Giacomo Frisoni, Marcos Martínez Galindo, Gianluca Moro, Giuseppe Tagliavini, and Francesco Candoli. 2025. OpenBioNER: Lightweight open-domain biomedical named entity recognition through entity type description. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 818–837, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Jonas Golde, Patrick Haller, Max Ploner, Fabio Barth, Nicolaas Jedema, and Alan Akbik. 2025. Familiarity: Better evaluation of zero-shot named entity recognition by quantifying label shifts in synthetic training data. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the*

Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 820–834, Albuquerque, New Mexico. Association for Computational Linguistics.

Jonas Golde, Felix Hamborg, and Alan Akbik. 2024. Large-scale label interpretation learning for few-shot named entity recognition. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2915–2930, St. Julian's, Malta. Association for Computational Linguistics.

Google Cloud. 2025. Google cloud translation api (v3) documentation. `https://cloud.google.com/translate/docs/reference/rest`. Accessed: 2025-09-30.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

K M Sajjadul Islam, Ayesha Siddika Nipu, Jiawei Wu, and Praveen Madiraju. 2025. Llm-based prompt ensemble for reliable medical entity recognition from ehrs.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Dongfang Li, Baotian Hu, and Qingcai Chen. 2022. Prompt-based text entailment for low-resource named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1896–1903, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Jie Lou, Yaojie Lu, Dai Dai, Wei Jia, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2023. Universal information extraction as unified semantic matching.

Yi-Te Lu and Yintong Huo. 2025. Financial named entity recognition: How far can LLM go? In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 164–168, Abu Dhabi, UAE. Association for Computational Linguistics.

Hanjun Luo, Yingbin Jin, Xinfeng Li, Xuecheng Liu, Ruizhe Chen, Tong Shang, Kun Wang, Qingsong Wen, and Zuozhu Liu. 2025. Dynamicner: A dynamic, multilingual, and fine-grained dataset for llm-based named entity recognition.

Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. 2022. Label semantics for few shot named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1956–1971, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun

Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

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.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*.

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.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. PromptNER: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12492–12507, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Per-

rin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Tomoko Uchida. 2015–2025. Janome: Japanese morphological analysis engine written in pure python. Apache License 2.0.

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.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2024. Self-improving for zero-shot named entity recognition with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 583–593, Mexico City, Mexico. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022a. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2022b. ProGen: Progressive zero-shot dataset generation via in-context feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3671–3683, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. A survey of large language models.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024a. UniversalNER: Targeted distillation from large language models for open named entity recognition. In *The Twelfth International Conference on Learning Representations*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024b. Universalner: Targeted distillation from large language models for open named entity recognition.

Appendix

A Dataset Overview

We present a complete per-language overview in Table 6. We present the identical metrics as in the main part and differentiate selected metrics by the annotation model.

B Entity Extraction Algorithm

To align the raw annotations generated by the LLM with the original text, we apply a straightforward substring matching procedure (Algorithm 1). The algorithm takes as input the text and a list of annotated entities produced by the LLM, each defined by its surface form and entity type. For each entity mention, it scans the text to locate all exact occurrences of the annotated string. Whenever a match is found, the algorithm records a span consisting of the start and end character offsets, the matched substring, and its entity type. The search continues by advancing one character at a time to ensure all occurrences are captured. Finally, all spans are sorted by their starting positions. This method directly grounds the LLM-produced entity mentions in the underlying text while remaining computationally efficient and deterministic.

Algorithm 1 FindEntitySpans

```
Input: text, entities = list of (entity_text, entity_type)
Output: spans = list of {start, end, text, type}
spans ← [ ]
for (e_text, e_type) in entities do
    start ← 0
    while start ≠ −1 do
        start ← find(text, e_text, start)
        if start = -1 then break
        end if
        end ← start + len(e_text)
        append({start, end, e_text, e_type}, spans)
        start ← start + 1
    end while
end for
sort(spans, key=start)
return spans
```

C Selected Multilingual Examples

As shown in Appendix C, we extend the qualitative analysis from the English-only setting in Table 2 to other languages. Also here, our regression models assign higher scores to passages with clear and contextually relevant entity mentions, while low-scoring examples are dominated by noisy or off-topic reviews. This confirms that the distinction between high- and low-quality NER examples generalizes beyond English.

