Representation Learning Enables Language Identification in *Zero*- and *Few*-Shot Settings

Rocco Tripodi
Department of Environmental Sciences, Informatics and Statistics
Ca' Foscari University of Venice
Via Torino 155, 30170 Venice, Italy
rocco.tripodi@unive.it

Abstract

Language Identification (LID) is a key component of multilingual NLP pipelines and multilingual corpus construction. LID errors can create spurious text–language associations, harming downstream model quality, especially for low-resource languages and fine-grained taxonomies involving closely related languages and dialects. We propose LIBE, a representation-based approach that embeds input texts and natural-language descriptions into a shared semantic space with a bi-encoder, enabling LID via similarity search. This formulation supports data-efficient learning and the addition of new languages through their descriptions. To improve discrimination in ambiguous or low-resource cases, we also introduce an optional cross-encoder re-ranking stage. Across multiple benchmarks, LIBE achieves competitive performance with strong LID systems when initialized from pretrained multilingual encoders, while requiring substantially less task-specific LID supervision in several adaptation settings. The hybrid bi-encoder–cross-encoder formulation further improves robustness in challenging open-set scenarios, achieving up to 90% F1 with as few as five labeled examples. Code and data are available at https://github.com/cafoscari-nlp/LIBE.

1 Introduction

Language identification (LID) is the task of automatically determining the language of a given piece of text (Jauhiainen et al., 2019). It has a long history dating back to the 1960s when LID was used for machine translation purposes (Jauhiainen et al., 2019). Nowadays, it is mainly utilized to develop multilingual corpora, scraped from the web, such as Common Crawl (Common Crawl, 2026), OSCAR (Abadji et al., 2022), and many others used to train multilingual language models.

Recent studies (Caswell et al., 2020; Kreutzer et al., 2022) have shown the limitations of this practice. Many current LID techniques require large numbers of examples of the same language to accurately identify it. This situation leads to a vicious cycle in which to obtain language data, the same kind of data is required. These issues in recent web-crawled multilingual datasets have been analyzed by Kreutzer et al. (2022), whose audit of 205 languages found that at least 15 lack usable text and that many others contain no more than 50% sentences of acceptable quality. While this disproportionately affects low-resource languages through misidentification and underrepresentation, it also impacts high-resource languages: incorrect language labeling can lead to misleading claims of cross-lingual transfer that are in fact driven by the unnoticed presence of low-resource languages in ostensibly monolingual corpora (Blevins and Zettlemoyer, 2022).

To address these challenges, we present LIBE, a model based on a Bi-Encoder (Mazaré et al., 2018; Reimers and Gurevych, 2019) and an optional Cross-Encoder (Wu et al., 2020; Thakur et al., 2021) that adapts pretrained multilingual representations to LID using substantially less task-specific supervision in low-resource and extensible settings. The Bi-Encoder is trained to represent a language description and texts written in that language as close points in a vector space. This representation allows us to formulate LID as a semantic search task in which the sentences to be labeled are embedded and used as queries over the vector space to retrieve the most similar language description from it (see Figure 1). The cross-encoder is trained to re-rank the candidates retrieved by the bi-encoder, yielding more accurate predictions in ambiguous cases. The central tenet of our methodology stems from the recognition that linguistic descriptions are accessible for a significant number of low-resource languages,

Transactions of the Association for Computational Linguistics, vol. 14, pp. 2555–2573, 2026. https://doi.org/10.1162/TACL.a.815
Action Editor: Antoine Bosselut. Submission batch: 5/2026; Revision batch: 8/2026; Published 10/2026.
©2026 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

notwithstanding the limited availability of training data. Incorporating language information into training pipelines offers a viable strategy for addressing data scarcity, with the potential to improve classification by leveraging explicit information about languages as a learning signal.

LIBE also addresses several key *desiderata* of a modern LID model identified by Kargaran et al. (2023): i) broad language coverage, since candidate languages are represented through descriptions rather than only through learned classifier weights; ii) open-source availability, as both the model and associated resources are released; iii) ease of use, because inference primarily requires encoding the input text and performing similarity search over cached language-description embeddings; iv) uncertainty estimation, since similarity scores and the confidence threshold τ provide a natural mechanism for rejecting low-confidence predictions or routing them to the cross-encoder; v) practical efficiency, both in terms of limited task-specific supervision in adaptation settings and manageable inference cost when description embeddings are cached; vi) flexibility in language granularity, since the inventory can include languages, varieties, or script-specific labels whenever suitable descriptions are available; and vii) rapid inference, supported by pre-computing language-description embeddings and using efficient nearest-neighbor search. Among these properties, extensibility is central to LIBE: new candidate languages can be introduced through their descriptions without changing the classifier architecture, and few-shot examples can be used when additional adaptation is needed.

Experimental evaluations on three benchmarks demonstrate that: i) when initialized from pre-trained multilingual encoders, its performance is comparable to strong LID systems while using fewer task-specific LID examples; ii) it exhibits strong performance in *0*-shot and *few*-shot settings across both novel domains and previously unseen languages; iii) in particular, LIBE performs especially well in settings where the test languages are not known in advance and classification must therefore be carried out over all languages seen during training, in contrast to many existing approaches whose performance degrades under these conditions; iv) moreover, LIBE gains 12 F1 points when a cross-encoder is used to re-rank and refine the base model's predictions.

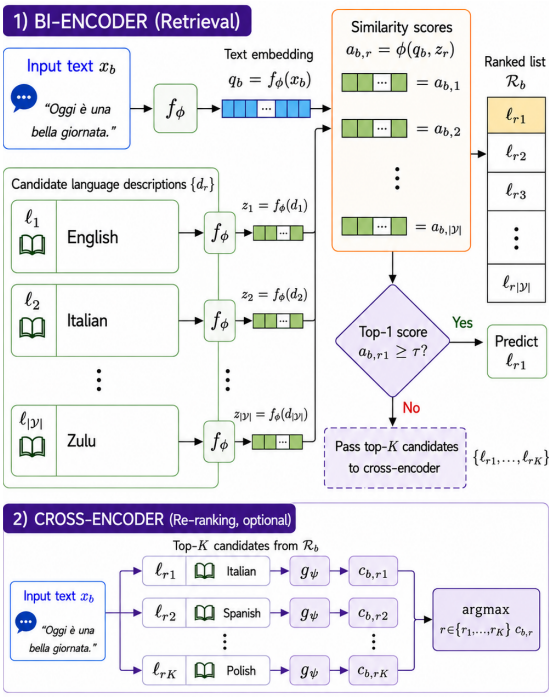

Figure 1: Overview of the proposed two-stage language identification pipeline. A bi-encoder retrieves and ranks candidate languages by comparing text and language-description embeddings; when the top prediction is below the confidence threshold τ, the top-K candidates are re-ranked by an optional cross-encoder.

2 Related Work

LID has evolved significantly over time. Early approaches relied on rule-based methods, characterized by heuristic rules (Lins and Gonçalves, 2004) and character frequency analysis (Nakamura, 1971). These methods were later superseded by statistical techniques, which gained popularity for extracting n-gram patterns from training texts and identifying languages by matching these patterns to new texts (Gyawali et al., 2013; Brown, 2013). Techniques such as Naïve Bayes classifiers also became widely used, leveraging probabilities derived from character or word frequencies (Lui and Baldwin, 2011; Peng and Schuurmans, 2003). More recently, deep learning techniques have revolutionized LID. Among these, FastText (Bojanowski et al., 2017) has emerged as a particularly effective approach due to its efficiency and accuracy. FastText has significantly advanced the field, especially for applications requiring real-time processing and robust performance across diverse languages. Its architecture extends traditional word embedding methods, such as word2vec (Mikolov et al., 2013), by incorporat-

ing subword information. This provides a crucial advantage for handling morphologically rich languages and rare words, as it enables the model to capture language-specific patterns at the subword level. Additionally, it is optimized for speed, allowing it to process large datasets efficiently and perform real-time language identification.

We argue that the dominance of FastText in recent state-of-the-art LID research stems from the scarcity of standardized benchmarks for the task. Notably, widely used LID models, such as CLD3 (Caswell et al., 2020), are trained on proprietary datasets, limiting the development of alternative methodologies. Moreover, the absence of standard evaluation benchmarks has hindered both the development and rigorous assessment of new models. However, we believe this *impasse* is coming to an end. The recent release of large, curated training sets such as LID-201 (Burchell et al., 2023) and GlotLID (Kargaran et al., 2023), along with new test sets like Flores-200 (NLLB Team et al., 2022) and MCS-350 (Agarwal et al., 2023), is expected to drive significant progress in this field as also exemplified by specialized workshops such as WMDQS (Ortiz Suarez et al., 2025).

3 Task Formulation

Notation. Let $\mathcal{Y} = \{\ell_1, \ldots, \ell_N\}$ be the language inventory, where N is the number of candidate languages. We use r to index languages, so that $\ell_r \in \mathcal{Y}$ denotes the r-th language. Each language ℓ_r is associated with a textual description d_r. We use x for an input text and $y \in \mathcal{Y}$ for its gold language label. Each text is assumed to have exactly one gold language.

Language Identification. Given an input text x and a language inventory $\mathcal{Y}$, a LID model assigns a score $a_r(x)$ to each candidate language ℓ_r. LID is formulated here as a single-label classification task: for each input text, the model predicts exactly one language label,

$$\hat{\ell}(x) = \arg\max_{\ell_r \in \mathcal{Y}} a_r(x),$$

corresponding to the highest-scoring candidate.

0-shot LID. We also consider zero-shot LID. Let $\mathcal{Y}_{\text{train}}$ and $\mathcal{Y}_{\text{test}}$ denote the language inventories used during training and testing, respectively. In the zero-shot setting, some test languages are not observed during training, i.e., $\mathcal{Y}_{\text{test}} \setminus \mathcal{Y}_{\text{train}} \neq \emptyset$. At inference time, descriptions are assumed to be available for the candidate languages in $\mathcal{Y}_{\text{test}}$.

