AfroScope: A Framework for Studying the Linguistic Landscape of Africa

Anonymous ACL submission

Figure 1: The ***AfroScope*** framework begins with **Dataset & Coverage**, where we maximize language coverage by aggregating multilingual datasets and metadata (language families, scripts) to construct the *AfroScope-Data*. We also employ rigorous decontamination to produce high-quality Evaluation Data. In **Model & Analysis**, we fine-tune baseline architectures introducing *AfroScope-Models*. To address fine-grained distinctions between closely related languages, we introduce *Mirror-Serengeti* for hierarchical classification. Finally, we evaluate these components through extensive cross-lingual analysis.

Abstract

Language Identification (LID), the task of determining the language of a given text, is a fundamental preprocessing step that affects the reliability of downstream NLP applications. While recent work has expanded African LID, existing approaches remain limited in *(i)* the number of supported languages and *(ii)* support for fine-grained distinctions among closely related varieties. We introduce ***AfroScope***, a unified framework for African LID that includes *AfroScope-Data*, a dataset covering 713 languages. We also present *AfroScope-Models*, a suite of strong LID models with broad African language coverage. To better separate highly confusable languages, we propose a hierarchical classification approach that leverages our new specialized embedding model, *Mirror-Serengeti*, that targets 29 closely related or geographically proximate languages. This approach improves macro-F_1 by 4.55 on this confusable subset compared to our best base model. Finally, we analyze cross-linguistic transfer and domain effects, offering guidance for building robust African LID systems. We position African LID as an enabling technology for large-scale measurement of Africa's linguistic landscape in digital text and release *AfroScope-Data* and *AfroScope-Models* online.[1]

1 Introduction

Scaling model size and web-scale pretraining have driven strong performance in modern Large Language Models (LLMs) (Raffel et al., 2020; Penedo et al., 2024, 2025). Yet, model behavior is tightly coupled to the distribution and quality of pretraining data (Grosse et al., 2023; Razeghi et al., 2022). As a result, these advances disproportionally benefit high-resource languages, while most of the world's $\sim$ 7,000 languages remain underserved (Eberhard et al., 2021; Grattafiori et al., 2024). Beyond data scarcity, low-resource languages also lack the mature curation pipelines and tools that are now mature for high-resource languages (Bi et al., 2024).

Language Identification (LID), the task of determining the language of a given text, is a foundational step in curating multilingual corpora from

[1] anonymous-link/

web crawls (Penedo et al., 2025). LID errors propagate to downstream stages such as tokenization (Duvenhage et al., 2017), filtering (Grattafiori et al., 2024; Li et al., 2024), and data scheduling for multilingual pretraining (Conneau et al., 2020; de Gibert et al., 2024; Laurençon et al., 2022). Crucially, LID systems determine not only how reliably each language is predicted but also the *scope* of identifiable languages. If a language is out of scope, its text is either dropped or misattributed to an in-scope language, distorting corpus composition and downstream evaluation (Costa-Jussà et al., 2022; Adebara et al., 2022a).

These issues are particularly acute for African languages, where major web-crawled corpora exhibit systematic quality problems (Kreutzer et al., 2022). Existing collections often contain substantial amounts of unusable or noisy text, including documents incorrectly attributed to African languages (Alabi et al., 2020). Such artifacts degrade downstream performance and can inflate apparent progress through superficial coverage gains (i.e.,*representation washing*) (Burchell et al., 2023), reinforcing persistent performance disparities (Blasi et al., 2022).

Domain skew further compounds the problem, as available data for many African languages is concentrated in religious texts and translations that do not reflect actual language use, hindering model robustness (Kargaran et al., 2023). With over $2,000$ African languages spanning diverse dialects, orthographies, and multilingual contexts (Eberhard et al., 2021; Hammarström et al., 2024), effective LID must support broad coverage while also distinguishing closely related languages and varieties.

Recent LID systems (Kargaran et al., 2023; Foroutan et al., 2025), including work targeting African languages (Adebara et al., 2022a; Ojo et al., 2025) have expanded the set of supported languages, but gaps remain in both *(i) scope* and *(ii) granularity*, i.e., separating closely related languages and their varieties. Moreover, dedicated analyses of multilingual transfer dynamics for African languages remain limited, despite their practical importance in data-scarce regimes (Longpre et al., 2025). In this work, we introduce ***AfroScope***, a unified framework for African LID that addresses these challenges and enables systematic study of cross-linguistic transfer. Figure 1 illustrates the entire framework. The ***AfroScope*** framework comprises three key contributions:

(i) Dataset and Models. We curate *AfroScope-Data*, a large-scale multilingual dataset spanning 713 African languages, with coverage across multiple orthographies and domains, offering a rich representation of the continent's linguistic breadth (§3). Using *AfroScope-Data*, we train *AfroScope-Models*, a family of LID models that improves over prior African LID baselines in our evaluation setting (§4).

(ii) Hierarchical disambiguation of closely related languages. We propose a hierarchical approach that leverages our new contrastive embedding model, *Mirror-Serengeti*, to better separate genetically related and geographically proximate languages that are frequently confused (§6.2). This design targets fine-grained discrimination among closely related language groups while retaining broad coverage.

(iii) Transfer and robustness analysis. Leveraging *AfroScope-Data*, we analyze performance by resource level, domain, and script (§5), and study multilingual transfer effects–including positive transfer and negative interference– as a function of language family structure and script overlap (§6.3). These analyses provide practical guidance for building and curating African LID systems.

2 Related Works

Linguistic diversity in Africa. Africa is among the most linguistically diverse regions globally, spanning many language families and typological profiles (Heine and Nurse, 2000; Eberhard et al., 2021). For NLP systems, this diversity manifests in phenomena that directly stress corpus curation and LID, including rich morphology, orthographic variation, and pervasive multilingual practices such as code-switching (Abdulmumin et al., 2024; Hussen et al., 2025). In addition, ISO macrolanguage groupings and closely related varities with fluid boundaries complicate labeling and evaluation, as distinct labels can correspond to highly similar surface forms and overlapping usage (Alabi et al., 2025). Recent work has responded with new resources and benchmarks (Ojo et al., 2023; Adelani et al., 2024; Olaleye et al., 2025; Adebara et al., 2025; Elmadany et al., 2025) and with African-focused models (Adebara et al., 2022a,b, 2024), highlighting sensitivity to data quality and coverage.

Data authenticity and corpus quality. Large multilingual corpora frequently contain misattributed text, ambiguous language codes, and other quality issues that disproportionately affect low-resource settings (Kreutzer et al., 2022). Prior studies of widely used multilingual resources and pipelines document systematic noise and labeling errors (Bañón et al., 2020; Schwenk et al., 2021; Xue et al., 2021), and emphasize the role of LID quality and preprocessing in mitigating such artifacts (Agarwal et al., 2023). These issues are often exacerbated in web-scale curation (Penedo et al., 2024, 2025), where resources may be incorporated with limited validation (Alabi et al., 2020; Lau et al., 2025). Improving authenticity is therefore central to building reliable and culturally representative language technologies (Ojo et al., 2023; Zhong et al., 2024; Alhanai et al., 2025), motivating dataset construction that explicitly controls for coverage, domain diversity, and contamination.

Progress in African language identification. Recent African LID research spans both efficient classifiers and transformer-based models, including FastText-based systems and African-focused pretrained models (Joulin et al., 2016b; Kargaran et al., 2023; Burchell et al., 2023; Adebara et al., 2022a,b, 2024). Methodologically, contrastive learning frameworks (Foroutan et al., 2025) and hierarchical approaches (Agarwal et al., 2023) that model confusion patterns have proven effective for distinguishing closely related languages and improving domain generalization. Despite this progress, existing systems still face limitations, and broad coverage that is robust across scripts, domains, and fine-grained confusable labels remains challenging.

3 *AfroScope-Data*

Building robust LID systems for African languages poses distinctive data challenges. We address these through a curation strategy guided by two objectives: *(i) maximizing language coverage* to reduce *out-of-model 'cousin'* errors (Caswell et al., 2020; Kreutzer et al., 2022), i.e., cases where text in an unsupported language is misattributed to the closest supported relative; and *(ii) ensuring domain diversity* to mitigate the narrow domain concentration in available African language data (Kargaran et al., 2023; Burchell et al., 2023).

To this end, we introduce *AfroScope-Data* (Table 1), a large-scale dataset spanning 713 African languages across 9 language families, 7 scripts, and 9 domains. *AfroScope-Data* is compiled from 11 publicly available datasets and contains 19,799,636 unique sentences. To the best of our knowledge, *AfroScope-Data* provides the broadest combined coverage of African languages and writing systems among publicly described sources for African LID.

3.1 Data Curation

We compile *AfroScope-Data* from published sources, prioritizing datasets that provide metadata enabling domain attribution. We treat `GlotLID` (Kargaran et al., 2023), `AfroLID` (Adebara et al., 2022a), and `SimbaText` (Elmadany et al., 2025) as *primary* sources due to their breadth and metadata availability, and augment them with eight additional *secondary* datasets to maximize coverage (Table 1).

We standardize language labels using ISO 639-3 codes and associate each label with *(i)* a language family hierarchy and *(ii)* writing system/script labels using metadata based on Ethnologue's writing database (Eberhard et al., 2021). We describe our preprocessing and splitting procedure in §4.

3.2 Coverage Across Family, Script, and Domain

A central objective of our curation is to support robus African LID by capturing diversity along three axes that strongly affect surface form and transfer: language family, writing system, and domain.

Language Family. *AfroScope-Data* covers languages spanning 9 high-level genealogical groupings (Figure 2) and cover contact/typological categories used in African contexts: *Afro-Asiatic, Austronesian, Creole, Indo-European, Khoe-Kwadi, Kx'a, Mixed language, Niger-Congo, Nilo-Saharan*. This breadth reduces reliance on cues specific to a single dominant family (e.g., *Niger-Congo*) and enables evaluation of cross-family generalization. We use the resulting hierarchy in our transfer analyses (§6.3).