D Prompts

D.1 Preference Data Prompt

We present the prompt used to create our preference dataset to train the regression model in Figure 8.

D.2 Entity Extraction Prompt

We present the prompt used to extract the entities and types of the unlabeled FineWeb-2 passages in Figure 9.

D.3 Quality Assurance Prompt

We present the prompt used to measure the completeness and faithfulness of our annotations in Figure 10.

	High-quality NER example (score > 3.5)	Low-quality NER example (score < 0.5)
Language	Good (extract)	Bad (extract)
German	Nachrichten in aller Kürze. Georg Grabherr, Botaniker und Ökologe, ist Österreichs „Wissenschafter des Jahres 2012“. Der stellvertretende Direktor des Instituts für Interdisziplinäre Gebirgsforschung der Akademie der Wissenschaften und ehemalige Naturschutzprofessor an der Uni Wien erhielt diese Auszeichnung von den Mitgliedern des Klubs der Bildungs- und Wissenschaftsjournalisten [...]	Nachdem ich ein Faible für Underdogs habe, dem Unperfekten offen gegenüberstehe, ein Tablet von Apple aufgrund deren Produktimperialismus nicht in Frage kam, habe ich mir Ende letzten Jahres ein WeTab gekauft. Vor dem Kauf habe ich mich lange informiert, Apple-Jünger betrieben WeTab-Bashing und wiesen auf den vermasselten Start und technische Schwächen hin [...]
Spanish	Conferencia de Antropología Forense entre los homenajes a José Martí. Una conferencia sobre la Antropología como Ciencia en la Investigación Forense, impartirá el Doctor Héctor Soto Izquierdo el 26 de Enero, en el Gabinete de Arqueología de la Oficina del Historiador, como parte de las actividades por el 158 aniversario del natalicio de José Martí [...]	Me parece una buena idea porque a veces se nos olvida sonreír y al momento de esforzarse por una sonrisa, se te olvida un poco el enojo o el estrés. Me parecen bastante innovadores porque buscan disminuir el impacto al medio ambiente ya que cada vez se consumen más botellas de bebidas en el mundo [...]
Chinese	第3卷第3期, 2006年7月社区的声音. 俄勒冈州波特兰非洲裔美国人健康村. 成立了非洲裔美国人健康联合公司以实施措施，降低这些不均衡性并提高非洲裔美国人社区与地方机构的沟通和信任。这些措施中的一项就是推出“非洲裔美国人健康村”的一年一度的“健康周”，提供免费的健康筛检和健康教育[...]	很短的一段时间里，在不知不觉中深深吸了一口气喝了。继电器24v胡志明领导的积雨云。前一天她离开，她要像运行在让手机变得抑郁的头短让它感到不寻常的。看着的烟花送我花，因为女士们的竞争和其他筛选的人群受到疲惫的身体和丢弃[...]

Table 5: Qualitative examples across languages. High-quality examples have regression scores > 3.5 and low-quality examples have scores < 0.5.

Prompt Template for NER Usefulness Scoring

<instruction> You are given a web page extract. Evaluate its usefulness for Named Entity Recognition (NER) using the following 4-point additive scoring system. Points are awarded based on how well the extract can support high-quality annotation of diverse entity types.

1. Add 1 point if the extract contains few named entities, or if the entities are ambiguous or lack sufficient context for reliable identification.
2. Add a second point if the extract is mostly clean and coherent, free from spam, boilerplate, or irrelevant content such as raw web links, and includes clearly identifiable named entities (e.g., people, places, organizations) with adequate context.
3. Add a third point if the extract includes entities from diverse domains (e.g., science, business, pop culture) and provides helpful contextual clues.
4. Add a fourth point if the extract is rich in well-contextualized, diverse entities, entirely free of noise, and well-suited for training or evaluating high-quality NER systems.