4 Methodology

Our approach uses a bi-encoder to map input texts and language descriptions into a shared embedding space. The bi-encoder retrieves a ranked list of candidate languages for each input text. In low-resource or low-confidence settings, an optional cross-encoder re-ranks the top-K candidates.

Bi-encoder. Let f_θ denote the encoder, composed of a Transformer backbone followed by a pooling operation. The same encoder is applied to both input texts and language descriptions, following a siamese architecture (Schroff et al., 2015).

Given a training example (x_b, y_b), let r_b be the index of the gold language, so that $y_b = \ell_{r_b}$. The input text and its corresponding language description are encoded as

$$q_b = f_\theta(x_b), \qquad z_{r_b} = f_\theta(d_{r_b}),$$

where q_b is the text embedding and z_{r_b} is the embedding of the gold language description. More generally, for any candidate language ℓ_r, we write $z_r = f_\theta(d_r)$. The bi-encoder score for text x_b and language ℓ_r is $a_{b,r} = \phi(q_b, z_r)$, where ϕ is a similarity function. Training encourages the score of the correct language, a_{b,r_b}, to be higher than the scores of incorrect candidate languages.

Because the model relies on a Transformer architecture, its computational complexity scales quadratically with the input length. In this work, we do not use optimized variants such as sparse attention (Beltagy et al., 2020), fast attention (Dao et al., 2022), low-rank approximations (Xu et al., 2024), or quantization (Xiao et al., 2023a). We did not experience time limitations when training our largest model on a single 48GB GPU. The runtime results reported in Section 6 can therefore be considered an upper bound that could be reduced with further optimization. In this work, we instead focus on low-resource training scenarios.

Inference. At inference time, the descriptions of all candidate languages are first encoded and cached as $z_r = f_\theta(d_r)$ for all $\ell_r \in \mathcal{Y}$. For each input text x_b, the model computes its embedding $q_b = f_\theta(x_b)$ and scores it against every language description: $a_{b,r} = \phi(q_b, z_r)$. The bi-encoder returns the ranked list $\mathcal{R}_b = \text{rank}_{\ell_r \in \mathcal{Y}}\, a_{b,r}$. The top-ranked language is selected as prediction if its bi-encoder score is above a confidence threshold τ.

Cross-encoder. When the prediction is unreliable or when additional discrimination is needed, the top-K candidates in $\mathcal{R}_b$ are passed to a cross-encoder. For each candidate language ℓ_r, the cross-encoder jointly encodes the pair (x_b, d_r) and produces a score $c_{b,r} = g_\omega(x_b, d_r)$, where g_ω is the cross-encoder scoring function. The final prediction is the candidate with the highest cross-encoder score among all candidates.

4.1 Implementation Details

Language selection and descriptions. We used the Glottolog 5.3 taxonomy (Nordhoff and Hammarström, 2011) to construct the language inventory $\mathcal{Y}$. We discarded non-natural languages by filtering out entries annotated with the following labels instead of a proper language family: *Pidgin*, *Unattested*, *Unclassifiable*, *Artificial Language*, *Sign Language*, *Speech Register*, *Mixed Language*, and *Bookkeeping*. This resulted in $N = 6{,}826$ languages.

For each language $\ell_r \in \mathcal{Y}$, we collected a textual description d_r from Wikipedia when available. For the 73 languages not covered by Wikipedia, we used Ethnologue.[1] We extract the lead textual content of the corresponding page, excluding tables, navigation elements, references, and infobox metadata. These descriptions contain typological, genealogical, geographic, and sociolinguistic information, all of which may provide useful signals for distinguishing related languages.[2]

We considered two description variants. The first uses the beginning of the description up to the encoder input limit. The second applies extractive compression under a fixed token budget. In the latter case, sentences are encoded with the BGE model (Xiao et al., 2023b), ranked by similarity to the full description, and selected until the token budget is reached. Descriptions are tokenized with the tokenizer of the corresponding encoder. Their length is highly skewed: the median description has 151 tokens, while the mean is 525 tokens; 31% of descriptions exceed the encoder maximum length of 512 tokens (using BERT's tokenizer). This design allows us to test whether naturally occurring descriptions provide useful semantic anchors for LID, rather than relying on other manually curated resources.

[1] https://www.ethnologue.com accessed 9/7/26.

[2] Representative examples of the language descriptions used by LIBE are provided in Appendix A.

Transformer backbone T_θ. We tested both randomly initialized Transformers and pretrained models, including monolingual and multilingual models. The dimensionality h of the text embeddings q_b and language-description embeddings z_r is inherited from the Transformer backbone T_θ.

Bi-encoder loss. The bi-encoder is trained using the Multiple Negatives Ranking Loss (Henderson et al., 2017). Consider a training mini-batch

$$\mathcal{B}_{\text{tr}} = \{(x_b, \ell_{r_b}, d_{r_b}) \mid b = 1, \ldots, B_{\text{tr}}\},$$

where B_{tr} is the training batch size. Each input text x_b is paired with its positive language description d_{r_b}. The other language descriptions in the same batch, $\{d_{r_c} \mid c \neq b\}$, are used as in-batch negatives. To ensure that these negatives are valid, batches are constructed so that no two examples in the same batch have the same gold language index, i.e., $r_b \neq r_c$ for $b \neq c$.

The loss function is computed as:

$$\mathcal{J}_{\text{bi}} = -\frac{1}{B_{\text{tr}}} \sum_{b=1}^{B_{\text{tr}}} \log \frac{\exp\left(\phi(q_b, z_{r_b})/\rho\right)}{\sum_{c=1}^{B_{\text{tr}}} \exp\left(\phi(q_b, z_{r_c})/\rho\right)},$$

where $\rho > 0$ is a temperature parameter. This objective encourages each input text embedding q_b to be closer to its corresponding language-description embedding z_{r_b} than to the other description embeddings in the same batch.

Cross-encoder loss The cross-encoder is trained as a re-ranking model to refine the predictions of the bi-encoder. For a training text x_b with gold language ℓ_{r_b}, the positive pair is (x_b, d_{r_b}). Hard negative descriptions are selected from the top-K bi-encoder candidates. Let $\mathcal{N}_b \subseteq \mathcal{C}_b \setminus \{r_b\}$ denote the set of hard negative language indices for x_b, where $\mathcal{C}_b$ is the set of top-K candidate indices returned by the bi-encoder.

The cross-encoder scoring function $g_\omega(x, d)$ jointly encodes the input text and a candidate language description. The model is trained with binary cross-entropy:

$$\mathcal{J}_{\text{ce}} = -\frac{1}{B_{\text{ce}}} \sum_{b=1}^{B_{\text{ce}}} \log \sigma(g_\omega(x_b, d_{r_b})) + \frac{1}{|\mathcal{N}_b|} \sum_{s \in \mathcal{N}_b} \log\left(1 - \sigma(g_\omega(x_b, d_s))\right),$$

where B_{ce} is the cross-encoder training batch size and σ is the sigmoid activation. This loss encourages the cross-encoder to assign a higher relevance

score to the correct language description than to hard negative ones.

Training regimes. We trained LIBE by varying both the number of positive text–description training pairs and the training dataset, using either LID-201 (Burchell et al., 2023) or GlotLID (Kargaran et al., 2023) (see Section 5.1). All models were trained for 5 epochs, with the learning rate linearly warmed up to a maximum value of 5×10^{-5}. The weight decay parameter was set to 0.01, and the temperature parameter was fixed at $\rho = 0.05$. The resulting models were then fine-tuned for 10 epochs in the *few*-shot experiments using only $k = \{1, 3, 5\}$ examples for each language not observed during training.

The training data for the cross-encoder (LIBE$_\chi$) was drawn selecting 10K sentences from GlotLID using temperature-based sampling with $\alpha = 0.5$. This sampling strategy smooths the original language distribution by reducing the probability of selecting examples from high-resource languages while increasing the probability of selecting examples from low-resource languages, thereby producing a more balanced multilingual training set.

For each language $\ell \in \mathcal{Y}$ with count $c_\ell > 0$, a sampling weight is computed as $w_\ell = c_\ell^{1-\alpha}$. The weights are then normalized to obtain the target number of samples for each language:

$$t_\ell = N \cdot \frac{w_\ell}{\sum_{\ell' \in \mathcal{Y}} w_{\ell'}},$$

where N is the total number of samples to draw and t_ℓ is the target allocation for language ℓ.

5 Experimental Setup

5.1 Training Datasets

LID-201 (Burchell et al., 2023) is a curated and open dataset comprising 121 million sentences in 201 languages. Its large size makes it suitable for training, and its openness allows researchers to conduct realistic experiments. In contrast, popular LID models are often trained on proprietary data, making comparison difficult. The dataset's sources include news sites, Wikipedia, and religious texts, resulting in mainly formal language.

GlotLID-C (Kargaran et al., 2023) comprises 289 million sentences spanning 1,941 languages (v3.1); however, only a subset of the corpus is publicly released. Of these, 145M sentences were designated for training purposes. For our experiments, we sampled sentences from the available data (∼60M), resulting in a dataset with 1940 script-language pairs covering 1,871 languages.

5.2 Test and Validation Datasets

Flores-200 (NLLB Team et al., 2022) consists of 842 articles sourced from English Wikimedia. Each sentence in them has been translated into 204 unique language-script combinations, representing 196 distinct languages, and has undergone human verification. The dataset includes 997 sentences for development, 1,012 for dev-test, and 992 for testing (not available). The Flores-200 test set is not publicly accessible. Consistent with previous research, we use the dev-test set as our Flores-200 test set and the dev split to tune the τ parameter of the models developed by us.