Script. We include 7 writing systems: *Latin* (`Latn`), *Arabic* (`Arab`), *Ge'ez* (`Ethi`), *N'Ko* (`Nkoo`), *Tifinagh* (`Tfng`), *Coptic* (`Copt`), and *Vai* (`Vaii`). We explicitly label scripts for each language (*language_script*), as individual languages may employ multiple writing systems (e.g., *gof, ttq*), strenghtening model robustness across orthographic variations.

	Dataset	#Sent.	#Lang.	#Family.	#Script.	#Domain.
Primary	GlotLID (Kargaran et al., 2023)	30,682,541	451	7	5	Religious, Web, Stories, News
	AfroLID (Adebara et al., 2022a)	12,682,541	513	7	5	Religious, Web, Government, Benchmarks, News, Stories, Wikipedia, Health
	SimbaText (Elmadany et al., 2025)	382,541	101	5	4	Speech, Stories, News, Web, Religious
Secondary	FineWeb2 (Penedo et al., 2025)	31,424	466	9	6	Religious, Web, Stories, News
	Flores+ (NLLB Team et al., 2024)	108,486	51	5	4	Wikipedia, Benchmarks
	Mafand (Adelani et al., 2022)	54,795	21	4	2	News
	Smol (Caswell et al., 2025)	10,872	62	6	4	Benchmarks
	MCS-350 (Agarwal et al., 2023)	94,894	151	6	3	Stories
	Openlid (Burchell et al., 2023)	197,487	55	6	4	Wikipedia
	BLOOM (Leong et al., 2022)	686	133	4	3	Stories
	UDHR (Kargaran et al., 2023)	6,696	106	6	4	Government, Benchmarks
	Train	19,682,541				
	Dev	50,697	713	9	7	Religious, Speech, Government, Benchmarks, Stories, News, Health, Wikipedia, Web
	Test	66,398				

Speech Government Benchmarks Stories News Health Wikipedia Religious Web

Table 1: Summary statistics of the constituent datasets used to construct and evaluate *AfroScope-Data*. The table compares the ***Primary*** and ***Secondary*** sets across size, diversity, and domain distribution, alongside the final aggregated statistics for *Afroscope*. **# Sent.** refers to total number of sentences, **# Lang.** Number of languages, **# Family.** Number of Language Family **# Script.** Number of scripts, **# Domain.** Number of domain. Details regarding how various sources they are derived from are provided in Appendix A.

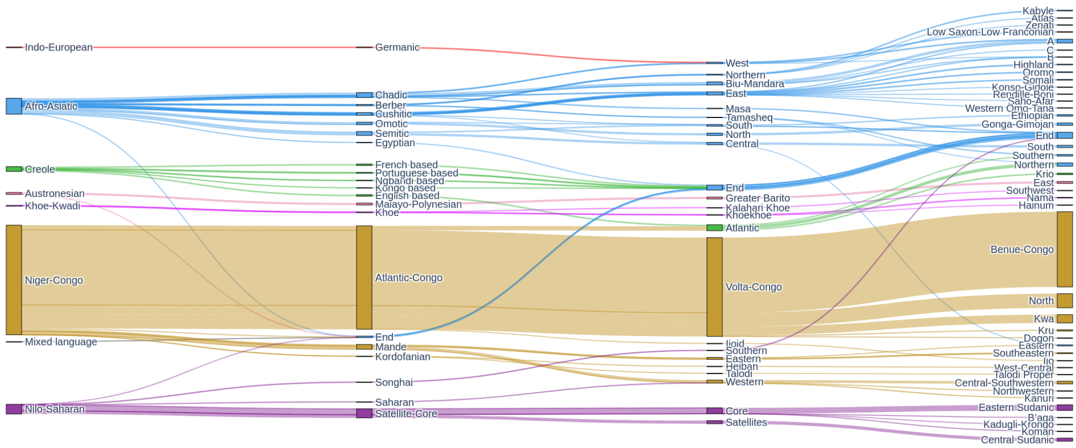

Figure 2: Distribution of languages across major language groupings, intermediate sub-families, and finer-grained groupings, capturing their genetic relationships.

Domain. To mitigate the narrow domain concentration in African language resources, we define 9 domain categories: *Speech, Government, Benchmarks, Stories, News, Health, Wikipedia, Religious* and *Web*. We assign domains using dataset metadata (e.g., URLs and source file names), mapping keywords to categories (Table A.1 in Appendix A.2). These labels enable controlled evaluation by domain and identify domain-specific challenges (§5).

4 *AfroScope-Models*

Using our *AfroScope-Data*, we fine-tune and evaluate a collection of LID models, *AfroScope-Models*. This section describes the experimental setup, evaluation settings, and baseline systems.

4.1 Experimental Setup

To balance representation across sources and reduce dominance by high-resource languages, we cap each language in *AfroScope-Data* at up to $100K$ sentences for training and 100 sentences for testing. Prior to sampling, we remove duplicate sentences across all sources to ensure the dataset contains only unique examples. We construct splits via a two-stage sampling procedure: we first sample from our *primary* sources, and for languages with fewer than 100K sentences available, we supplement using *secondary* sources (Table 1).

Internal vs. external evaluation. We report results under two complementary settings: *(i) internal* evaluation on the blind *AfroScope-Data* test split, and *(ii) external* evaluation on each constituent *secondary* dataset after filtering for potential leakage (Table 1). Specifically, when a language is sourced from a given secondary dataset for training, we exclude that language from the external evaluation set derived from the same dataset to avoid source-level contamination.

Contamination analysis. To quantify residual overlap between training and evaluation data, we measure 4-gram containment: we consider a test sentence contaminated if all of its 4-grams appear in a single training sentence. We observe minimal overlap (Table 2). Table 1 summarizes the resulting split statistics for *AfroScope-Data*[2].

Dataset	#Sent.	Contam.%	#Lang.	0–10%	≥10%
Afroscope	65,503	0	713	0	0
FineWeb2	28,236	0.02	416	414	2
Flores+	5,000	**19.86**	50	0	24
Mafand	2,000	5.20	20	1	2
MCS-350	9,179	4.16	105	69	19
SmolSent	4,800	0.02	48	1	0
OpenLID	5,200	**29.23**	52	3	31
BLOOM	5,200	2.23	131	12	90
UDHR	5,263	1.24	85	3	5

Table 2: Contamination rates of evaluation datasets against *AfroScope-Data train* split. We exclude ***OpenLID*** and ***Flores+*** from evaluation due to high data contamination rates.

4.2 Baselines and Metrics

We evaluate diverse LID approaches ranging from FastText classifiers to transformer-based models. Unless noted otherwise, all neural models are fine-tuned on *AfroScope-Data train* with the *dev* split used for best checkpoint selection. For FastText, we train on the merged set (*train* + *dev*) following common practice. We evaluate performance using macro-F_1, an aggregate measure of precision and recall.

FastText models. We train a custom *FastText* classifier (Joulin et al., 2016b) and evaluate *ConLID* (Foroutan et al., 2025), a recent FastText-based LID model that incorporates supervised contrastive learning to enhance robustness on out-of-domain data. Training hyperparameters are in Appendix B.1.

Neural Models. We evaluate three transformer-based models developed for African languages: *AfroLID* (Adebara et al., 2022a), *Serengeti* (Adebara et al., 2022b) (XLM-RoBERTa variant), and *Cheetah* (Adebara et al., 2024) (T5-based). Fine-tuning hyperparameters are in Appendix B.2.

[2]We examine the relationship between contamination rates and model performance in Appendix C.1

5 Results

Table 3 reports language-level performance on the *AfroScope-Data test* split, and Table 4 summarizes external evaluation on the constituent *secondary* datasets. Unless stated otherwise, analysis in this section refer to our best-performing model.

Performance by language resource level. Figure 3 plots per-language performance against the number of available training sentences and suggests an inflection point at $\sim$980 sentences, beyond which average performance approaches 95 macro-F_1. Motivated by this trend (and using thresholds on a log scale), we partition languages into three groups: *low-resource languages* ($<$ 98 sentences, n=47), *medium-resource languages* (98–980 sentences, n=22), and *high-resource languages* ($>$ 980 sentences, n=644). *Low-resource languages* (e.g., *Bamun* `(bax)`, and *Gichuka* `(cuh)`) exhibit high variance and low average performance (avg. macro-F_1: 41.60), consistent with severe data scarcity. *Medium-resource languages* (e.g., *Wongo* `(won)` and *Saya* `(say)`) improve rapidly (avg. macro-F_1: 89.10), indicating that on the order of 10^3 sentences can yield strong LID performance for many languages. *High-resource languages* (e.g., *Mbay* `(myb)` and *Karaboro* `(xrb)`) (avg. macro-F_1: 97.68) largely plateau, showing diminishing returns as training data increases; in some cases aditional data correlates with small degradations, which may be consistent with greater heterogeneity (e.g., non-standard orthography or noisier sources).

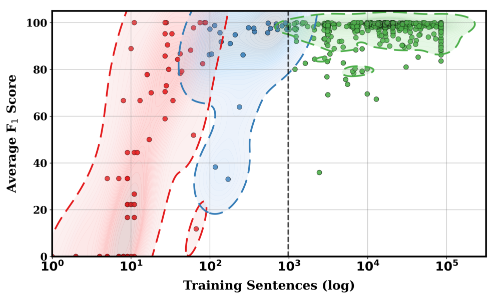

Figure 3: Relationship between training data size (log scale) and average macro-F_1 across low-resource, medium-resource, and high-resource languages.