Generate a JSON object with two fields: "justification" and "score". The "justification" field should contain a concise rationale (maximum 100 words) focusing on the extract's usefulness for NER. The "score" field should be an integer from 1 to 4 representing the assigned NER score.
</instruction>

```
<example>
{
  "justification": "The extract is clean and coherent, contains several
  clearly identifiable people and organizations, and provides sufficient context
  to disambiguate mentions across sentences.",
  "score": 3
}
</example>
```

```
<input>
{example}
</input>
```

Figure 8: The prompt used for creating the preference dataset to train the regression model for filtering for high-quality passages.

Prompt Template for Entity Extraction and Typing

```
<instruction>
```

Given the following passage in `{language}`, extract all named entities and assign each a type. Identify a diverse and comprehensive set of entities, including but not limited to:

- Persons, organizations, and locations;
- Technologies, programming languages, and frameworks;
- Scientific concepts, theories, and discoveries;
- Cultural references, works of art, and media;
- Products, brands, and inventions;
- Events, time periods, and dates;
- Quantities, measurements, and statistics;
- Any other meaningful or contextually significant named entities.

For each entity, determine the most appropriate and specific type label. Go beyond general categories when possible (e.g., use *Quantum Theory* instead of just *Scientific Concept*).

```
</instruction>

<example>
```

The output should be a list of tuples in the following format:

```
[("entity 1", "type of entity 1"), . . . ]
</example>

<input>
Passage: {input_passage}
</input>

Annotations:
```

Figure 9: The prompt used to extract named entities and its types from the filtered FineWeb-2 passages.

Prompt template for LLM based quality assessment of NER annotations

```
<instruction>
```

You are an expert in multilingual named entity recognition. You will be given a sentence in `{language}` and a list of entity types. Your task is to evaluate the quality of the annotations along the following dimensions:

- *faithfulness*: how well the entity types match the entity mentions;
- *completeness*: whether all important entities present in the sentence are annotated;
- *missing_annotations* (if any): a list of tuples of missing annotations;
- *extra_annotations* (if any): a list of tuples of extra annotations;
- *wrong_annotations* (if any): a list of tuples of wrong annotations.

The annotations are provided in the following format: `[ { "entity mention": <entity mention>, "entity type": <entity type> }, . . . ]`
Please answer strictly in JSON with the format:

```
{
"explanation": "<a short summary of the quality of the annotations with max. 100 words>",
"faithfulness": <a score between 0 and 5>,
"completeness": <a score between 0 and 5>,
"missing_annotations": <a list of tuples of missing annotations>,
"extra_annotations": <a list of tuples of extra annotations>,
"wrong_annotations": <a list of tuples of wrong annotations>
}
</instruction>

<input>
{test_input}
</input>

<annotations>
{annotations}
</annotations>

Quality assessment:
```

Figure 10: Prompt used for automatic quality assessment of NER annotations. The model reads a sentence and a set of proposed annotations and returns JSON with scores for faithfulness and completeness, together with lists of missing, extra, and wrong annotations.

Table 6: Full overview of FINERWEB.