UDHR (Vatanen et al., 2010) comprises over 500 translations of the *Universal Declaration of Human Rights*. 419 of them have an ISO 639-3 code. In this paper we used the UDHR version curated by (Kargaran et al., 2023) that filtered out short sentences, that accounted for the 35% of sentences for each language.

MCS-350 (Agarwal et al., 2023) is a collection of 50,000 multilingual and parallel children's stories available in over 350 languages. This dataset was created using resources from the African Storybooks Initiative and Pratham Books' Storyweaver, both of which are available under Creative Commons licenses. It features both original and human-translated stories spanning various language families, including Niger-Congo, Indo-European, Nilo-Saharan, Sino-Tibetan, and Austronesian. About 70% of the languages in the dataset use the Latin script, though it also includes many stories in non-Latin scripts.

5.3 Language distributions

These datasets differ substantially in how training examples are distributed across languages. Although LID-201 and Flores-200 contain approximately the same languages, Flores-200 is balanced across languages, whereas LID-201 reflects a more realistic long-tail distribution with both high-resource and low-resource languages. UDHR exhibits a similarly skewed distribution, characterized by a sharp decline in the number of examples for low-resource languages. MCS-350 also represents a realistic setting in which high-resource and low-resource languages coexist.

5.4 Model Selection

We developed a fixed setup to isolate the effect of a small number of controlled parameters on the training of LIBE. Each configuration varies along two axes: (i) the encoder backbone, which determines model capacity and language coverage, and (ii) the type of positive supervision used during training. The first axis captures the impact of model size and multilingual pretraining, while the second evaluates how language descriptions can be leveraged within LIBE.

Encoder backbones. We evaluate a diverse set of text encoders covering different model sizes and pretraining strategies. *LaBSE* (Feng et al., 2022) is a large-scale multilingual encoder trained for language-agnostic sentence embeddings using contrastive learning. *EmbeddingGemma-300M* (Vera et al., 2025) is a recent multilingual embedding model derived from the Gemma family. *e5* models (Wang et al., 2024) (*e5-small* and *e5-large*) are based on XLM-R (Conneau et al., 2019) and optimized for retrieval-style embedding tasks. *MiniLM* (Wang et al., 2021) and *DistilUSE* (Yang et al., 2020) are lightweight multilingual encoders designed for efficiency, relying on knowledge distillation. *RoBERTa-large* (Liu et al., 2019) represents a high-capacity monolingual backbone. Finally, *BERTrand* is a randomly initialized BERT-base architecture that serves as a not pre-trained baseline. Together, these models span a wide range of parameter sizes and pretraining regimes, enabling analysis of the trade-offs between capacity, efficiency, and cross-lingual generalization.

Type of supervision. We consider different forms of supervision: *Full*, where complete language descriptions are used up to the model maximum sequence length (512 tokens); *Sum*, where descriptions are summarized to fit the same limit; and Sum_{50}, where descriptions are summarized to 50 tokens, allowing a lower training cost. We also include a *Shuffled* condition, where descriptions are randomly reassigned to languages, to test whether the model relies on semantic content rather than spurious associations. Finally, we evaluate a *Clf* setting, where a classification head is trained on top of the encoder, to assess the effect of the architecture choice in LIBE. All other parameters are kept fixed.

Evaluation uses macro precision (P), recall (R), F1, and false positive rate (FPR). We also report performance separately for seen (S) and unseen (U) languages, defined by pretraining coverage. For each subset, we report its size (n_S, n_U) and F1 scores ($F1_S$, $F1_U$).

Metrics. To assess model performance, we use precision (P), recall (R), and F1. To avoid down-weighting low-resource languages (Kreutzer et al., 2022), we report macro-averages. Following prior LID evaluations (Caswell et al., 2020), we define FPR as the number of predictions assigned to an incorrect language divided by the total number of test items. This dataset-level FPR differs from a per-class binary false-positive rate, but is useful for comparing how often systems produce spurious language assignments in large-scale filtering scenarios.

Model	Desc	θ	F1	FPR	n_S	$F1_S$	n_U	$F1_U$
BERTrand	Full	110	0.744	0.0013	1	0.870	195	0.744
BERTrand	Label	110	0.006	0.0051	1	0.014	195	0.006
BERTrand	Sum	110	0.777	0.0011	1	0.943	195	0.776
BERTrand	Sum_{50}	110	0.560	0.0021	1	0.449	195	0.560
RoBERTa	Full	355	0.836	0.0008	1	0.974	195	0.836
RoBERTa	Sum	355	0.832	0.0009	1	0.977	195	0.832
RoBERTa	Sum_{50}	355	0.813	0.0009	1	0.958	195	0.813
DistilUSE	Full	135	0.829	0.0009	12	0.941	184	0.822
DistilUSE	Sum	135	0.828	0.0009	12	0.936	184	0.820
DistilUSE	Sum_{50}	135	0.776	0.0011	12	0.897	184	0.768
MiniLM	Full	118	0.846	0.0008	41	0.943	155	0.821
MiniLM	Sum	118	0.845	0.0008	41	0.943	155	0.819
MiniLM	Sum_{50}	118	0.773	0.0011	41	0.876	155	0.745
e5-large	Full	560	0.885	**0.0006**	75	0.980	121	0.826
e5-large	Sum	560	**0.886**	**0.0006**	75	0.982	121	0.827
e5-large	Sum_{50}	560	0.879	**0.0006**	75	0.978	121	0.817
e5-small	Full	270	0.871	**0.0006**	75	0.973	121	0.808
e5-small	Sum	270	0.870	0.0007	75	0.974	121	0.806
e5-small	Sum_{50}	270	0.832	0.0008	75	0.966	121	0.749
Gemma	Full	300	0.807	0.0010	89	0.925	107	0.709
Gemma	Sum	300	0.811	0.0010	89	0.919	107	0.720
Gemma	Sum_{50}	300	0.784	0.0011	89	0.916	107	0.675
LaBSE	Full	470	**0.886**	**0.0006**	89	0.980	107	0.809
LaBSE	Sum	470	0.883	**0.0006**	89	0.976	107	0.805
LaBSE	Sum_{50}	470	0.872	**0.0006**	89	0.978	107	0.783
$LaBSE_{Clf}$	-	470	0.864	0.0007	89	0.970	107	0.776
LaBSE	Label	470	0.185	0.0039	89	0.272	107	0.112
LaBSE	Shuff	470	0.056	0.0048	89	0.060	107	0.054
$LaBSE_{1M}$	Full	470	0.899	0.0005	89	0.979	107	0.833
$LaBSE_{10M}$	Full	470	0.906	0.0005	89	0.981	107	0.844
$LaBSE_{1Mg}$	Full	470	0.890	0.0005	89	0.970	107	0.824
$LaBSE_{10Mg}$	Full	470	0.901	0.0005	89	0.971	107	0.843

Table 1: Model comparison on Flores-200 dev. in three blocks: backbone selection with 100K examples, LaBSE ablations, and scaling on LID-201/GlotLID-C. Params are in millions; metrics are macro F1 and FPR. S/U denote seen/unseen languages.

Results. Table 1 is divided into three blocks. The first block reports a controlled model-selection experiment in which all encoder backbones are trained under the same conditions, using the same 100K training examples and the same description-based retrieval formulation. This block isolates the effect of the encoder archi-

tecture and pretraining regime. Under this controlled setting, LaBSE and e5-large obtain the strongest overall results, with F1 scores around 0.885–0.886, while smaller multilingual encoders such as MiniLM and DistilUSE remain competitive but consistently underperform the larger models. BERTrand and RoBERTa also perform reasonably well, but show weaker generalization to languages not covered by their pretraining.

The second block focuses on LaBSE and tests alternative ways of using the same backbone. This block includes the standard description-based bi-encoder, a classifier baseline, a label-only retrieval condition, and a shuffled-description control. The classifier baseline, LaBSE$_{Clf}$, replaces the retrieval formulation with a conventional classification head trained to predict language labels directly from text embeddings. This comparison is important because it isolates whether the gains come simply from using a strong LaBSE encoder or from the retrieval-based formulation itself. LaBSE$_{Clf}$ performs competitively, confirming that LaBSE provides a strong multilingual representation space, but it remains below the description-based bi-encoder. This suggests that the retrieval formulation provides an advantage over a standard closed-set classifier, especially because it represents candidate languages through textual descriptions rather than through a fixed classifier. The label-only and shuffled-description controls further clarify the role of language descriptions. In the label-only condition, the model receives language names instead of full descriptions; in the shuffled condition, descriptions are randomly reassigned to languages. Both settings perform substantially worse than the description-based bi-encoder. This shows that the model does not merely exploit the presence of language identifiers, nor does it benefit from arbitrary description uniqueness. Instead, the alignment between each language and its actual description provides a meaningful semantic signal.

The third block evaluates the effect of increasing the amount of task-specific training data and changing the training corpus. Starting from the selected LaBSE configuration, we train models with 1M and 10M examples from LID-201, and with 1M and 10M examples from GlotLID-C. The LID-201 experiments test the effect of scaling the number of examples while keeping the language inventory close to that of Flores-200. The GlotLID-C experiments instead use a much broader training inventory, covering many more languages than those present in Flores-200 development data. Increasing the number of examples improves performance, with LaBSE reaching 0.899 F1 with 1M LID-201 examples and 0.906 F1 with 10M examples. Training on GlotLID-C remains competitive despite the much larger candidate-language space, suggesting that exposure to a broader set of languages improves robustness, although the reduced number of examples per language makes the comparison different from the LID-201 setting.

Overall, the results show that model capacity, multilingual pretraining, description-based supervision, and the size and diversity of the training inventory all contribute to performance. The comparison with LaBSE$_{Clf}$ indicates that a strong pretrained encoder alone is not sufficient to explain the best results: the retrieval-based formulation offers additional benefits over a conventional classifier. The seen/unseen split further shows that languages covered by pretraining are easier to identify, but the strong results on unseen languages indicate that pretrained multilingual encoders also provide transferable representations useful for languages not directly observed during pretraining.