Performance by domain. Figure 4 shows that domain substantially affects both average performance and stability. *Religious* and *News* tend to

	Language	Afrolid	Seregenti	Cheetah	Fine-Tuned on *Afroscope*				
					Afrolid	Serengeti	Cheetah	FastText	Conlid
High	Abé (aba)	91.54	92.37	96.05	94.95	95.38	95.88	95.83	95.80
	Afar (aar)	84.92	87.91	90.65	98.49	99.50	100	82.56	98.37
	Abidji (abi)	0	0	0	100	100	100	99.50	100
	...	...	...	...	...	...	...	...	...
Mid	Kom (bkm)	0	0	0	98.63	98.67	96.10	70.59	91.42
	Sherbro (bun)	90.00	90.76	91.25	95.24	97.67	95.45	83.54	89.74
	Bullom So (buy)	95.34	97.96	98.48	98.45	98.97	98.45	93.94	95.09
	...	...	...	...	...	...	...	...	...
Low	Ghotuo (aaa)	0	0	0	0	0	0	0	0
	Adangbe (adq)	0	0	0	0	0	0	0	0
	Esimbi (ags)	0	0	0	0	100	100	60.06	100
	...	...	...	...	...	...	...	...	...

Table 3: Per-language macro-F_1 scores comparing baseline models and models fine-tuned on *AfroScope-Data* across resource levels. We provide full results per-langauge in Appendix C.

Dataset	# Lang.	Transformers			FastText	
		Afrolid	Serengeti	Cheetah	Fasttext	ConLID
Afroscope	713	97.16	**97.83**	97.73	78.30	87.17
BLOOM	55	95.76	92.43	**94.63**	85.00	87.95
FineWeb2	416	94.25	94.18	**94.52**	84.00	89.03
Mafand	20	91.02	92.92	**93.54**	73.32	85.43
MCS-350	105	66.33	69.78	**70.38**	56.55	63.54
Smol	48	88.70	**90.02**	89.44	78.20	81.55
UDHR	85	87.88	**89.68**	89.07	79.22	82.12

Table 4: Model performance (macro-F_1) across *secondary* datasets. **Bold** indicates best performance per dataset.

ß

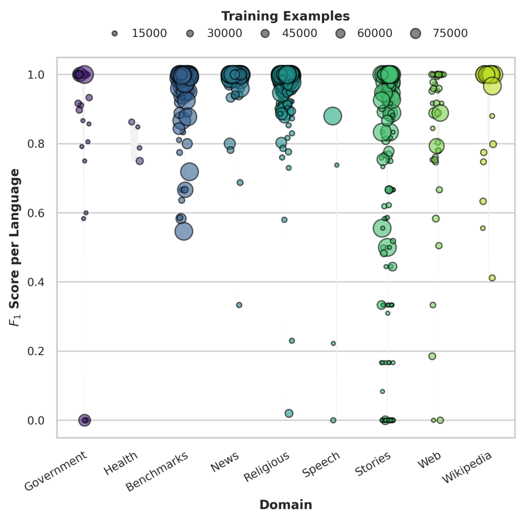

Figure 4: Per-language macro-F_1 scores across domains. Bubble size corresponds to training examples.

achieve high scores across many languages, while more specialized categories such as *Benchmarks*, *Stories*, and *Government* exhibit higher variance across languages. Finally, *Web* and *Wikipedia* are generally string but more dispersed, consistent with the heterogeneous and less standardized nature of open-domain text.

Performance by script. Languages written in less prevalent scripts in our data (e.g., *Coptic*, *Tifinagh*, and *N'Ko*) underperform relative to those written in *Latin* and *Ethiopic* scripts on average, suggesting that limited script coverage may be hindering generalization. Arabic-script languages achieve moderate performance (macro-F_1: 89.03) but show increased confusion among closely related varieties (e.g., *arz*, *ary*) within the Arabic (*ara*) macrolanguage, suggesting that high script similarity can make fine-grained discrimination more challenging.

6 Discussion

While LID performance typically improves with more training data, we observe systematic outliers: some languages remain difficult even beyond the data-size inflection point, whereas others achieve strong results despite limited supervision. We discuss both patterns and connect them to *(i)* confusability among closely related labels and *(ii)* cross-lingual transfer effects.

6.1 Low performance despite sufficient training data

Even above the inflection point, several languages fail to reach the expected performance. To investigate this, we isolate *underperformers*—high-resource languages with scores lower than macro-F_1 85—and identify the top three most frequent misclassifications for each to form confusion groups. These groups reveals that *label confusability* is a primary failure mode, where closely related languages and varieties with substantial lexical and orthographic overlap remain difficult to

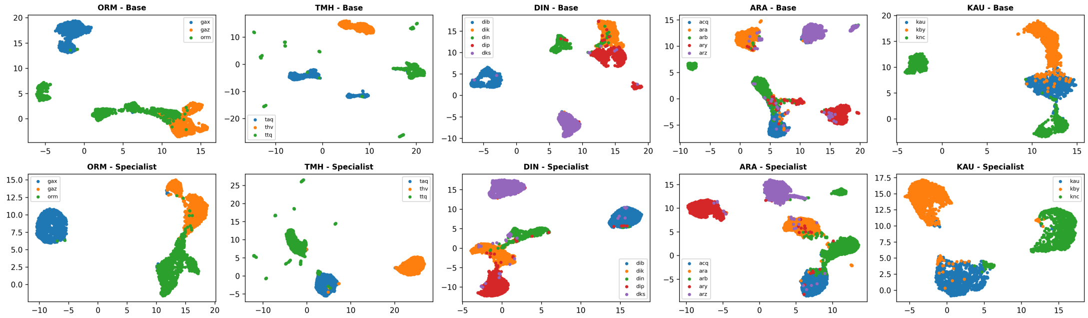

Figure 5: UMAP visualization comparing base *Serengeti* (top) and *Mirror-Serengeti*(bottom) embedding spaces. We visualize five groups representing macro-languages and confusion pairs. Specialized embeddings show improved separation between closely related language varieties.

Group	Language	Baseline F_1	F_1_0.75	**Δ_0.75**	F_1_0.8	**Δ_0.8**	F_1_0.85	**Δ_0.85**	F1_0.9	**Δ_0.9**	F_1_0.95	**Δ_0.95**
ful	fub	90.38	91.79	+1.40	**92.23**	**+1.85**	**92.23**	**+1.85**	91.26	+0.88	**92.23**	**+1.85**
swa	swh	87.18	87.76	+0.58	87.76	+0.58	**88.32**	**+1.15**	**88.32**	**+1.15**	87.31	+0.13
kon	kng	93.90	95.65	+1.76	96.12	+2.22	96.08	+2.18	**98.00**	**+4.10**	97.49	+3.59
	kon	80.00	84.44	+4.44	85.08	+5.08	86.49	+6.49	**88.89**	**+8.89**	88.42	+8.42
	ktu	**93.66**	94.58	+0.92	94.58	+0.92	94.58	+0.92	94.12	+0.46	**93.66**	**+0.00**
	kwy	96.15	96.62	+0.46	96.62	+0.46	**97.09**	**+0.93**	**97.09**	**+0.93**	**97.09**	**+0.93**
		......	...	...	...	...	...	...	...	...	...	...
		...	...	...	...	...	...	...	...	...	...	...
Average		–	–	+3.30	–	+3.93	–	+4.22	–	+4.20	–	**+4.55**

Table 5: Hierarchical classification results using *Mirror-Serengeti* embeddings across confidence thresholds (75%, 80%, 85%, 90%, 95%). Baseline F_1 shows base Serengeti performance; **Δ** columns show improvement over baseline. Bold indicates best performance per language. We provide full results on the full confusion groups in Appendix D.1.

separate, leading to persistent confusions. We frequently observe misclassifications stemming from macrolanguage structures, such as among Dinka (*din*) varieties (e.g., *dik* vs. *dip*), and between languages that occur in similar geographic and linguistic contexts (e.g., Konni [*kma*] vs. Farefare [*gur*] in Ghana). These patterns suggest that errors often reflect genuine linguistic similarity and label granularity rather than random noise. We provide further examples and analysis of these confusion groups in Appendix D.1.

6.2 Targeted disambiguation with *Mirror-Serengeti*

Motivated by these confusions, we introduce *Mirror-Serengeti*, a specialized embedding model trained to improve separation among frequently confused groups. We build *Mirror-Serengeti* on top of *Serengeti*, our strongest base model on the *AfroScope-Data* test split, and train it with **Mirror-BERT** (Liu et al., 2021), an unsupervised contrastive learning objective that pulls semantically similar representations together while pushing unrelated ones apart. Detailed training procedures and hyperparameters are in Appendix D.2.

Figure 5 compares the embedding spaces produced by *Serengeti* and *Mirror-Serengeti*, showing clearer separation for confusable labels and tighter within-label clustering. To leverage this improved separation, we implement a hierarchical inference scheme where low-confidence predictions trigger a specialized, group-specific disambiguation step utilizing *Mirror-Serengeti* embeddings. We evaluate this strategy on 29 confusable languages. Table 5 reports macro-F_1 over this confusable subset for confidence thresholds from 75% to 95%. Across thresholds, we observe consistent gains, increasing from +3.30 at 75% to +4.55 at 95%, indicating that target embeddings are particularly effective at resolving the most ambiguous cases.

At the language level, we observe improvements both for closely related varieties within macrolanguage groupings, e.g., Kongo (*kon*) varieties *kng* and *kwy* improve +4.10 and +0.93 respectively, Swahili (*swc*) variety *swh* gains +1.15, and Fulah (*ful*) variety *fub* improves +1.85. We also observe improvements for confusable regional pairs: *kma* and *kmy* improve 0.37 and 7.61 respectively,

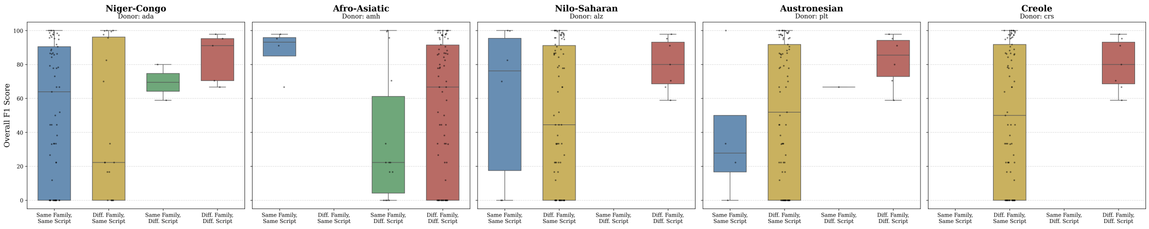

Figure 6: Transfer learning performance across language families and script compatibility. Box plots show macro-F_1 scores for languages grouped by their relationship to anchor languages.

and *gur* improves 0.50. However, a small number of languages decline (e.g., *ewo*: -0.98, and *kau*: -3.97), suggesting that hierarchical routing can add unnecessary complexity when the base model is already confident.