ISO639	Language	Script	# Sent.	∅Length	∅Label/Sent.	# Unique Labels
					4o | Gemma | Merged	4o | Gemma | Merged
afr	Afrikaans	Latn	2495.0	1180.0	19.9 | 23.0 | 26.3	1420.0 | 4200.0 | 7213.0
als	Tosk Albanian	Latn	2498.0	1039.7	14.0 | 16.5 | 19.7	758.0 | 2403.0 | 3971.0
amh	Amharic	Ethi	2478.0	1286.0	34.8 | 30.7 | 49.0	1490.0 | 3816.0 | 6626.0
arb	Standard Arabic	Arab	2498.0	1114.3	20.4 | 24.0 | 29.3	1869.0 | 4600.0 | 8422.0
asm	Assamese	Beng	2495.0	1592.8	27.3 | 31.7 | 41.5	1082.0 | 3805.0 | 6175.0
azj	North Azerbaijani	Latn	2494.0	1189.1	19.2 | 21.9 | 26.5	897.0 | 3251.0 | 5272.0
bel	Belarusian	Cyrl	2475.0	1233.0	12.7 | 16.9 | 20.0	940.0 | 2827.0 | 4661.0
ben	Bengali	Beng	2500.0	1544.5	29.3 | 33.2 | 41.5	1142.0 | 3648.0 | 6131.0
bos	Bosnian	Latn	2491.0	1154.0	13.4 | 17.5 | 20.4	943.0 | 3280.0 | 5237.0
bre	Breton	Latn	2496.0	969.4	19.5 | 23.3 | 26.1	701.0 | 3875.0 | 6152.0
bul	Bulgarian	Cyrl	2485.0	1187.9	17.9 | 20.0 | 23.6	1674.0 | 3734.0 | 6816.0
cat	Catalan	Latn	2499.0	1081.8	15.6 | 17.1 | 20.2	1209.0 | 3553.0 | 6014.0
ces	Czech	Latn	2497.0	1215.5	14.9 | 17.4 | 20.4	1492.0 | 4132.0 | 6857.0
ckb	Central Kurdish	Arab	2493.0	1146.9	23.1 | 26.0 | 36.2	924.0 | 3842.0 | 6068.0
cmn	Mandarin Chinese	Hani	2495.0	1496.5	54.0 | 70.9 | 83.8	2488.0 | 8719.0 | 14572.0
cym	Welsh	Latn	2499.0	1073.2	16.6 | 18.4 | 21.9	1321.0 | 3851.0 | 6485.0
dan	Danish	Latn	2498.0	1171.0	18.7 | 20.7 | 23.8	1378.0 | 3998.0 | 6691.0
deu	German	Latn	2500.0	1335.8	19.8 | 22.2 | 26.1	1764.0 | 4885.0 | 8403.0
ekk	Standard Estonian	Latn	2491.0	1317.5	19.3 | 23.5 | 28.0	1324.0 | 3995.0 | 6623.0
ell	Greek	Grek	2496.0	1176.0	13.6 | 15.6 | 18.5	1215.0 | 3662.0 | 5963.0
eng	English	Latn	2498.0	1241.2	20.4 | 22.3 | 26.4	2034.0 | 4096.0 | 8002.0
epo	Esperanto	Latn	2497.0	1126.7	20.4 | 24.8 | 27.6	1227.0 | 5006.0 | 7853.0
eus	Basque	Latn	2500.0	1357.4	18.4 | 21.9 | 27.7	1004.0 | 4014.0 | 6381.0
fas	Persian	Arab	2498.0	1012.8	21.0 | 23.2 | 29.3	1640.0 | 4538.0 | 7859.0
fil	Filipino	Latn	2498.0	1135.4	17.2 | 19.5 | 22.4	1140.0 | 4186.0 | 6706.0
fin	Finnish	Latn	2498.0	1433.5	16.8 | 21.3 | 24.6	1108.0 | 4328.0 | 6674.0
fra	French	Latn	2500.0	1159.1	17.2 | 19.3 | 22.3	1534.0 | 4113.0 | 7080.0
fry	Western Frisian	Latn	2497.0	892.9	16.4 | 19.4 | 21.4	933.0 | 3103.0 | 5252.0
gaz	West Central Oromo	Latn	2500.0	1254.1	17.5 | 20.6 | 25.7	647.0 | 2553.0 | 4053.0
gla	Scottish Gaelic	Latn	2497.0	1132.1	16.9 | 19.4 | 22.4	1187.0 | 4132.0 | 6747.0