5.5 Confidence Threshold

We use a confidence threshold τ to determine whether a bi-encoder prediction is reliable. The threshold is tuned on the Flores-200 development set (NLLB Team et al., 2022) by selecting the value that maximizes F1. In the bi-encoder-only setting, predictions below τ are assigned a __REJECT__ label; in the bi-encoder–cross-encoder setting, they are passed to the cross-encoder for reranking. Threshold sensitivity experiments with LaBSE models trained on 100K, 1M and 10M examples from LID-201 and GlotLID-C show that the optimal τ is model-dependent. LID-201 models remain stable for $\tau \leq 0.7$, whereas GlotLID-C models, especially GlotLID-10M, degrade more sharply at higher thresholds. We therefore tune τ separately for each model and split on Flores-200. For datasets without development sets, we report results with $\tau = 0.0$ and 0.5, following prior work.

6 Evaluation

The evaluation of LIBE was conducted on three benchmarks with different sizes and distribution of languages. This evaluation is aimed at inspect-

ing the behavior of the model in extreme situations, i.e., i) scarcity of training samples on all languages, and ii) scarcity of training samples only on some languages. These settings allows us to simulate cases in which extremely low-resource languages have to be added to a pretrained model and the effect that few samples have on the prediction of languages never seen by LIBE during training.

6.1 Models

CLD3. Google's Compact Language Detector v3 (CLD3) is based on the FastText (Bojanowski et al., 2017; Joulin et al., 2017; Grave et al., 2018) model. It employs a neural network model based on n-grams (where $1 \leq n \leq 3$). The inference process involves extracting character n-grams from the input text and calculating the frequency of each n-gram's occurrence. The model averages the embeddings of each n-gram type based on their frequencies, and these averaged embeddings are concatenated to form the embedding layer. The network's remaining components include a hidden Rectified Linear Unit (ReLU) layer and a softmax layer for multi-class classification. To predict the language of the input text, a forward pass through the network is performed.

NLLB. NLLB project was developed as part of the *No Language Left Behind* (NLLB Team et al., 2022) framework. It is a FastText model trained using publicly available datasets, with data cleaning performed to eliminate spurious correlations, and professional translators employed to create *bitext* data for low-resource languages. Threshold optimization is used to discard predictions with low confidence. Additionally, due to class imbalance in the training set, up-sampling was performed to increase the representativeness of low-resource languages. Sentences from a language l representing the proportion pl of the dataset are sampled proportionally to $pl^{0.3}$.

GlotLID-M. GlotLID-M (Kargaran et al., 2023) is a FastText model trained on GlotLID-C. The same techniques used to train NLLB were used also to train GlotLID-M.

OpenLID. OpenLID (Burchell et al., 2023) is a FastText model trained on the LID-201 dataset using, also in this case, the same parameters employed by (NLLB Team et al., 2022).

FastText. Our fastText models were trained on the same datasets used for LIBE tuning τ in the same way. The hyperparameters were those reported in (Burchell et al., 2023), i.e., $lr = 0.05$; $dim = 200$; $neg = 5$; and $ws = 5$.

6.2 Evaluation Methodologies

Consistent with (Kargaran et al., 2023), we use SET! for the oracle closed-set condition, where predictions are restricted to languages present in the test set, and SET? for the realistic open-set condition, where models may predict any language in their inventory. In the closed set scenario, to enable 0-shot prediction of languages not observed during training, we augment the candidate set at inference time with additional language representations, yielding the LIBE+0 configuration.

0- and few-shot Setting. Since the primary objective of this study is to assess the performance of LID models in real-world scenarios, we incorporate both 0-shot and few-shot learning into our evaluation. In the 0-shot setting, the evaluation set retains languages not seen during training. This enables us to evaluate the generalization capabilities of our approach. In the few-shot setting, we fine-tune several models using $[1, 3, 5]$ examples exclusively for the languages not covered during the initial training phase. This experiment provides insight into the adaptability to new languages.

6.3 Results

Flores-200. The Flores-200 benchmark largely overlaps with the language coverage of our models trained on LID-201, while differing in domain, as Flores-200 consists exclusively of formal-style texts. Results are reported in Table 2 for seven variants of LIBE, evaluated on three standard subsets corresponding to the language coverage of OpenLID, NLLB, and CLD3.

Under the closed-set evaluation, LIBE achieves competitive or superior performance across all subsets. On OpenLID, LIBE trained on GlotLID with 10M sentences attains the highest F1-score, outperforming FastText and matching or exceeding models trained on substantially larger corpora. On the NLLB subset, OpenLID performs best overall, but LIBE trained on 10M sentences achieves closely comparable results. On the CLD3 subset, LIBE consistently reaches near-perfect F1-scores across training sizes, including in low-data regimes. Under matched task-specific LID

					Flores-200 ∩ OpenLID 201 langs					Flores-200 ∩ NLLB 193 langs					Flores-200 ∩ CLD3 95 langs				
Model	**Train DS**	**|train|**	**|L|**	**SET**	τ	**F1↑**	**FPR↓**	**P↑**	**R↑**	τ	**F1↑**	**FPR↓**	**P↑**	**R↑**	τ	**F1↑**	**FPR↓**	**P↑**	**R↑**
CLD3	unk.	unk.	107	!	-	-	-	-	-	-	-	-	-	-	-	0.968	0.03	-	-
NLLB	unk.	unk.	218	!	-	-	-	-	-	-	0.95	0.023	-	-	-	0.985	0.019	-	-
OpenLID	OpenLID	121M	201	!	-	0.927	0.033	-	-	-	**0.959**	0.02	-	-	-	0.989*	0.011	-	-
FastText	OpenLID	1M	201	!	0.5	0.802	0.0008	0.872	0.814	0.45	0.814	0.0009	0.881	0.828	0.5	0.904	0.0007	0.934	0.910
FastText	OpenLID	10M	201	!	0.55	0.851	0.0006	0.892	0.855	0.55	0.870	0.0005	0.903	0.877	0.5	0.935	0.0005	0.942	0.936
FastText	GlotLID	10M	1871	!	0.45	0.752	0.0011	0.822	0.760	0.5	0.776	0.0009	0.846	0.780	0.5	0.859	0.0009	0.896	0.855
GlotLID-M	GlotLID-C	245M	2102	!	0.45	0.918	**0.0002**	0.926	0.918	0.5	0.934	**0.0001**	0.937	0.936	0.5	0.902	0.0009	0.922	0.911
LIBE	OpenLID	1K	201	!	0.1	0.685	0.0015	0.724	0.711	0.1	0.706	0.0015	0.745	0.731	0.1	0.972	0.0003	0.974	0.973
LIBE	OpenLID	10K	201	!	0.2	0.850	0.0007	0.857	0.863	0.2	0.874	0.0006	0.882	0.886	0.3	0.990	**0.0001**	0.990	0.990
LIBE	OpenLID	100K	201	!	0.25	0.895	0.0005	0.903	0.905	0.25	0.919	0.0004	0.927	0.928	0.4	0.990	**0.0001**	0.990	0.990
LIBE	OpenLID	1M	201	!	0.35	0.911	0.0004	0.925	0.919	0.35	0.931	0.0004	0.942	0.939	0.35	0.990	**0.0001**	0.990	0.990
LIBE	OpenLID	10M	201	!	0.45	0.918*	0.0004	0.931	0.926	0.4	0.937	0.0003	0.945	0.945	0.45	**0.991**	**0.0001**	**0.991**	**0.991**
LIBE	GlotLID	1M	1871	!	0.35	0.900	0.0005	0.913	0.909	0.35	0.925	0.0004	0.939	0.934	0.35	0.980	0.0002	0.984	0.984
LIBE	GlotLID	10M	1871	!	0.45	**0.928**	0.0003	**0.935**	**0.934**	0.4	0.949*	0.0002	**0.956**	**0.955**	0.45	0.981	0.0002	0.979	0.985
FastText	GlotLID	10M	1871	?	0.45	0.752	0.0011	0.822	0.760	0.5	0.778	0.0009	0.847	0.783	0.5	0.859	0.0009	0.896	0.855
GlotLID-M	GlotLID-C	245M	2102	?	0.45	0.917	**0.0002**	0.925	0.918	0.5	0.934	**0.0001**	0.937	0.936	0.5	0.956	0.0001	0.958	0.956
LIBE	OpenLID	10M	201	?	0.45	**0.918**	0.0004	**0.931**	**0.926**	0.4	**0.937**	0.0003	**0.945**	**0.945**	0.45	**0.991**	0.0001	**0.991**	**0.991**
LIBE	GlotLID	1M	1871	?	0.35	0.882	0.0004	0.888	0.888	0.35	0.907	0.0002	0.912	0.912	0.35	0.950	**0.0000**	0.954	0.950
LIBE	GlotLID	10M	1871	?	0.45	0.904*	0.0002	0.904	0.909	0.4	0.925*	**0.0001**	0.925	0.930	0.45	0.949*	**0.0000**	0.948	0.950

Table 2: Results on the Flores-200 dataset presented as F1, false positive rate (FPR), precision (P) and recall (R). The experiments were conducted on three subsets of languages: languages in Flores-200 and also in OpenLID, languages in Flores-200 and also in NLLB, and languages in Flores-200 and also in CLD3. The best result on each column is reported in bold and the first F1 result with a statistically significant difference from it is marked with * (χ^2, $p < 0.05$). For each model we reported also the training dataset, the number of training examples ($|train|$), the number of languages ($|L|$) and the evaluation strategy (SET).

training data, LIBE outperforms fastText[3]: with as few as 100K sentences, LIBE matches or surpasses FastText models trained on tens of millions of examples. Moreover, LIBE benefits from training on GlotLID, which spans over 1,800 languages, whereas FastText degrades in this setting due to reduced data per language.