6.3 High performance under limited supervision

We also find cases where languages achieve strong performance despite having fewer than 980 training sentences. Building on recent work on multilingual transfer (Longpre et al., 2025), we investigate whether such gains are associated with language-family proximity and script compatibility. For each family, we select a high-resource *anchor* language (the language with the largest training set in that family) and group lower-resource recipient languages by their relationship to the anchor: same family, same script, both, or neither.

Figure 6 suggest that transfer patterns vary by family. For **Niger-Congo**, family proximity is strongly associate with positive transfer: recipients within the same family (blue) tend to outperform unrelated languages even when the unrelated languages share the same script (yellow). This indicates that shared linguistic hierarchy allows languages within large subfamilies such as *Volta-Congo* (e.g., *bkm*, *koq*) and *Benue-Congo* to leverage deep structural similarities, enabling them to overperform relative to their training data size. **Nilo-Saharan family** shows a similar tendency but with higher variance, potentially reflecting weaker or less uniform subfamily structure in the available data.

In contrast, for **Afro-Asiatic**, script compatibility appears to be a dominant factor: recepients sharing the anchor's script (e.g., *Ethi* in Ethiopic and *Tfng* in Tifinagh) (blue) show substantially stronger performance than recipients with script mismatch (green), suggesting that orthographic alignment is critical for transfer when scripts are highly distinctive.

Finally, **Austronesian** languages show limited evidence of positive transfer in our setting, with some same-family recipients (blue) underperforming relative to unrelated languages (yellow), consistent with negative interference when suitable donor signals are weak or mismatched.

7 Conclusion

We introduce ***AfroScope***, a unified framework for African language identification (LID) that combines broad coverage data, strong baselines, targeted disambiguation, and analysis. In addition, we present *AfroScope-Data*, a large-scale dataset spanning 713 language labels, and used to train *AfroScope-Models*, which outperforms prior African-focused LID baselines in our internal and external evaluations.

To mitigate persistent confusions among closely related languages and varieties, we propose a hierarchical inference approach based on a new specialized embedding model *Mirror-Serengeti*. On the identified confusion groups, our results show this approach improves macro-F_1 by $+4.55\%$ on average, with larger gains at higher-confidence thresholds.

Finally, our transfer analysis suggest that genealogical proximity and script are key correlates of positive multilingual transfer, helping some low-resource languages achieve strong LID performance with limited supervision. We hope these resources and findings support more robust and inclusive African NLP and enable future work on finer-grained varieties, domains shifts, and mixed-language text.

Limitations

We note several limitations:

1. **Mixed-language and code-switched text.** Our formulation treats each instance as belonging to a single language label. This does

not capture important phenomna in Africa's linguistic landscape such as code-switching mixed-language documents, and contact varieties (pidgins and creoles). Extending LID to multi-label or span-level identification is an important direction for future research.

2. **Language metadata and classification choices.** *AfroScope-Data*language, language family and script information is solely based

   We relay on external catalogs (primarily Ethnologue) to assign language identifiers, genealogical groupings, and script metadata. Alternative resources (e.g., Glottlog) may differ in classification and naming, which could affect analyses that depend on family hierarchy. Future work should evaluate sensitivity to these metadata choices and provide mappings across catalog standards.

3. **Confidence-based routing and calibration.** Our hierarchical disambiguation method relies on model confidence to decide when to invoke the group-specific refinement step. While this improves performance on many confusable labels, gains are not uniform across languages and thresholds, and some cases exhibit degradations. Improving probability calibration and learning routing policy (rather than using fixed thresholds) may further increase robstness.

References

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. Correcting FLORES evaluation dataset for four African languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 570–578, Miami, Florida, USA. Association for Computational Linguistics.

Ife Adebara, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Cheetah: Natural language generation for 517 african languages. *arXiv preprint arXiv:2401.01053*.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022a. Afrolid: A neural language identification tool for african languages. *arXiv preprint arXiv:2210.11744*.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022b. Serengeti: Massively multilingual language models for africa. *arXiv preprint arXiv:2212.10785*.

Ife Adebara, Hawau Olamide Toyin, Nahom Tesfu Ghebremichael, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2025. Where are we? evaluating llm performance on african languages. *arXiv preprint arXiv:2502.19582*.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, et al. 2024. Irokobench: A new benchmark for african languages in the age of large language models. *arXiv preprint arXiv:2406.03368*.

Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. Limit: Language identification, misidentification, and translation using hierarchical models in 350+ languages. *arXiv preprint arXiv:2305.14263*.

Jesujoba O. Alabi, Kwabena Amponsah-Kaakyire, David I. Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

Jesujoba O Alabi, Michael A Hedderich, David Ifeoluwa Adelani, and Dietrich Klakow. 2025. Charting the landscape of african nlp: Mapping progress and shaping the road ahead. *arXiv preprint arXiv:2505.21315*.

Tuka Alhanai, Adam Kasumovic, Mohammad M Ghassemi, Aven Zitzelberger, Jessica M Lundin, and Guillaume Chabot-Couture. 2025. Bridging the gap: enhancing llm performance for low-resource african languages with new benchmarks, fine-tuning, and cultural adjustments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27802–27812.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

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.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Moussa Doumbouya, Djibrila Diané, et al. 2025. Smol: Professionally translated parallel data for 115 under-represented languages. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1103–1123.

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.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Bernardt Duvenhage, Mfundo Ntini, and Phala Ramonyai. 2017. Improved text language identification for the south african languages. In *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*, pages 214–218. IEEE.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2021. *Ethnologue: Languages of the World*, 24 edition. SIL International, Dallas, Texas.

AbdelRahim Elmadany, Sang Yun Kwon, Hawau Olamide Toyin, Alcides Alcoba Inciarte, Hanan Aldarmaki, and Muhammad Abdul-Mageed. 2025. Voice of a continent: Mapping africa's speech technology frontier. *arXiv preprint arXiv:2505.18436*.

Negar Foroutan, Jakhongir Saydaliev, Ye Eun Kim, and Antoine Bosselut. 2025. Conlid: Supervised contrastive learning for low-resource language identification. *arXiv preprint arXiv:2506.15304*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. 2023. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.1. Accessed on 2025-04-03.

Bernd Heine and Derek Nurse, editors. 2000. *African Languages: An Introduction*. Cambridge University Press.

Kedir Yassin Hussen, Walelign Tewabe Sewunetie, Abinew Ali Ayele, Sukairaj Hafiz Imam, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025. The state of large language models for african languages: Progress and challenges. *arXiv preprint arXiv:2506.02280*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

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, Alex Wang, Abdul Wahab, Naman Goyal, Noah Constant, Xinying Chen, Guillaume Wenzek, Vishrav Chaudhary, Francisco Guzmán, Philipp Koehn, Ondřej Bojar, Christian Federmann, Nizar Habash, Yulia Tsvetkov, Holger Schwenk, and Alexis Conneau. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6825–6843.

Mingfei Lau, Qian Chen, Yeming Fang, Tingting Xu, Tongzhou Chen, and Pavel Golik. 2025. Data quality issues in multilingual speech datasets: The need for sociolinguistic awareness and proactive language planning. *arXiv preprint arXiv:2506.17525*.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. 2024. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. *arXiv preprint arXiv:2104.08027*.

Shayne Longpre, Sneha Kudugunta, Niklas Muennighoff, I Hsu, Isaac Caswell, Alex Pentland, Sercan Arik, Chen-Yu Lee, Sayna Ebrahimi, et al. 2025. Atlas: Adaptive transfer scaling laws for multilingual pretraining, finetuning, and decoding the curse of multilinguality. *arXiv preprint arXiv:2510.22037*.

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 Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic 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. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Jessica Ojo, Zina Kamel, and David Ifeoluwa Adelani. 2025. Divers-bench: Evaluating language identification across domain shifts and code-switching. *arXiv preprint arXiv:2509.17768*.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David Ifeoluwa Adelani. 2023. How good are large language models on african languages? *arXiv preprint arXiv:2311.07978*.

Kayode Olaleye, Arturo Oncevay, Mathieu Sibue, Nombuyiselo Zondi, Michelle Terblanche, Sibongile Mapikitla, Richard Lastrucci, Charese Smiley, and Vukosi Marivate. 2025. AfroCS-xs: Creating a compact, high-quality, human-validated code-switched dataset for African languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33391–33410, Vienna, Austria. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

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. *arXiv preprint arXiv:2506.20920*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot reasoning. *arXiv preprint arXiv:2202.07206*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2021. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Ansari Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, et al. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv preprint arXiv:2412.04497*.

Appendices

The following appendices provide comprehensive supplementary material supporting the main findings of this work. We include detailed descriptions of the datasets used, models, and experimental setup.

- §A: *AfroScope-Data* constituent datasets
- §B: Baseline Models
- §C: Evaluation Results
- §D: Discussions

A Data Collection and Corpus Curation

A.1 Constituent Datasets.

Below, we list the constituent datasets that make up *AfroScope-Data*. We prioritize primary datasets as they contain rich metadata for domain attribution. To reach the requirement of 100K training examples and 100 test examples, we supplement certain languages with data from secondary datasets.

Primary Datasets

GlotLID. We extract 528 African languages from `GlotLID-C`, a collection spanning 2,099 languages globally (Kargaran et al., 2023).

AfroLID. A manually curated multi-domain web dataset covering 516 African languages (Adebara et al., 2022a).