gle	Irish	Latn	2496.0	1152.5	18.0 | 20.0 | 24.0	1271.0 | 4953.0 | 7609.0
glg	Galician	Latn	2497.0	1133.3	17.4 | 19.0 | 22.3	1116.0 | 3588.0 | 6093.0
guj	Gujarati	Gujr	2489.0	1200.1	21.8 | 26.9 | 32.8	1264.0 | 3550.0 | 6191.0
heb	Hebrew	Hebr	1725.0	1083.6	21.5 | 21.9 | 28.6	1705.0 | 3816.0 | 6654.0
hin	Hindi	Deva	2498.0	1487.5	26.7 | 30.8 | 36.4	1155.0 | 3812.0 | 6495.0
hrv	Croatian	Latn	2496.0	1223.5	14.4 | 17.7 | 21.1	1223.0 | 3850.0 | 6136.0
hun	Hungarian	Latn	2498.0	1321.8	17.9 | 22.1 | 26.2	1396.0 | 4449.0 | 7419.0
hye	Armenian	Armn	2487.0	1747.0	19.5 | 25.1 | 30.6	919.0 | 4857.0 | 6719.0
ind	Indonesian	Latn	2500.0	1373.8	22.6 | 25.7 | 30.4	1444.0 | 4676.0 | 7749.0
isl	Icelandic	Latn	2499.0	1157.2	16.8 | 19.9 | 23.8	1099.0 | 4088.0 | 6459.0
ita	Italian	Latn	2498.0	1178.0	16.6 | 18.2 | 21.1	1438.0 | 3769.0 | 6483.0
jav	Javanese	Latn	2498.0	1343.3	22.5 | 26.6 | 30.1	1417.0 | 5458.0 | 8602.0
jpn	Japanese	Jpan	2497.0	1581.1	43.5 | 56.8 | 66.1	3051.0 | 9830.0 | 16592.0
kan	Kannada	Knda	2493.0	1357.9	22.3 | 26.7 | 33.6	1147.0 | 3649.0 | 6168.0
kat	Georgian	Geor	2480.0	1283.0	14.1 | 18.4 | 22.2	942.0 | 3202.0 | 5149.0
kaz	Kazakh	Cyrl	2485.0	1355.6	19.8 | 23.7 | 29.0	1091.0 | 4400.0 | 6897.0
khk	Halh Mongolian	Cyrl	2488.0	1276.6	18.0 | 22.0 | 27.4	1118.0 | 4505.0 | 6909.0
khm	Khmer	Khmr	2489.0	1967.4	32.8 | 40.2 | 54.4	1229.0 | 6793.0 | 9498.0
kir	Kirghiz	Cyrl	2487.0	1184.9	18.5 | 21.6 | 26.7	838.0 | 3273.0 | 5178.0
kor	Korean	Hang	2498.0	806.2	29.6 | 36.5 | 43.0	2113.0 | 6803.0 | 11457.0
lao	Lao	Laoo	2488.0	1665.3	29.0 | 34.9 | 47.3	1048.0 | 7049.0 | 9587.0
lat	Latin	Latn	2498.0	1353.0	19.1 | 24.6 | 29.6	869.0 | 5563.0 | 8041.0
lit	Lithuanian	Latn	2499.0	1359.8	13.1 | 17.8 | 20.8	1255.0 | 3592.0 | 5858.0
lvs	Standard Latvian	Latn	2490.0	1307.6	15.3 | 19.6 | 23.4	1285.0 | 3766.0 | 6212.0
mal	Malayalam	Mlym	2493.0	1573.7	16.5 | 22.5 | 27.0	857.0 | 3457.0 | 5332.0
mar	Marathi	Deva	2500.0	1326.4	23.2 | 26.2 | 32.1	1160.0 | 3545.0 | 5992.0
mkd	Macedonian	Cyrl	2500.0	1092.3	16.5 | 18.0 | 21.3	1144.0 | 2830.0 | 5092.0
mya	Burmese	Mymr	2488.0	2962.5	28.9 | 50.3 | 59.5	1157.0 | 9267.0 | 11890.0
nld	Dutch	Latn	2499.0	1217.7	18.3 | 19.8 | 23.0	1564.0 | 4296.0 | 7316.0
npi	Nepali	Deva	2499.0	1633.7	30.0 | 35.8 | 44.3	964.0 | 3972.0 | 6533.0
ory	Odia	Orya	2496.0	1352.4	30.3 | 30.3 | 42.0	1117.0 | 4064.0 | 6534.0
pan	Panjabi	Guru	2497.0	1522.9	26.4 | 29.9 | 37.5	1210.0 | 4545.0 | 7188.0
pbt	Southern Pashto	Arab	2498.0	928.5	18.8 | 23.3 | 27.9	642.0 | 2704.0 | 4155.0