We report results under a realistic open-set evaluation, where models may predict any language seen during training. Although performance drops as expected, LIBE remains more robust than FastText. Trained on OpenLID with 10M sentences, LIBE performs strongly across subsets; trained on GlotLID, it maintains competitive F1-scores even with thousands of candidate languages. These results show that LIBE scales to realistic LID scenarios with large, open candidate sets.

Trained on GlotLID-C with 245M sentences and over 2,100 languages, GlotLID-M achieves high F1-scores and extremely low FPR, showing the value of large-scale multilingual training for robust language discrimination.

UDHR. Table 3 reports precision, recall, F1, and false positive rate (FPR) as a function of the number (k) of fine-tuning examples for the 54 languages not seen during training and the decision threshold τ, under both the closed-set (SET=!) and realistic open-set (SET=?) evaluation settings. For completeness, we included also the results of GlotLID-M in 0-shot although it was trained on substantially more data. It was not possible to train this model with *few*-shot examples because part of its training dataset is not available.

In the closed-set scenario, GlotLID-M achieves strong performance with consistently low FPR, reflecting the benefits of large-scale multilingual training when the set of target languages is fixed. LIBE exhibits a consistent improvement as k increases: both precision and recall steadily improve, leading to higher F1-scores while maintaining low FPR. Increasing τ to 0.5 further improves precision and reduces false positives, yielding the best overall trade-off for LIBE.

The proposed hybrid LIBE$_\chi$ architecture variant achieves the highest F1-scores among all evaluated methods, showing that the cross-encoder can resolve ambiguous cases and improve both precision and recall with only a marginal increase in computational cost. Overall, these results indicate that combining bi-encoder efficiency with cross-encoder expressiveness yields a favorable accuracy–robustness trade-off.

Under open-set evaluation, performance degrades for all models, as expected, since predictions are no longer restricted to test-set languages. This drop is especially pronounced for fastText, whose F1-scores fall substantially across values of k, reflecting the difficulty of discriminating among

[3]LIBE uses pretrained representations, whereas fastText is trained from scratch; the comparison therefore concerns task-specific LID supervision.

Model	—train—	SET	τ	P ↑				R ↑				F1 ↑				FPR ↓			
				k=0	k=1	k=3	k=5	k=0	k=1	k=3	k=5	k=0	k=1	k=3	k=5	k=0	k=1	k=3	k=5
GlotLID-M	245M	!	0.0	0.701	–	–	–	0.772	–	–	–	0.715	–	–	–	0.0005	–	–	–
GlotLID-M	245M	!	0.5	0.759	–	–	–	0.769	–	–	–	0.756*	–	–	–	0.0002	–	–	–
LIBE	10M+k	!	0.0	0.750	0.786	0.805	0.814	0.778	0.799	0.813	0.819	0.737	0.764*	0.782*	0.790*	0.0005	0.0005	0.0005	0.0004
LIBE	10M+k*54	!	0.5	**0.787**	**0.816**	**0.826**	0.845	**0.783**	**0.805**	**0.821**	0.831	**0.768**	**0.790**	**0.804**	0.817	**0.0003**	**0.0003**	**0.0003**	**0.0003**
LIBE$_\chi$	10M+10K	!	0.0	-	-	-	0.832	-	-	-	0.836	-	-	-	0.809	-	-	-	0.0004
LIBE$_\chi$	10M+10K	!	0.5	-	-	-	**0.861**	-	-	-	**0.849**	-	-	-	**0.840**	-	-	-	**0.0003**
fastText	10M+k*54	!	0.0	0.578	0.591	0.591	0.587	0.532	0.543	0.559	0.552	0.495	0.513	0.524	0.509	0.0011	0.0011	0.0011	0.0011
fastText	10M+k*54	!	0.5	0.614	0.606	0.640	0.618	0.587	0.571	0.606	0.586	0.575	0.564	0.595	0.576	0.0005	0.0006	0.0005	0.0007
GlotLID-M	245M	?	0.0	0.749	–	–	–	0.765	–	–	–	0.747	–	–	–	0.0003	–	–	–
GlotLID-M	245M	?	0.5	0.759	–	–	–	0.769	–	–	–	0.756	–	–	–	0.0002	–	–	–
LIBE+0	10M	?	0.0	0.775	–	–	–	0.773	–	–	–	0.754	–	–	–	0.0003	–	–	–
LIBE+0	10M	?	0.5	**0.780**	–	–	–	**0.777**	–	–	–	**0.761**	–	–	–	0.0002	–	–	–
LIBE	10M+k*54	?	0.0	0.740	0.803	0.821	0.836	0.747	**0.796**	0.807	0.822	0.725	0.776*	0.790	0.807	0.0003	0.0003	0.0003	0.0003
LIBE	10M+k*54	?	0.5	0.748	**0.806**	**0.821**	0.843	0.749	**0.796**	**0.811**	0.827	0.732*	**0.782**	**0.796**	0.815*	**0.0002**	**0.0002**	**0.0002**	**0.0003**
LIBE$_\chi$	10M+10K	?	0.0	-	-	-	0.836	-	-	-	0.839	-	-	-	0.813	-	-	-	0.0004
LIBE$_\chi$	10M+10K	?	0.5	-	-	-	**0.860**	-	-	-	**0.848**	-	-	-	**0.836**	-	-	-	**0.0003**
fastText	10M+k*54	?	0.0	0.612	0.611	0.645	0.622	0.568	0.551	0.599	0.565	0.552	0.544	0.586	0.551	0.0007	0.0008	0.0006	0.0008
fastText	10M+k*54	?	0.5	0.614	0.606	0.640	0.618	0.587	0.571	0.606	0.586	0.575	0.564	0.595*	0.576	0.0005	0.0006	0.0005	0.0007

Table 3: Results on UDHR presented as precision (P), recall (R), F1 and FPR. For each model we report the number of training examples ($|train|$), the evaluation setting (SET; ! = closed set, ? = open set), and the number of examples k. The best result in each column is highlighted in bold, and the first F1 result with a statistically significant difference from it is marked with an asterisk * (χ^2, $p < 0.05$).

Model	—train—	SET	τ	P ↑				R ↑				F1 ↑				FPR ↓			
				k=0	k=1	k=3	k=5	k=0	k=1	k=3	k=5	k=0	k=1	k=3	k=5	k=0	k=1	k=3	k=5
GlotLID-M	245M	!	0.0	0.405	–	–	–	0.481	–	–	–	0.410	–	–	–	0.0010	–	–	–
GlotLID-M	245M	!	0.5	0.501	–	–	–	0.523	–	–	–	0.492*	–	–	–	**0.0006**	–	–	–
LIBE	10M+k*86	!	0.0	0.463	0.466	0.482	0.492	0.532	0.544	0.558	0.569	0.457	0.460*	0.476*	0.484*	0.0009	0.0009	0.0009	0.0008
LIBE	10M+k*86	!	0.5	**0.544**	**0.553**	**0.574**	0.573	**0.579**	**0.593**	**0.606**	0.621	**0.534**	**0.538**	**0.556**	0.560	0.0007	**0.0007**	**0.0007**	0.0006
LIBE$_\chi$	10M+10K	!	0.0	-	-	-	0.512	-	-	-	0.599	-	-	-	0.497	-	-	-	0.0008
LIBE$_\chi$	10M+10K	!	0.5	-	-	-	**0.645**	-	-	-	**0.670**	-	-	-	**0.609**	-	-	-	**0.0005**
fastText	10M+k*86	!	0.0	0.325	0.322	0.323	0.311	0.215	0.216	0.205	0.221	0.198	0.196	0.184	0.198	0.0021	0.0021	0.0022	0.0021
fastText	10M+k*86	!	0.5	0.404	0.403	0.393	0.408	0.355	0.333	0.340	0.363	0.338	0.307	0.313	0.332	0.0010	0.0010	0.0011	0.0010
GlotLID-M	245M	?	0.0	0.479	–	–	–	0.509	–	–	–	0.472	–	–	–	0.0007	–	–	–
GlotLID-M	245M	?	0.5	0.501	–	–	–	0.523	–	–	–	0.492	–	–	–	**0.0006**	–	–	–
LIBE+0	10M	?	0.0	0.521	–	–	–	0.554	–	–	–	0.503	–	–	–	0.0007	–	–	–
LIBE+0	10M	?	0.5	**0.550**	–	–	–	**0.573**	–	–	–	**0.531**	–	–	–	0.0006	–	–	–
LIBE	10M+k*86	?	0.0	0.461	0.523	0.538	0.550	0.527	0.560	0.580	0.591	0.464	0.502*	0.518*	0.531*	0.0007	0.0007	0.0007	0.0007
LIBE	10M+k*86	?	0.5	0.491	**0.561**	**0.575**	0.581	0.543	**0.584**	**0.593**	0.613	0.492*	**0.537**	**0.548**	0.561	**0.0006**	**0.0006**	**0.0006**	**0.0006**
LIBE$_\chi$	10M+10K	?	0.0	-	-	-	0.578	-	-	-	0.632	-	-	-	0.556	-	-	-	0.0007
LIBE$_\chi$	10M+10K	?	0.5	-	-	-	**0.622**	-	-	-	**0.674**	-	-	-	**0.596**	-	-	-	**0.0006**
fastText	10M+k*86	?	0.0	0.404	0.391	0.377	0.393	0.321	0.288	0.300	0.344	0.303	0.267	0.274	0.307	0.0014	0.0015	0.0015	0.0012
fastText	10M+k*86	?	0.5	0.404	0.403	0.393	0.408	0.355	0.333	0.340	0.363	0.338	0.307	0.313	0.332	0.0010	0.0010	0.0011	0.0010

Table 4: Results on MCS-350 presented as precision (P), recall (R), F1 and FPR. For each model we report the number of training examples ($|train|$), the evaluation setting (SET; ! = closed set, ? = open set), and the number of examples k. The best result in each column is highlighted in bold, and the first F1 result with a statistically significant difference from it is marked with an asterisk * (χ^2, $p < 0.05$).

many languages with n-gram representations. In contrast, LIBE remains more robust, benefiting from larger k values and higher thresholds. Notably, LIBE+0 enables 0-shot prediction by expanding the candidate set and maintains competitive performance despite the larger label space. It remains robust as the label space expands.