SimbaText. Speech-derived text data spanning 103 African languages, originally collected for speech and language identification (Elmadany et al., 2025).

Secondary Datasets

OpenLID. Manually audited data from news, Wikipedia, and religious texts across 55 African languages (Burchell et al., 2023).

Bloom Stories. A multimodal dataset for language modeling and visual storytelling covering 133 African languages (Leong et al., 2022).

MCS-350. Parallel children's stories across 151 African languages, drawn from a multilingual collection of 50k texts in over 350 languages (Agarwal et al., 2023).

Domain	Associated Keywords
Speech	Speech, CommonVoice, TTS, Audio
Government	Human Rights, Autshumato, Legal, GOV, Parliament, Gazette
Benchmarks	Flores, NLB, mt560, Tatoeba, UD, ai4d, lti, Benchmark, Human, Madar, iadd
Stories	Story, Stories, Fiction, Bloom, Lyrics
News	News, xlsum, Vukuzenzele, CBC, BBC, Afriqa, Masakha, Goud
Health	Health, Covid, Medical, Med
Wikipedia	Wiki, Leipzig, Wili, Encyclopedia
Religious	Bible, JW, Tanzil, PBC, Quran, Scripture, Religion
Web	Oscar, CC, CommonCrawl, Web, Dialect, Social, Forum

Table A.1: Keywords extracted from dataset metadata to map sources into domain categories. We provide examples of metadata in Appendix A.2.

SMOL. Professionally translated parallel data for 115 under-represented languages (Caswell et al., 2025).

A.2 Domain classification.

We assign domains by matching keywords found in the metadata associated with each sentence, following the categorization scheme from (Kargaran et al., 2023). Table A.1 lists the specific keyword mappings.

Sample Metadata
`Bible-aar_line94`
`Bible-aar_line392`
`Bible-aar_line115`
`CC100_zu.txt.tsv_f17_line64983`
`CC100_zu.txt.tsv_f17_line40162`
`JW-zul_line3295`

Table A.2: Examples of sentence-level metadata identifiers used for domain attribution.

B Hyperparameters for Baseline Models

B.1 FastText Models

FastText. (Joulin et al., 2016a) We follow the hyperparameters shown in Table B.1 to train our FastText model.

ConLID. (Foroutan et al., 2025): We follow the training procedure from the official GitHub repository[3] with hyperparameters detailed in Table B.2.

B.2 Neural Models

We use the hyperparameters in Table B.3 for *Afrolid*(Adebara et al., 2022a) and *Serengeti*(Adebara et al., 2022b) (both XLM-R variants) and Table B.4 for *Cheetah* (Adebara et al., 2024).

[3] https://github.com/epfl-nlp/ConLID

argument	description	value
-minCount	minimal number of word occurrences	1000
-minCountLabel	minimal number of label occurrences	0
-wordNgrams	max length of word ngram	1
-bucket	number of buckets	1e6
-minn	min length of char ngram	2
-maxn	max length of char ngram	5
-loss	loss function	softmax
-dim	size of word vectors	256
-epoch	number of epochs	2
-lr	learning rate	.8

Table B.1: FastText training hyperparameters

argument	description	value
-model_type	model architecture variant	conlid_s
-contrastive_temperature	contrastive loss temperature	0.05
-bank_size	memory bank size	2048
-optim	optimizer	adamw_torch
-lr_scheduler_type	learning rate scheduler	linear
-learning_rate	learning rate	0.004
-per_device_train_batch_size	batch size (per device)	128
-num_train_epochs	number of training epochs	1
-seed	random seed	42

Table B.2: ConLID training hyperparameters.

C Evaluation Results

C.1 Data Contamination and Model Performance

We observe no direct correlation between contamination rates and F_1 scores. For instance, `FineWeb2` and `Smol` exhibit negligible contamination (0.02%) yet achieve high performance (94.5% and 90.0%, respectively). Conversely, `MCS-350` has a higher contamination rate (4.16%) than most external datasets but records the lowest performance (70.4%). This suggests that the high scores on external benchmarks are not driven by training data leakage.

C.2 Results Per-Language

Full results per-language is presented in Table C.1, C.2,C.3, C.4, C.5.

D Discussions

D.1 Extended analysis of confusion groups

To investigate persistent errors, we isolate *underperformers*—high-resource languages with scores below F_1 85—and identify the top three most frequent misclassifications for each to form confusion groups. This analysis yields 14 distinct groups comprising 29 languages in total. We find that these confusions primarily stem from either macrolanguage structures (e.g., *ful* vs. *fub*) or geographic proximity (e.g., *bsq* vs. *bas*). Table D.1 details the composition of all confusion groups, and we report the corresponding performance improvements for each individual language.

argument	description	value
-max_seq_length	max input sequence length	128
-per_device_train_batch_size	training batch size (per device)	64
-learning_rate	learning rate	2e-5
-num_train_epochs	number of training epochs	10
-metric_for_best_model	evaluation metric	f1

Table B.3: Training hyperparameters for Afrolid and Serengeti.

argument	description	value
-max_target_length	max target sequence length	128
-per_device_train_batch_size	training batch size (per device)	32
-learning_rate	learning rate	5e-5
-num_train_epochs	number of training epochs	10
-metric_for_best_model	evaluation metric	f1

Table B.4: Training hyperparameters for the Cheetah model.

D.2 Mirror-BERT training procedure

We follow the procedures from the official github repository for Mirror-BERT[4]. We use *Serengeti* for this experiment as it being our best performing model. Table D.2 shows the hyperparamerters we use to train *Mirror-Serengeti*.