Continued on next page

ISO639	Language	Script	# Sent.	∅Length	∅Label/Sent.	# Unique Labels
plt	Plateau Malagasy	Latn	2496.0	1318.4	18.5 | 20.5 | 24.4	971.0 | 3040.0 | 5252.0
pol	Polish	Latn	2498.0	1264.5	13.3 | 16.7 | 19.4	1293.0 | 3890.0 | 6264.0
por	Portuguese	Latn	2498.0	1116.9	17.2 | 18.7 | 22.0	1426.0 | 3633.0 | 6629.0
ron	Romanian	Latn	2500.0	1153.6	16.6 | 20.1 | 22.7	1356.0 | 3684.0 | 6212.0
rus	Russian	Cyrl	2494.0	1255.5	13.8 | 17.4 | 20.2	1555.0 | 3962.0 | 6792.0
san	Sanskrit	Deva	2489.0	2038.1	30.4 | 38.2 | 51.9	1297.0 | 10777.0 | 13842.0
sin	Sinhala	Sinh	2490.0	1143.6	18.5 | 21.9 | 29.5	1269.0 | 4709.0 | 7318.0
slk	Slovak	Latn	2498.0	1229.4	13.7 | 17.3 | 19.9	1341.0 | 3820.0 | 6224.0
slv	Slovenian	Latn	2499.0	1186.1	14.0 | 17.9 | 20.6	1214.0 | 3523.0 | 5838.0
snd	Sindhi	Arab	2495.0	943.4	19.6 | 23.0 | 28.0	968.0 | 3502.0 | 5780.0
som	Somali	Latn	2500.0	1156.0	17.2 | 21.9 | 24.7	627.0 | 2737.0 | 4414.0
spa	Spanish	Latn	2499.0	1128.0	16.1 | 17.8 | 20.5	1251.0 | 3400.0 | 5972.0
srp	Serbian	Latn	2498.0	1088.6	10.6 | 12.9 | 15.8	740.0 | 2196.0 | 3649.0
sun	Sundanese	Latn	2494.0	1330.9	21.3 | 25.5 | 29.4	1424.0 | 4839.0 | 7856.0
swe	Swedish	Latn	2500.0	1217.9	18.2 | 19.9 | 23.3	1426.0 | 4173.0 | 6976.0
swh	Swahili	Latn	2500.0	1182.0	19.2 | 22.2 | 25.9	948.0 | 3221.0 | 5271.0
tam	Tamil	Taml	2496.0	1457.8	19.7 | 22.5 | 27.7	1039.0 | 3026.0 | 5112.0
tel	Telugu	Telu	2497.0	1419.9	22.2 | 27.5 | 33.2	1262.0 | 3694.0 | 6242.0
tha	Thai	Thai	2496.0	2376.7	32.3 | 40.7 | 49.0	1899.0 | 8333.0 | 12727.0
tur	Turkish	Latn	2499.0	1312.7	21.0 | 24.1 | 27.5	1438.0 | 4343.0 | 7135.0
uig	Uighur	Arab	2488.0	1531.4	26.0 | 30.6 | 41.1	1023.0 | 6260.0 | 8494.0
ukr	Ukrainian	Cyrl	2482.0	1203.8	12.9 | 16.0 | 18.8	1299.0 | 3190.0 | 5458.0
urd	Urdu	Arab	2493.0	1915.3	31.5 | 36.8 | 44.2	1216.0 | 5109.0 | 8022.0
uzn	Northern Uzbek	Latn	2492.0	1305.1	18.0 | 20.9 | 25.1	932.0 | 3274.0 | 5245.0
vie	Vietnamese	Latn	2499.0	943.4	17.1 | 19.5 | 22.7	1300.0 | 3785.0 | 6283.0
xho	Xhosa	Latn	2498.0	1634.7	19.7 | 23.6 | 27.2	917.0 | 3711.0 | 5946.0
ydd	Eastern Yiddish	Hebr	2217.0	1249.8	17.4 | 21.9 | 27.0	1080.0 | 4664.0 | 7094.0
zsm	Standard Malay	Latn	2500.0	1370.4	19.6 | 23.0 | 26.5	1232.0 | 4030.0 | 6616.0