Across both evaluation settings, LIBE$_\chi$ consistently achieves the best overall F1 while preserving low FPR. This suggests that the cross-encoder mainly improves difficult cases without sacrificing robustness. Overall, these results show that retrieval-based language identification scales effectively with the number of candidate languages and input ambiguity, and that selectively incorporating cross-encoder scoring benefits both closed- and open-set conditions.

MCS-350. Table 4 reports results in a substantially more challenging setting, characterized by lower absolute performance across all models. In this regime, GlotLID-M shows limited effectiveness despite its large training corpus, indicating that scale alone is insufficient when label ambiguity increases. fastText exhibits the weakest performance overall, with consistently low F1-scores and higher FPR, and benefits only marginally from increasing k or adjusting the threshold.

In contrast, LIBE demonstrates clear gains as k increases for the 86 languages not seen during training, with both precision and recall improving steadily. This trend holds in both the closed-set and realistic evaluation settings, confirming the importance of candidate expansion in difficult scenarios. Raising the decision threshold further im-

proves robustness, leading to the best trade-offs between accuracy and false positives.

The hybrid model LIBE$_\chi$ consistently achieves the strongest results, confirming that selectively invoking the cross-encoder is especially beneficial in low-confidence and ambiguous cases. It maintains low FPR while delivering the highest F1-scores in both evaluation settings.

DS	SET = ?			SET = !		
	Bi	Cross	GlotLID-M	Bi	Cross	GlotLID-M
udhr-lid	0.884	**0.895**	0.806	0.840	0.920	**0.935**
limit	0.865	**0.912**	0.793	0.826	0.914	**0.940**

Table 5: Coverage of the models with $\tau = 0.5$.

6.4 Analysis

Coverage. Table 5 reports the fraction of inputs for which the model actually makes a prediction over the total number of samples. Across both datasets, the cross-encoder consistently attains higher coverage than the bi-encoder, indicating that it is able to confidently handle a larger fraction of inputs when applied selectively. This effect is particularly pronounced in the closed-set setting, where cross-encoder refinement substantially reduces abstentions. GlotLID-M exhibits the highest coverage in the realistic setting, reflecting its strong confidence when predicting over a large label space, albeit at the cost of lower precision as discussed previously. Notably, the cross-encoder achieves coverage values close to those of GlotLID-M while retaining the advantages of the bi-encoder-based filtering strategy.

Generalization to Unseen Languages. To further analyze the generalization capabilities of the models, we restrict the evaluation to languages that were not observed during training. In this setting, models are required to predict languages for which no direct supervision was available during pretraining, and, in the *few*-shot case, are supported only by a very small number of labeled examples. This scenario represents an extreme but realistic low-resource condition. Figure 2 summarizes the macro F1 scores obtained by LIBE and LIBE$_\chi$ on UDHR and MCS-350 when considering only unseen languages. In the 0-shot setting ($k = 0$), performance is predictably low. Nevertheless, the fact that LIBE is able to produce meaningful predictions in this setting is in itself remarkable, given that no examples of the target languages were observed during training. This

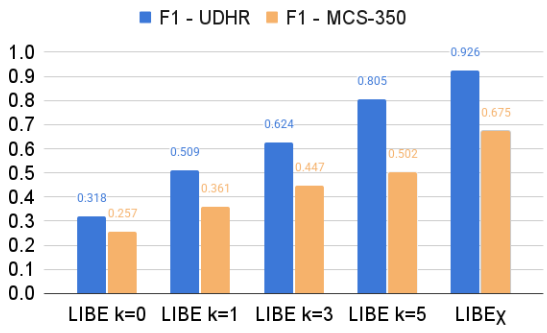

Figure 2: LIBE F1 results on the unseen languages in UDHR and MCS-350.

suggests that the model learns transferable representations that capture cross-linguistic regularities beyond simple memorization of features.

Introducing even minimal supervision leads to substantial gains. With just one labeled example per unknown language ($k = 1$), performance increases by more than 20 F1 points on UDHR and over 10 points on MCS-350. Further increasing the number of examples yields consistent and large improvements, reaching 0.926 F1 on UDHR and 0.675 F1 on MCS-350 with only a few additional examples. In contrast, FastText is unable to meaningfully predict any unseen language in this setting, even when *few*-shot examples are provided. This behavior can be attributed to its reliance on fixed label representations and shallow classification mechanisms, which limit its ability to adapt to new classes with minimal supervision. As a result, FastText fails to generalize beyond the set of languages observed during training.

The strong *few*-shot gains observed for LIBE highlight a key advantage of its architecture. By learning a shared representation space that captures language-level similarities, LIBE can effectively leverage a handful of examples to rapidly adapt to new languages. This capability is particularly beneficial in low-resource scenarios, where collecting large annotated datasets is infeasible. The results also show that *few*-shot learning is more impactful on UDHR than on MCS-350, reflecting the additional difficulty posed by shorter sentences and stronger domain shift in the latter. Overall, this analysis confirms that *few*-shot adaptation plays a crucial role in enabling language identification in realistic low-resource settings.

Error Analysis. We analyze extreme error cases (error rates > 0.9) on UDHR and MCS-350 to

characterize the limitations of the bi-encoder and cross-encoder models. In both datasets, most errors arise among closely related languages and dialect continua, where fine-grained distinctions are inherently difficult. On UDHR, frequent confusions occur within the Quechuan languages (e.g., `que`, `qvn`, `qva`, `qug`) and among Sinitic varieties (e.g., `hsn`, `wuu`, `cjy`, `yue`, `cmn`), indicating that many misclassifications correspond to linguistically plausible alternatives.

The same trend is observed on the more challenging MCS-350 dataset, where the bi-encoder exhibits a larger number of extreme cases, often confusing typologically or geographically proximate languages (e.g., `prs` with `fas/pus`, or `tgl` with other Austronesian languages). In both settings, the cross-encoder substantially reduces the number of extreme cases and produces more constrained confusions, typically limited to closely related languages. Overall, these errors reflect intrinsic linguistic ambiguity rather than systematic model deficiencies, while cross-encoder refinement consistently mitigates severe failures.

Model	Seq. Len	Throughput	F1
fastText	sent. len.	7,998.8	0.794
LIBE	16	16,136.2	0.817
LIBE	32	9,125.1	0.885
LIBE	64	4,765.9	0.900
LIBE	96	3,173.4	0.901
LIBE	112	2,743.9	0.901

Table 6: Throughput expressed in texts per second and F1 results on FLORES-200 DEV, varying the max sequence length of LIBE.

6.5 Training Duration and Inference Latency

All experiments used a workstation with an Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz, 1.5 TB RAM, and an NVIDIA Quadro RTX 8000 GPU (48 GB). The largest FastText model, trained on 10M sentences, completed within one hour, whereas LIBE training on the same dataset required 36 hours.

Table 6 reports the trade-off between throughput and F1 score on FLORES-200 DEV when varying the maximum sequence length of LIBE. The results show that very short sequences provide the highest throughput but lower accuracy: with a maximum length of 16 tokens, LIBE processes 16,136.2 texts per second, more than twice as fast as fastText and also more accurate. Increasing the maximum sequence length substantially improves performance. Setting the limit to 64 tokens yields 0.900 F1, after which the gains saturate. Overall, the table shows that a maximum sequence length of 64 provides a strong accuracy–efficiency trade-off, while longer inputs bring negligible F1 improvements at a higher computational cost.

7 Conclusion

We introduced LIBE, a representation-based approach to language identification that embeds texts and natural-language descriptions of languages into a shared vector space. By formulating LID as retrieval over language descriptions, LIBE supports flexible candidate-language expansion and few-shot adaptation to languages with limited task-specific supervision. Across Flores-200, UDHR, and MCS-350, LIBE shows strong zero- and few-shot adaptation when initialized from pretrained multilingual encoders. It can incorporate new candidate languages through descriptions and adapt with limited task-specific supervision.

The revised analyses show that both components of the approach are important. Multilingual pretraining provides transferable representations, especially for languages covered during pretraining, while the semantic content of language descriptions is essential: label-only and shuffled-description controls perform substantially worse. The optional cross-encoder further improves robustness in ambiguous cases involving closely related languages and dialect continua.

The current formulation remains limited to sentence-level, single-label LID and does not directly address code-switching or multi-label language identification. Future work will therefore focus on extending the framework to mixed-language inputs, improving inference efficiency, and investigating richer or more standardized forms of language descriptions for fine-grained language identification.

Limitations. The data-efficiency results should be interpreted as task-specific LID supervision efficiency, not total training-data efficiency. LIBE relies on pretrained multilingual encoders and thus benefits from representation learning before LID training. Comparisons with fastText models trained only on LID corpora therefore do not show that LIBE uses less total training data; rather, they show that, given a pretrained encoder, the retrieval-based formulation can adapt to LID with fewer labeled examples.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. LIMIT: Language identification, misidentification, and translation using hierarchical models in 350+ languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14496–14519, Singapore. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. ArXiv:2004.05150v2.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ralf D. Brown. 2013. Selecting and weighting n-grams to identify 1100 languages. In *Text, Speech, and Dialogue - 16th International Conference, TSD 2013, Pilsen, Czech Republic, September 1-5, 2013. Proceedings*, volume 8082 of *Lecture Notes in Computer Science*, pages 475–483. Springer.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Common Crawl. 2026. Common Crawl: Open repository of web crawl data. `https://commoncrawl.org`. Accessed: 2026-01-15.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. ArXiv:1911.02116v2.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Binod Gyawali, Gabriela Ramirez, and Thamar Solorio. 2013. Native language identification: a simple n-gram based approach. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–231, Atlanta, Georgia. Association for Computational Linguistics.