[4] https://github.com/cambridgeltl/mirror-bert

ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid	ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid	ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid
aaa	0.00%	0.00%	1.00%	0.00%	2.13%	bax	80.00%	92.31%	7.00%	92.31%	94.44%	bsp	99.00%	99.50%	100.00%	99.50%	100.00%
aar	98.49%	99.50%	100.00%	99.50%	100.00%	bba	99.50%	100.00%	100.00%	100.00%	100.00%	bsq	76.85%	81.00%	100.00%	81.00%	83.13%
aba	94.95%	95.38%	100.00%	95.38%	97.51%	bbj	94.00%	95.48%	100.00%	95.48%	97.61%	bss	99.50%	99.50%	100.00%	99.50%	100.00%
abi	100.00%	100.00%	100.00%	100.00%	100.00%	bbk	98.49%	98.04%	100.00%	98.04%	100.00%	bst	99.50%	99.50%	100.00%	99.50%	100.00%
abn	100.00%	100.00%	100.00%	100.00%	100.00%	bbo	100.00%	99.50%	100.00%	99.50%	100.00%	btt	100.00%	100.00%	100.00%	100.00%	100.00%
acd	100.00%	100.00%	100.00%	100.00%	100.00%	bce	100.00%	100.00%	2.00%	100.00%	100.00%	bud	100.00%	100.00%	100.00%	100.00%	100.00%
ach	96.55%	93.84%	100.00%	93.84%	95.97%	bci	98.99%	100.00%	100.00%	100.00%	100.00%	bum	97.54%	98.52%	100.00%	98.52%	100.00%
acq	91.30%	95.29%	100.00%	95.29%	97.42%	bcn	99.50%	99.50%	100.00%	99.50%	100.00%	bun	95.24%	97.67%	44.00%	97.67%	99.80%
ada	98.52%	100.00%	100.00%	100.00%	100.00%	bcw	98.04%	99.50%	100.00%	99.50%	100.00%	bus	99.50%	99.50%	100.00%	99.50%	100.00%
ade	100.00%	100.00%	100.00%	100.00%	100.00%	bcy	94.31%	95.08%	64.00%	95.08%	97.21%	buy	98.45%	98.97%	97.00%	98.97%	100.00%
adh	99.00%	99.50%	100.00%	99.50%	100.00%	bdh	100.00%	99.50%	100.00%	99.50%	100.00%	bwq	98.52%	100.00%	100.00%	100.00%	100.00%
adj	99.50%	100.00%	100.00%	100.00%	100.00%	bds	99.50%	100.00%	100.00%	100.00%	100.00%	bwr	98.51%	99.50%	100.00%	99.50%	100.00%
adq	0.00%	0.00%	1.00%	0.00%	2.13%	bec	0.00%	0.00%	1.00%	0.00%	2.13%	bwt	50.00%	0.00%	3.00%	0.00%	2.13%
aeb	86.11%	91.35%	100.00%	91.35%	93.48%	bem	96.55%	97.54%	100.00%	97.54%	99.67%	bwu	100.00%	100.00%	100.00%	100.00%	100.00%
afr	100.00%	100.00%	100.00%	100.00%	100.00%	beq	98.51%	99.50%	100.00%	99.50%	100.00%	bxk	91.28%	95.38%	100.00%	95.38%	97.51%
agq	100.00%	100.00%	100.00%	100.00%	100.00%	ber	91.49%	91.01%	100.00%	91.01%	93.14%	byf	84.42%	84.38%	100.00%	84.38%	86.51%
ags	0.00%	100.00%	1.00%	100.00%	100.00%	bex	100.00%	99.50%	100.00%	99.50%	100.00%	byv	94.69%	91.24%	100.00%	91.24%	93.37%
aha	99.50%	100.00%	100.00%	100.00%	100.00%	bez	99.50%	98.48%	100.00%	98.48%	100.00%	bza	100.00%	100.00%	100.00%	100.00%	100.00%
ajg	97.46%	97.98%	100.00%	97.98%	100.00%	bfa	97.98%	98.00%	100.00%	98.00%	100.00%	bze	40.00%	100.00%	4.00%	100.00%	100.00%
aka	85.71%	90.71%	100.00%	90.71%	92.84%	bfd	100.00%	100.00%	100.00%	100.00%	100.00%	bzw	99.50%	99.00%	100.00%	99.00%	100.00%
akp	100.00%	100.00%	100.00%	100.00%	100.00%	bfm	0.00%	0.00%	2.00%	0.00%	2.13%	cce	98.00%	98.02%	100.00%	98.02%	100.00%
Akuapim-twi	67.65%	70.77%	28.00%	70.77%	72.90%	bfo	99.50%	100.00%	100.00%	100.00%	100.00%	cgg	95.38%	95.96%	100.00%	95.96%	98.09%
ald	99.50%	99.50%	100.00%	99.50%	100.00%	bgf	0.00%	0.00%	1.00%	0.00%	2.13%	chw	99.00%	98.52%	100.00%	98.52%	100.00%
alz	98.51%	99.50%	100.00%	99.50%	100.00%	bhs	0.00%	0.00%	2.00%	0.00%	2.13%	cjk	99.00%	100.00%	100.00%	100.00%	100.00%
amf	99.50%	99.50%	100.00%	99.50%	100.00%	bib	100.00%	100.00%	100.00%	100.00%	100.00%	cko	99.00%	99.50%	100.00%	99.50%	100.00%
amh	99.01%	99.01%	100.00%	99.01%	100.00%	bim	98.48%	98.48%	100.00%	98.48%	100.00%	cme	99.50%	100.00%	100.00%	100.00%	100.00%
ann	100.00%	99.00%	100.00%	99.00%	100.00%	bin	100.00%	99.50%	100.00%	99.50%	100.00%	cop	78.54%	78.05%	100.00%	78.05%	80.18%
anu	98.99%	98.99%	100.00%	98.99%	100.00%	biv	100.00%	100.00%	100.00%	100.00%	100.00%	cou	98.99%	100.00%	100.00%	100.00%	100.00%
anv	100.00%	99.50%	100.00%	99.50%	100.00%	bjv	99.50%	100.00%	100.00%	100.00%	100.00%	cri	96.37%	96.91%	100.00%	96.91%	99.04%
any	98.52%	100.00%	100.00%	100.00%	100.00%	bkc	0.00%	0.00%	1.00%	0.00%	2.13%	crs	99.50%	99.00%	100.00%	99.00%	100.00%
apd	93.26%	94.42%	100.00%	94.42%	96.55%	bkh	0.00%	0.00%	1.00%	0.00%	2.13%	csk	100.00%	100.00%	100.00%	100.00%	100.00%
ara	75.71%	84.38%	100.00%	84.38%	86.51%	bkm	98.63%	98.67%	37.00%	98.67%	100.00%	cuh	0.00%	0.00%	2.00%	0.00%	2.13%
arb	87.50%	89.10%	100.00%	89.10%	91.23%	bkv	100.00%	100.00%	100.00%	100.00%	100.00%	cuv	0.00%	0.00%	1.00%	0.00%	2.13%
arq	94.30%	92.86%	100.00%	92.86%	94.99%	bky	100.00%	99.50%	100.00%	99.50%	100.00%	cwe	97.06%	96.15%	100.00%	96.15%	98.28%
ary	95.65%	97.56%	100.00%	97.56%	99.69%	blh	100.00%	100.00%	100.00%	100.00%	100.00%	cwt	98.99%	100.00%	100.00%	100.00%	100.00%
arz	84.93%	90.55%	100.00%	90.55%	92.68%	blo	98.51%	100.00%	100.00%	100.00%	100.00%	daa	98.04%	99.50%	100.00%	99.50%	100.00%
asa	97.98%	98.00%	100.00%	98.00%	100.00%	bmo	100.00%	100.00%	100.00%	100.00%	100.00%	dag	98.51%	99.50%	100.00%	99.50%	100.00%
Asante-twi	29.63%	44.44%	21.00%	44.44%	46.57%	bmq	98.52%	100.00%	100.00%	100.00%	100.00%	dav	95.88%	96.94%	100.00%	96.94%	99.07%
asg	98.99%	98.99%	100.00%	98.99%	100.00%	bmv	99.01%	98.99%	100.00%	98.99%	100.00%	dbq	98.52%	100.00%	100.00%	100.00%	100.00%
atg	99.00%	99.50%	100.00%	99.50%	100.00%	bob	0.00%	0.00%	1.00%	0.00%	2.13%	ddn	97.96%	99.50%	100.00%	99.50%	100.00%
ati	99.00%	98.00%	100.00%	98.00%	100.00%	bom	100.00%	99.50%	100.00%	99.50%	100.00%	dga	98.00%	99.01%	100.00%	99.01%	100.00%
avn	99.50%	99.50%	100.00%	99.50%	100.00%	bov	99.00%	99.50%	100.00%	99.50%	100.00%	dgd	100.00%	100.00%	100.00%	100.00%	100.00%
avu	100.00%	100.00%	100.00%	100.00%	100.00%	box	100.00%	100.00%	100.00%	100.00%	100.00%	dgi	100.00%	100.00%	100.00%	100.00%	100.00%
ayl	89.11%	95.48%	100.00%	95.48%	97.61%	boz	85.71%	100.00%	4.00%	100.00%	100.00%	dhm	99.00%	98.99%	100.00%	98.99%	100.00%
azo	98.99%	100.00%	100.00%	100.00%	100.00%	bqc	100.00%	99.01%	100.00%	99.01%	100.00%	dib	100.00%	100.00%	100.00%	100.00%	100.00%
bag	0.00%	0.00%	1.00%	0.00%	2.13%	bqj	100.00%	100.00%	100.00%	100.00%	100.00%	did	99.01%	99.50%	100.00%	99.50%	100.00%
bam	88.89%	90.23%	100.00%	90.23%	92.36%	bqm	66.67%	66.67%	2.00%	66.67%	68.80%	dig	97.51%	96.48%	100.00%	96.48%	98.61%
bas	91.74%	93.46%	100.00%	93.46%	95.59%	bqp	99.50%	100.00%	100.00%	100.00%	100.00%	dik	81.82%	85.34%	100.00%	85.34%	87.47%
bav	99.00%	99.00%	100.00%	99.00%	100.00%	bri	0.00%	0.00%	2.00%	0.00%	2.13%	din	74.21%	80.70%	100.00%	80.70%	82.83%
baw	60.00%	50.00%	4.00%	50.00%	52.13%	bsc	99.50%	100.00%	100.00%	100.00%	100.00%	dip	98.49%	96.97%	100.00%	96.97%	99.10%

Table C.1: Per-Language Results Part 1

ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid	ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid	ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid
diu	97.00%	98.48%	100.00%	98.48%	100.00%	fuq	98.48%	95.61%	100.00%	95.61%	97.74%	ibb	94.30%	96.48%	100.00%	96.48%	98.61%
dje	97.46%	98.99%	100.00%	98.99%	100.00%	fuv	88.89%	92.82%	100.00%	92.82%	94.95%	ibo	99.50%	100.00%	100.00%	100.00%	100.00%
dks	99.01%	98.51%	100.00%	98.51%	100.00%	fvr	100.00%	100.00%	9.00%	100.00%	100.00%	idu	99.01%	99.50%	100.00%	99.50%	100.00%
dnj	99.50%	100.00%	100.00%	100.00%	100.00%	gaa	100.00%	99.50%	100.00%	99.50%	100.00%	ife	100.00%	100.00%	100.00%	100.00%	100.00%
dop	100.00%	100.00%	100.00%	100.00%	100.00%	gax	97.44%	97.46%	100.00%	97.46%	99.59%	igb	93.88%	97.98%	100.00%	97.98%	100.00%
dos	100.00%	100.00%	100.00%	100.00%	100.00%	gaz	90.20%	93.40%	100.00%	93.40%	95.53%	ige	99.50%	99.50%	100.00%	99.50%	100.00%
dov	97.46%	95.29%	100.00%	95.29%	97.42%	gbo	99.50%	99.50%	100.00%	99.50%	100.00%	igl	98.99%	99.50%	100.00%	99.50%	100.00%
dow	99.50%	99.50%	100.00%	99.50%	100.00%	gbr	98.51%	99.00%	100.00%	99.00%	100.00%	ijc	100.00%	100.00%	4.00%	100.00%	100.00%
dsh	100.00%	100.00%	100.00%	100.00%	100.00%	gde	99.50%	100.00%	100.00%	100.00%	100.00%	ijn	99.50%	100.00%	100.00%	100.00%	100.00%
dts	99.50%	100.00%	100.00%	100.00%	100.00%	gej	95.15%	99.00%	100.00%	99.00%	100.00%	ijs	100.00%	100.00%	10.00%	100.00%	100.00%
dua	96.48%	98.49%	100.00%	98.49%	100.00%	gez	85.71%	100.00%	4.00%	100.00%	100.00%	ikk	99.50%	99.50%	100.00%	99.50%	100.00%
dug	97.51%	99.01%	100.00%	99.01%	100.00%	gid	99.50%	99.01%	100.00%	99.01%	100.00%	ikw	100.00%	100.00%	100.00%	100.00%	100.00%
dur	99.50%	100.00%	100.00%	100.00%	100.00%	giz	99.50%	100.00%	100.00%	100.00%	100.00%	ilb	97.09%	96.62%	100.00%	96.62%	98.75%
dwr	99.50%	99.50%	100.00%	99.50%	100.00%	gjn	99.50%	99.00%	100.00%	99.00%	100.00%	iqw	97.00%	97.96%	100.00%	97.96%	100.00%
dyi	100.00%	100.00%	100.00%	100.00%	100.00%	gkn	98.00%	99.00%	100.00%	99.00%	100.00%	iri	99.50%	100.00%	100.00%	100.00%	100.00%
dyo	99.50%	99.00%	100.00%	99.00%	100.00%	gkp	68.93%	72.73%	100.00%	72.73%	74.86%	irk	100.00%	99.01%	100.00%	99.01%	100.00%
dyu	93.19%	92.47%	100.00%	92.47%	94.60%	gmv	99.50%	98.02%	100.00%	98.02%	100.00%	ish	99.01%	100.00%	100.00%	100.00%	100.00%
ebr	100.00%	100.00%	100.00%	100.00%	100.00%	gna	98.99%	100.00%	100.00%	100.00%	100.00%	iso	98.52%	98.52%	100.00%	98.52%	100.00%
ebu	98.49%	99.50%	100.00%	99.50%	100.00%	gnd	99.50%	100.00%	100.00%	100.00%	100.00%	isu	66.67%	0.00%	2.00%	0.00%	2.13%
efi	98.04%	99.01%	100.00%	99.01%	100.00%	gng	99.50%	100.00%	100.00%	100.00%	100.00%	iyx	98.48%	99.50%	100.00%	99.50%	100.00%
ego	99.50%	100.00%	100.00%	100.00%	100.00%	goa	99.50%	100.00%	100.00%	100.00%	100.00%	izr	100.00%	99.50%	100.00%	99.50%	100.00%
eka	100.00%	99.50%	100.00%	99.50%	100.00%	gof	98.52%	98.52%	100.00%	98.52%	100.00%	izz	96.97%	96.52%	100.00%	96.52%	98.65%
ekm	0.00%	0.00%	2.00%	0.00%	2.13%	gog	97.51%	99.50%	100.00%	99.50%	100.00%	jab	57.14%	88.89%	5.00%	88.89%	91.02%
eko	98.51%	98.99%	100.00%	98.99%	100.00%	gol	100.00%	100.00%	100.00%	100.00%	100.00%	jbu	100.00%	100.00%	100.00%	100.00%	100.00%
emk	100.00%	100.00%	12.00%	100.00%	100.00%	gou	0.00%	0.00%	2.00%	0.00%	2.13%	jen	96.00%	100.00%	12.00%	100.00%	100.00%
enb	99.50%	100.00%	100.00%	100.00%	100.00%	gqr	99.50%	100.00%	100.00%	100.00%	100.00%	jgo	99.50%	99.50%	100.00%	99.50%	100.00%
eot	85.71%	85.71%	4.00%	85.71%	87.84%	gso	99.50%	99.50%	100.00%	99.50%	100.00%	jib	100.00%	100.00%	100.00%	100.00%	100.00%
eto	97.46%	98.99%	100.00%	98.99%	100.00%	gud	99.50%	99.00%	100.00%	99.00%	100.00%	jit	98.99%	100.00%	100.00%	100.00%	100.00%
ets	66.67%	66.67%	4.00%	66.67%	68.80%	guk	100.00%	99.50%	100.00%	99.50%	100.00%	jmc	96.59%	97.54%	100.00%	97.54%	99.67%
etu	99.01%	98.52%	100.00%	98.52%	100.00%	gur	98.99%	99.50%	100.00%	99.50%	100.00%	kab	89.91%	89.91%	100.00%	89.91%	92.04%
etx	100.00%	100.00%	100.00%	100.00%	100.00%	guw	100.00%	100.00%	100.00%	100.00%	100.00%	kam	98.00%	99.50%	100.00%	99.50%	100.00%
ewe	98.52%	97.54%	100.00%	97.54%	99.67%	gux	99.50%	99.50%	100.00%	99.50%	100.00%	kao	100.00%	100.00%	100.00%	100.00%	100.00%
ewo	89.00%	91.18%	100.00%	91.18%	93.31%	guz	99.50%	98.00%	100.00%	98.00%	100.00%	kau	80.00%	86.46%	100.00%	86.46%	88.59%
eza	99.01%	98.52%	100.00%	98.52%	100.00%	gvl	99.50%	100.00%	100.00%	100.00%	100.00%	kbn	100.00%	100.00%	100.00%	100.00%	100.00%
fak	90.32%	93.81%	100.00%	93.81%	95.94%	gwl	39.71%	32.79%	100.00%	32.79%	34.92%	kbo	100.00%	99.50%	100.00%	99.50%	100.00%
fal	99.01%	100.00%	100.00%	100.00%	100.00%	gwr	67.92%	70.50%	100.00%	70.50%	72.63%	kbp	100.00%	100.00%	100.00%	100.00%	100.00%
fan	92.09%	94.69%	100.00%	94.69%	96.82%	gya	99.50%	100.00%	100.00%	100.00%	100.00%	kbr	99.50%	99.50%	100.00%	99.50%	100.00%
fat	98.99%	100.00%	100.00%	100.00%	100.00%	hae	97.54%	99.50%	100.00%	99.50%	100.00%	kby	92.61%	93.47%	100.00%	93.47%	95.60%
ffm	88.54%	91.10%	100.00%	91.10%	93.23%	hag	99.00%	98.52%	100.00%	98.52%	100.00%	kcg	100.00%	100.00%	100.00%	100.00%	100.00%
fia	100.00%	99.22%	65.00%	99.22%	100.00%	har	90.00%	90.48%	22.00%	90.48%	92.61%	kck	97.03%	99.01%	100.00%	99.01%	100.00%
fip	96.94%	97.44%	100.00%	97.44%	99.57%	hau	99.01%	100.00%	100.00%	100.00%	100.00%	kcp	50.00%	80.00%	3.00%	80.00%	82.13%
fli	100.00%	100.00%	4.00%	100.00%	100.00%	hav	95.43%	96.59%	100.00%	96.59%	98.72%	kdc	93.78%	95.15%	100.00%	95.15%	97.28%
flr	99.00%	99.50%	100.00%	99.50%	100.00%	hay	92.39%	94.74%	100.00%	94.74%	96.87%	kde	96.55%	97.96%	100.00%	97.96%	100.00%
fon	98.51%	99.00%	100.00%	99.00%	100.00%	hbb	97.51%	99.50%	100.00%	99.50%	100.00%	kdh	100.00%	100.00%	100.00%	100.00%	100.00%
fub	91.94%	90.38%	100.00%	90.38%	92.51%	hdy	0.00%	100.00%	2.00%	100.00%	100.00%	kdi	99.50%	100.00%	100.00%	100.00%	100.00%
fuc	22.22%	66.67%	8.00%	66.67%	68.80%	heh	98.52%	98.04%	100.00%	98.04%	100.00%	kdj	99.50%	99.50%	100.00%	99.50%	100.00%
fue	96.00%	96.48%	100.00%	96.48%	98.61%	her	100.00%	100.00%	100.00%	100.00%	100.00%	kdl	99.50%	98.02%	100.00%	98.02%	100.00%
fuf	96.52%	99.50%	100.00%	99.50%	100.00%	hgm	96.37%	98.99%	100.00%	98.99%	100.00%	kdn	98.99%	98.48%	100.00%	98.48%	100.00%
fuh	87.44%	92.96%	100.00%	92.96%	95.09%	hig	99.00%	99.00%	100.00%	99.00%	100.00%	kea	97.56%	96.15%	100.00%	96.15%	98.28%
ful	77.17%	83.15%	100.00%	83.15%	85.28%	hna	94.79%	98.99%	100.00%	98.99%	100.00%	ken	99.00%	98.51%	100.00%	98.51%	100.00%