Matthew L. Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. ArXiv:1705.00652v1.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *J. Artif. Intell. Res.*, 65:675–782.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Rafael Dueire Lins and Paulo Gonçalves. 2004. Automatic language identification of written texts. In *Proceedings of the 2004 ACM Symposium on Applied Computing (SAC), Nicosia, Cyprus, March 14-17, 2004*, pages 1128–1133. ACM.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. ArXiv:1907.11692v1.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Yukio Nakamura. 1971. Identification of languages with short sample texts. *Library and Information Science*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey

Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. ArXiv:2207.04672v3.

Sebastian Nordhoff and Harald Hammarström. 2011. Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In *Proceedings of the First International Workshop on Linked Science*, volume 783 of *CEUR Workshop Proceedings*.

Pedro Ortiz Suarez, Sarah Luger, Laurie Burchell, Kenton Murray, and Catherine Arnett. 2025. 1st workshop on multilingual data quality signals. Workshop co-located with the Conference on Language Modeling (COLM 2025). Non-archival workshop.

Fuchun Peng and Dale Schuurmans. 2003. Combining naive bayes and n-gram language models for text classification. In *Advances in Information Retrieval, 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14-16, 2003, Proceedings*, volume 2633 of *Lecture Notes in Computer Science*, pages 335–350. Springer.

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.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Tommi Vatanen, Jaakko J. Väyrynen, and Sami Virpioja. 2010. Language identification of short text segments with n-gram models. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Ju-yeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yun-Hsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. 2025. EmbeddingGemma: Powerful and lightweight text representations. ArXiv:2509.20354v3.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. ArXiv:2402.05672v1.

Wenhui Wang, Hangbo Bao, Shaohan Huang,

Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Guangxuan Xiao, Ji Lin, Mickaël Seznec, Hao Wu, Julien Demouth, and Song Han. 2023a. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 38087–38099. PMLR.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023b. C-Pack: Packaged resources to advance general chinese embedding.

Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, Xiaopeng Zhang, and Qi Tian. 2024. QA-LoRA: Quantization-aware low-rank adaptation of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Acknowledgments

We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).

The authors used AI-assisted tools for language editing and for refining the presentation of Figure 1; all content, analyses, and claims were reviewed and verified by the authors.

This work is dedicated to the loving memory of Roberta.

A Language Descriptions

This appendix provides representative examples of the language descriptions used by LIBE. The examples are sampled from GlotLID-C and report the language name, ISO 639-3 code, most common script, top-level language family, normalized endangerment status derived from Wikidata, and description source. Descriptions are truncated to one line for presentation.

Abkhaz (abk) *script: Cyrl; family: Abkhaz-Adyge; status: Vulnerable; source: Wiki.*
Abkhaz,[b] also known as Abkhazian, is a Northwest Caucasian language most closely related to Abaza. It is spoken...

Acehnese (ace) *script: Arab; family: Austronesian; status: Threatened; source: Wiki.*
Acehnese or Achinese (Jawoë: بهسا اچيه) is an Austronesian language natively spoken by the Acehnese people in... b

Alur (alz) *script: Latn; family: Nilotic; status: Developing; source: Wiki.*
Alur is a variety of Southern Luo spoken by about 1.4 million people in northwestern Uganda and the north east of...

Ancient Hebrew (hbo) *script: Hebr; family: Afro-Asiatic; status: Dormant; source: Wiki.*
Ancient Hebrew. Ancient Hebrew, also known as Biblical Hebrew or Classical Hebrew, is the original language of...

Assyrian Neo-Aramaic (aii) *script: Syrc; family: Afro-Asiatic; status: Threatened; source: Wiki.*
Suret (Syriac: ܣܘܪܝܬ) (['suːrɪtʰ] or ['suːrɪθ]), also known as Assyrian refers to the varieties of Northeastern Neo...

Basque (eus) *script: Latn; family: Basque; status: Threatened; source: Wiki.*
Basque is the only surviving Paleo-European language spoken in Europe, predating the arrival of speakers of...

Bolivar-North Chimborazo Highland Quichua (qug) *script: Latn; family: Quechuan; status: Definitely endangered; source: Wiki.*
Chimborazo Highland Quichua is a stable indigenous language of Ecuador. It belongs to the Quechuan language...

Central Aymara (ayr) *script: Latn; family: Aymaran; status: Threatened; source: Wiki.*
Intermediate Aymara shares dialectical features with both Northern and Southern Aymara and is found in the...

Central Guerrero Nahuatl (ngu) *script: Latn; family: Uto-Aztecan; status: Vigorous; source: Wiki.*
The Guerrero Nahuatl language is a Nahuan language spoken by about 125,000 people in Mexico. It is also known...

Central Khmer (khm) *script: Khmr; family: Austroasiatic; status: National; source: Wiki.*
Khmer is an Austroasiatic language spoken by the Khmer people, and the official and national language of...

Central Mnong (cmo) *script: Khmr; family: Austroasiatic; status: Developing; source: Wiki.*
The Mnong language (also known as Pnong or Bunong) (Bunong: ษูธิษ) belongs to the Austro-Asiatic language...

Chechen (che) *script: Cyrl; family: Nakh-Daghestanian; status: Vulnerable; source: Wiki.*
Chechen (Нохчийн мотт, Noxçiyn mott, ['nɔxʧɪ̈ː muɔt]) is a Northeast Caucasian language spoken by approxi...

Cherokee (chr) *script: Cher; family: Iroquoian; status: Moribund; source: Wiki.*
ᏣᎳᎩ ᎦᏬᏂᎯᏍᏗ Cherokee or Tsalagi (Cherokee: ᏣᎳᎩ ᎦᏬᏂᎯᏍᏗ, romanized: Tsalagi Gawonihisdi, IPA: [dʒala'gî...

Classical Syriac (syc) *script: Syrc; family: Afro-Asiatic; status: Second-language only; source: Wiki.*
The Syriac language, also known as Syriac Aramaic (Syrian Aramaic, Assyrian Aramaic, Syro-Aramaic) and...

Coptic (cop) *script: Copt; family: Afro-Asiatic; status: Extinct; source: Wiki.*
Coptic (Bohairic Coptic: ⲧⲙⲉⲧⲣⲉⲙ̀ⲛ̀ⲭⲏⲙⲓ, Timetremǹkhēmi) is a group of closely related Egyptian dialects, represent...

Dawro (dwr) *script: Ethi; family: Ta-Ne-Omotic; status: Educational; source: Ethn.*
Dawro is a stable indigenous language of Ethiopia. It belongs to the Afro-Asiatic language family. Direct evidence...

Dzongkha (dzo) *script: Tibt; family: Sino-Tibetan; status: Vulnerable; source: Wiki.*
Dzongkha (རྫོང་ཁ; [ʦ͡sòŋkʰá]) is a Sino-Tibetan language that is the official and national language of Bhutan. It is...

Eastern Armenian (hye) *script: Armn; family: Indo-European; status: National; source: Wiki.*
Armenian (endonym: հայերեն (reformed), հայերէն (classical), hayeren, pronounced [hɑjɛˈɾɛn]) is an Indo...

Eastern Canadian Inuktitut (ike) *script: Cans; family: Eskimo-Aleut; status: Provincial; source: Wiki.*
Inuktitut , also known as Eastern Canadian Inuktitut, is one of the principal Inuit languages of Canada. It is spoken...

Eastern Mari (mhr) *script: Cyrl; family: Uralic; status: Definitely endangered; source: Wiki.*
Mari is a Uralic language natively spoken by about 320,000(data: 2020, 400,000 in total) people in Russia. The...

Eastern Panjabi (pan) *script: Guru; family: Indo-European; status: Safe; source: Wiki.*
Europe North America Oceania Punjabi , sometimes spelled Panjabi,[d] is an Indo-Aryan language native to the...

Gamo (gmv) *script: Ethi; family: Ta-Ne-Omotic; status: Educational; source: Wiki.*
Gamo is a stable indigenous language of Ethiopia. It belongs to the Afro-Asiatic language family. Direct evidence is,...

Garo (grt) *script: Beng; family: Sino-Tibetan; status: Provincial; source: Wiki.*
Garo, also referred to by its endonym A·chikku, is a Tibeto-Burman language spoken in the Northeast Indian states...

Georgian (kat) *script: Geor; family: Kartvelian; status: National; source: Wiki.*
Georgian (ქართული ენა, romanized: kartuli ena, pronounced ['kʰartʰuli 'ena]) is the most widely spoken...

Gujarati (guj) *script: Gujr; family: Indo-European; status: Provincial; source: Wiki.*
Gujarati is an Indo-Aryan language native to the Indian state of Gujarat and spoken predominantly by the Gujarati...

Halh Mongolian (khk) *script: Cyrl; family: Mongolic-Khitan; status: National; source: Wiki.*

Continued on next page

The Khalkha dialect is a dialect of central Mongolic widely spoken in Mongolia. According to some classifications,...

Hausa States Fulfulde (fuv) *script: Arab; family: Atlantic-Congo; status: Wider communication; source: Wiki.*
Nigerian Fulfulde is a language of wider communication that originated in Nigeria. It belongs to the Niger-Congo...

Ho (hoc) *script: Wara; family: Austroasiatic; status: Vulnerable; source: Wiki.*
Ho is a Munda language of the Austroasiatic language family spoken primarily in India by about 2.2 million people...

Japanese (jpn) *script: Jpan; family: Japonic; status: Safe; source: Wiki.*
Japanese (日本語, Nihongo, [ɲihoŋgo]) is the principal language of the Japonic language family spoken by the....

Javanese (jav) *script: Java; family: Austronesian; status: Safe; source: Ethn.*
Javanese is a stable indigenous language of Indonesia. It belongs to the Austronesian language family. Direct...

Juang (jun) *script: Orya; family: Austroasiatic; status: Definitely endangered; source: Wiki.*
The Juang language is a Munda language of the Austroasiatic language family spoken primarily by the Juang people...

K'iche' (quc) *script: Latn; family: Mayan; status: Educational; source: Wiki.*
K'iche' ([kʼiˈtʃʰeʔ], also known as Qatzijobʼal lit. 'our language' among its speakers), or Quiché , is a Mayan language...

Kabardian (kbd) *script: Cyrl; family: Abkhaz-Adyge; status: Vulnerable; source: Wiki.*
Surviving Destroyed or barely existing Kabardian [a] also known as East Circassian, is a Northwest Caucasian...

Kannada (kan) *script: Knda; family: Dravidian; status: Institutional; source: Wiki.*
Vijayanagara: (Origin. Empire. Musicological nonet. Medieval city. Military. Haridasa. Battle of Raichur. Battle of...

Korean (kor) *script: Hang; family: Koreanic; status: Safe; source: Wiki.*
Korean (South Korean: 한국어, hangugeo; North Korean: 조선말, chosŏnmal) is the native language for about 81...

Kurukh (kru) *script: Deva; family: Dravidian; status: Threatened; source: Wiki.*
Kurukh , also Kurux, Oraon or Uranw, is a North Dravidian language spoken by the Kurukh (Oraon) and Kisan...

Lisu (lis) *script: Lisu; family: Sino-Tibetan; status: Wider communication; source: Wiki.*
Lisu (Fraser alphabet: LI-SU, LI-SU ʌO: or LISU; Latin: Lisu ngot; Lisu syllabary: ; Chinese: 傈僳语; pinyin: Lìsùyǔ...

Malayalam (mal) *script: Mlym; family: Dravidian; status: Safe; source: Wiki.*
Malayalam is a Dravidian language spoken in the Indian state of Kerala and the union territories of Lakshadweep...

Manipuri (mni) *script: Mtei; family: Sino-Tibetan; status: Vulnerable; source: Wiki.*
Meitei (ꯃꯤꯇꯩ, Eastern Nagari script: মৈতৈলোন্, romanized: meiteilon), also known as Manipuri), is a Tibeto...

Mapudungun (arn) *script: Latn; family: Araucanian; status: Severely endangered; source: Wiki.*
Orange: rural Mapuche; Dark: urban Mapuche; White: non-Mapuche inhabitants Mapuche ; from mapu 'land' and...

Modern Greek (ell) *script: Grek; family: Indo-European; status: Institutional; source: Wiki.*
Modern Greek (endonym: Νέα Ελληνικά, Néa Elliniká, ['ne.a elini'ka] or Κοινή Νεοελληνική Γλώσσα, Kiní...

Modern Hebrew (heb) *script: Hebr; family: Afro-Asiatic; status: Safe; source: Wiki.*
Modern Hebrew (עִבְרִית חֲדָשָׁה ʿIvrīt ḥadašá [iv'ʁit χada'ʃa]), also called Israeli Hebrew or simply Hebrew, is the...

Nama (Namibia) (naq) *script: Latn; family: Khoe-kwadi; status: Threatened; source: Wiki.*
The Khoekhoe KOY-koy language (Khoekhoegowab, Khoekhoe pronunciation: [k̄xʰo͘e̯k̄xʰo͘e̯gowab]), also known by...

Northern East Cree (crl) *script: Cans; family: Algic; status: Vulnerable; source: Wiki.*
Northern East Cree is a stable indigenous language of Canada. It belongs to the Algiclanguage family and is part of...

Northern Gumuz (guk) *script: Ethi; family: Gumuz; status: Educational; source: Wiki.*
Gumuz (also spelled Gumaz) is a dialect cluster spoken along the border of Ethiopia and Sudan. It has been...

Northern Katang (ncq) *script: Laoo; family: Austroasiatic; status: Vigorous; source: Wiki.*
Northern Katang is a stable indigenous language of Laos. It belongs to the Austro-Asiatic language family. The...

Northern Khmer (kxm) *script: Thai; family: Austroasiatic; status: Developing; source: Wiki.*
Northern Khmer (พซาคแมร; Khmer: ខ្មែរខាងជើង), also called Surin Khmer (Khmer: ខ្មែរសុរិន្ទ), is the dialect of the Khmer...

Plains Cree (crk) *script: Cans; family: Algic; status: Vulnerable; source: Wiki.*
Plains Cree (endonym: ᓀᐦᐃᔨᐧᐃᐣ nēhiyawēwin; alternatively: <ᐢᑲ·Δ·ᕁ·ᓵ paskwâwinîmowin ""language of the...

Polish (pol) *script: Latn; family: Indo-European; status: National; source: Wiki.*
Polish (endonym: język polski, ['jɛ̃zɨk 'pɔlskʲi], polszczyzna [pɔl'ʂʧizna] or simply polski, ['pɔlskʲi]) is a West S...

Russia Buriat (bxr) *script: Cyrl; family: Mongolic-Khitan; status: Provincial; source: Wiki.*
Russia Buriat is an endangered indigenous language of the Russian Federation. It belongs to the Mongolic language...

Santali (sat) *script: Olck; family: Austroasiatic; status: Vulnerable; source: Wiki.*
Santali (Pronounced: [santaɽi], Ol Chiki: ᱥᱟᱱᱛᱟᱲᱤ, Bengali: সাঁওতালী, Odia: ସାନ୍ତାଳୀ, Devanagari: संताली, also known...

Shan (shn) *script: Mymr; family: Tai-Kadai; status: Wider communication; source: Wiki.*

Continued on next page

The Shan language is the native language of the Shan people and is mostly spoken in Shan State, Myanmar...

Sinhala (sin) *script: Sinh; family: Indo-European; status: National; source: Wiki.*
Sinhala , sometimes called Sinhalese (SIN-(h)ə-LEEZ, SING-(g)ə-LEEZ), is an Indo-Aryan language primarily spoken...

South Azerbaijani (azb) *script: Arab; family: Turkic; status: Wider communication; source: Wiki.*
South Azerbaijani is a language of wider communication that originated in Iran, Iraq, and Turkey. It belongs to the...

Southern East Cree (crj) *script: Cans; family: Algic; status: Moribund; source: Wiki.*
Southern East Cree is an endangered indigenous language of Canada. It belongs to the Algic language family and is...

Susu (sus) *script: Arab; family: Mande; status: Educational; source: Wiki.*
The Susu language (endonym: Susu: Sosoxui; French: Soussou) is the language of the Susu or Soso people of Guinea...

Swahili (swh) *script: Latn; family: Atlantic-Congo; status: National; source: Ethn.*
Swahili is the official national language of Tanzania. It belongs to the Niger-Congo language family and is part of...

Tabasaran (tab) *script: Cyrl; family: Nakh-Daghestanian; status: Vulnerable; source: Wiki.*
Tabasaran (also written Tabassaran) is a Northeast Caucasian language of the Lezgic branch. It is spoken by the...

Tamil (tam) *script: Taml; family: Dravidian; status: Safe; source: Wiki.*
Tamil[b] (தமிழ், Tamiḻ, pronounced [t̪amiɻ]) is a Dravidian language natively spoken by the Tamil people of South...

Tawallammat Tamajaq (ttq) *script: Tfng; family: Afro-Asiatic; status: Vigorous; source: Wiki.*
Tawellemmet (Tawəlləmmət) is the largest of the Tuareg languages in the Berber branch of the Afroasiatic family...

Telugu (tel) *script: Telu; family: Dravidian; status: Safe; source: Wiki.*
Telugu is a Dravidian language native to the Indian states of Andhra Pradesh and Telangana, where it is also the...

Thai (tha) *script: Thai; family: Tai-Kadai; status: Safe; source: Wiki.*
Thai,[a] or Central Thai[b] (historically Siamese;[c][d] Thai: ภาษาไทย), is a Tai language of the Kra-Dai language fam...

Tibetan (bod) *script: Tibt; family: Sino-Tibetan; status: Provincial; source: Wiki.*
Tibetan is the most widely spoken language in the Himalaya. It is spoken in Tibet, parts of China, northern Pakistan,...

Uighur (uig) *script: Arab; family: Turkic; status: Provincial; source: Wiki.*
Uyghur or Uighur is a Turkic language written in a Uyghur Perso-Arabic script with 8-13 million speakers, spoken...

Urak Lawoi' (urk) *script: Thai; family: Austronesian; status: Definitely endangered; source: Wiki.*
Urak Lawoi' or Urak Lawoc (Urak Lawoi': อูรักลาโวยจ, IPA: ['urʌk 'lawʊjʔ]) is a Malayic language spoken in southern....

Wolaytta (wal) *script: Ethi; family: Ta-Ne-Omotic; status: Wider communication; source: Wiki.*
Wolaitta or Wolayttatto Doonaa is a North Omotic language of the Ometo group spoken in the Wolayita Zone and...

Xibe (sjo) *script: Mong; family: Tungusic; status: Severely endangered; source: Wiki.*
The Sibe language (Xibe: ᠰᡞᠪᡝ ᡤᡳᠰᡠᠨ, romanized: sibe gisun, also Sibo, Xibe) is a Tungusic language spoken by...

Yue Chinese (yue) *script: Hani; family: Sino-Tibetan; status: Provincial; source: Wiki.*
Yue (Cantonese pronunciation: [jy:t̚]) is a branch of the Sinitic languages primarily spoken in Southern China,...