Table C.2: Per-Language Results 2

ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid	ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid	ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid
keo	100.00%	99.50%	100.00%	99.50%	100.00%	kwy	96.15%	96.15%	100.00%	96.15%	98.28%	lub	100.00%	99.50%	100.00%	99.50%	100.00%
ker	99.50%	100.00%	100.00%	100.00%	100.00%	kxc	100.00%	100.00%	100.00%	100.00%	100.00%	luc	100.00%	100.00%	100.00%	100.00%	100.00%
kez	99.50%	100.00%	100.00%	100.00%	100.00%	kyf	100.00%	99.50%	100.00%	99.50%	100.00%	lue	99.50%	100.00%	100.00%	100.00%	100.00%
khq	97.54%	98.99%	100.00%	98.99%	100.00%	kyq	99.50%	100.00%	100.00%	100.00%	100.00%	lug	95.61%	97.51%	100.00%	97.51%	99.64%
khy	99.01%	100.00%	100.00%	100.00%	100.00%	kzn	82.35%	86.36%	100.00%	86.36%	88.49%	lun	100.00%	99.00%	100.00%	99.00%	100.00%
kia	99.50%	100.00%	100.00%	100.00%	100.00%	kzr	98.99%	98.99%	100.00%	98.99%	100.00%	luo	98.52%	99.50%	100.00%	99.50%	100.00%
kik	97.56%	99.01%	100.00%	99.01%	100.00%	lai	99.50%	99.50%	100.00%	99.50%	100.00%	luy	0.00%	0.00%	1.00%	0.00%	2.13%
kin	99.50%	99.50%	100.00%	99.50%	100.00%	laj	99.00%	98.99%	100.00%	98.99%	100.00%	lwg	84.82%	92.31%	100.00%	92.31%	94.44%
kiz	99.50%	100.00%	100.00%	100.00%	100.00%	lam	97.49%	96.52%	100.00%	96.52%	98.65%	lwo	99.50%	100.00%	100.00%	100.00%	100.00%
kki	98.52%	99.50%	100.00%	99.50%	100.00%	lan	66.67%	66.67%	4.00%	66.67%	68.80%	maf	100.00%	100.00%	100.00%	100.00%	100.00%
kkj	99.01%	99.50%	100.00%	99.50%	100.00%	lap	98.49%	99.50%	100.00%	99.50%	100.00%	mas	97.00%	97.51%	100.00%	97.51%	99.64%
kln	93.68%	95.92%	100.00%	95.92%	98.05%	las	100.00%	100.00%	100.00%	100.00%	100.00%	maw	98.99%	98.51%	100.00%	98.51%	100.00%
klu	99.50%	99.01%	100.00%	99.01%	100.00%	ldi	98.49%	100.00%	100.00%	100.00%	100.00%	mbu	99.50%	98.99%	100.00%	98.99%	100.00%
kma	80.16%	80.16%	100.00%	80.16%	82.29%	lea	95.92%	98.48%	100.00%	98.48%	100.00%	mck	97.56%	99.50%	100.00%	99.50%	100.00%
kmb	99.50%	99.50%	100.00%	99.50%	100.00%	led	99.50%	99.01%	100.00%	99.01%	100.00%	mcn	99.01%	99.01%	100.00%	99.01%	100.00%
kmy	67.11%	67.11%	100.00%	67.11%	69.24%	lee	99.50%	99.01%	100.00%	99.01%	100.00%	mcp	98.52%	98.04%	100.00%	98.04%	100.00%
knc	91.08%	94.74%	100.00%	94.74%	96.87%	lef	99.50%	100.00%	100.00%	100.00%	100.00%	mcu	99.50%	100.00%	100.00%	100.00%	100.00%
knf	99.00%	99.50%	100.00%	99.50%	100.00%	leh	98.04%	99.50%	100.00%	99.50%	100.00%	mda	100.00%	100.00%	100.00%	100.00%	100.00%
kng	91.59%	93.90%	100.00%	93.90%	96.03%	lem	98.52%	100.00%	100.00%	100.00%	100.00%	mdm	99.01%	99.50%	100.00%	99.50%	100.00%
knk	99.00%	99.00%	100.00%	99.00%	100.00%	lfa	0.00%	0.00%	1.00%	0.00%	2.13%	mdy	100.00%	100.00%	100.00%	100.00%	100.00%
kno	99.50%	99.50%	100.00%	99.50%	100.00%	lgg	98.99%	98.99%	100.00%	98.99%	100.00%	men	99.50%	99.50%	100.00%	99.50%	100.00%
kny	99.50%	100.00%	100.00%	100.00%	100.00%	lgm	98.51%	99.50%	100.00%	99.50%	100.00%	meq	100.00%	99.01%	100.00%	99.01%	100.00%
kon	79.07%	80.00%	100.00%	80.00%	82.13%	lia	97.98%	98.99%	100.00%	98.99%	100.00%	mer	96.48%	97.00%	100.00%	97.00%	99.13%
koo	99.00%	99.50%	100.00%	99.50%	100.00%	lik	99.50%	99.00%	100.00%	99.00%	100.00%	mev	100.00%	98.99%	100.00%	98.99%	100.00%
koq	97.01%	97.78%	69.00%	97.78%	99.91%	lin	96.08%	97.54%	100.00%	97.54%	99.67%	mfe	99.01%	99.50%	100.00%	99.50%	100.00%
kpz	98.02%	100.00%	100.00%	100.00%	100.00%	lip	99.50%	99.50%	100.00%	99.50%	100.00%	mfg	99.50%	97.56%	100.00%	97.56%	99.69%
kqn	97.46%	98.99%	100.00%	98.99%	100.00%	lkb	0.00%	20.00%	8.00%	20.00%	22.13%	mfh	99.50%	99.50%	100.00%	99.50%	100.00%
kqo	100.00%	100.00%	100.00%	100.00%	100.00%	lke	81.97%	95.08%	31.00%	95.08%	97.21%	mfi	98.52%	100.00%	100.00%	100.00%	100.00%
kqp	100.00%	100.00%	100.00%	100.00%	100.00%	lko	36.36%	58.33%	14.00%	58.33%	60.46%	mfj	0.00%	0.00%	2.00%	0.00%	2.13%
kqs	98.99%	99.50%	100.00%	99.50%	100.00%	llb	95.24%	97.56%	100.00%	97.56%	99.69%	mfk	98.51%	98.51%	100.00%	98.51%	100.00%
kqy	99.50%	99.50%	100.00%	99.50%	100.00%	lln	100.00%	100.00%	100.00%	100.00%	100.00%	mfq	98.04%	99.01%	100.00%	99.01%	100.00%
kri	99.50%	99.50%	100.00%	99.50%	100.00%	lmd	99.50%	99.50%	100.00%	99.50%	100.00%	mfz	99.50%	99.00%	100.00%	99.00%	100.00%
krs	100.00%	100.00%	100.00%	100.00%	100.00%	lmp	100.00%	100.00%	100.00%	100.00%	100.00%	mgc	100.00%	100.00%	100.00%	100.00%	100.00%
krw	99.50%	100.00%	100.00%	100.00%	100.00%	lnl	100.00%	99.50%	100.00%	99.50%	100.00%	mgg	0.00%	0.00%	2.00%	0.00%	2.13%
krx	98.51%	97.98%	100.00%	97.98%	100.00%	lns	96.97%	98.49%	100.00%	98.49%	100.00%	mgh	99.00%	99.00%	100.00%	99.00%	100.00%
ksb	98.99%	99.00%	100.00%	99.00%	100.00%	lob	99.50%	100.00%	100.00%	100.00%	100.00%	mgo	98.00%	99.50%	100.00%	99.50%	100.00%
ksf	99.50%	99.00%	100.00%	99.00%	100.00%	log	97.51%	99.01%	100.00%	99.01%	100.00%	mgq	99.50%	99.50%	100.00%	99.50%	100.00%
ksp	98.49%	99.50%	100.00%	99.50%	100.00%	loh	0.00%	0.00%	1.00%	0.00%	2.13%	mgr	97.56%	97.56%	100.00%	97.56%	99.69%
kss	100.00%	100.00%	100.00%	100.00%	100.00%	lok	99.50%	99.50%	100.00%	99.50%	100.00%	mgw	97.49%	98.99%	100.00%	98.99%	100.00%
ktb	100.00%	100.00%	100.00%	100.00%	100.00%	lol	99.00%	100.00%	100.00%	100.00%	100.00%	mhi	99.00%	100.00%	100.00%	100.00%	100.00%
ktj	98.48%	98.49%	100.00%	98.49%	100.00%	lom	98.99%	99.50%	100.00%	99.50%	100.00%	mhw	95.29%	97.62%	85.00%	97.62%	99.75%
ktu	92.16%	93.66%	100.00%	93.66%	95.79%	loq	99.00%	99.50%	100.00%	99.50%	100.00%	mif	98.52%	98.04%	100.00%	98.04%	100.00%
ktz	97.62%	97.62%	43.00%	97.62%	99.75%	lot	97.51%	97.44%	100.00%	97.44%	99.57%	mkl	99.50%	100.00%	100.00%	100.00%	100.00%
kua	95.10%	97.09%	100.00%	97.09%	99.22%	loz	98.52%	99.50%	100.00%	99.50%	100.00%	mlg	79.52%	89.50%	100.00%	89.50%	91.63%
kub	99.01%	99.01%	100.00%	99.01%	100.00%	lro	100.00%	100.00%	100.00%	100.00%	100.00%	mlk	100.00%	66.67%	2.00%	66.67%	68.80%
kuj	99.50%	99.50%	100.00%	99.50%	100.00%	lsm	96.52%	98.99%	100.00%	98.99%	100.00%	mlr	96.37%	97.46%	100.00%	97.46%	99.59%
kus	99.50%	99.50%	100.00%	99.50%	100.00%	lth	95.92%	94.74%	100.00%	94.74%	96.87%	mlw	0.00%	0.00%	1.00%	0.00%	2.13%
kvj	99.50%	100.00%	100.00%	100.00%	100.00%	lto	100.00%	99.50%	100.00%	99.50%	100.00%	mmu	0.00%	66.67%	2.00%	66.67%	68.80%
kwn	98.02%	98.02%	100.00%	98.02%	100.00%	lts	0.00%	0.00%	1.00%	0.00%	2.13%	mmy	98.51%	100.00%	100.00%	100.00%	100.00%
kwu	66.67%	90.91%	6.00%	90.91%	93.04%	lua	98.52%	100.00%	100.00%	100.00%	100.00%	mne	0.00%	0.00%	1.00%	0.00%	2.13%

Table C.3: Per-Language Results 3

ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid	ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid	ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid
mnf	99.50%	100.00%	100.00%	100.00%	100.00%	ngn	81.97%	85.11%	100.00%	85.11%	87.24%	orm	87.88%	90.62%	100.00%	90.62%	92.75%
mnk	98.48%	98.49%	100.00%	98.49%	100.00%	ngp	98.49%	99.50%	100.00%	99.50%	100.00%	ozm	100.00%	99.50%	100.00%	99.50%	100.00%
mny	98.51%	98.51%	100.00%	98.51%	100.00%	nhr	99.50%	99.50%	100.00%	99.50%	100.00%	pae	0.00%	0.00%	1.00%	0.00%	2.13%
moa	100.00%	100.00%	100.00%	100.00%	100.00%	nhu	100.00%	100.00%	100.00%	100.00%	100.00%	pbi	99.01%	99.01%	100.00%	99.01%	100.00%
mor	100.00%	100.00%	100.00%	100.00%	100.00%	nih	97.00%	98.99%	100.00%	98.99%	100.00%	pcm	93.12%	94.30%	100.00%	94.30%	96.43%
mos	99.50%	99.50%	100.00%	99.50%	100.00%	nim	100.00%	100.00%	100.00%	100.00%	100.00%	pem	100.00%	100.00%	100.00%	100.00%	100.00%
moy	100.00%	100.00%	100.00%	100.00%	100.00%	nin	99.50%	100.00%	100.00%	100.00%	100.00%	pfe	99.50%	100.00%	100.00%	100.00%	100.00%
moz	99.50%	98.51%	100.00%	98.51%	100.00%	niq	87.44%	90.72%	100.00%	90.72%	92.85%	phm	99.50%	99.50%	100.00%	99.50%	100.00%
mpe	100.00%	100.00%	100.00%	100.00%	100.00%	niy	99.50%	100.00%	100.00%	100.00%	100.00%	pil	100.00%	100.00%	14.00%	100.00%	100.00%
mpg	100.00%	100.00%	100.00%	100.00%	100.00%	njd	0.00%	0.00%	2.00%	0.00%	2.13%	pkb	96.41%	96.45%	100.00%	96.45%	98.58%
mqb	99.50%	99.01%	100.00%	99.01%	100.00%	njy	66.67%	66.67%	2.00%	66.67%	68.80%	pko	98.00%	98.49%	100.00%	98.49%	100.00%
msc	98.51%	98.49%	100.00%	98.49%	100.00%	nka	99.00%	99.00%	100.00%	99.00%	100.00%	plt	86.21%	91.74%	100.00%	91.74%	93.87%
mse	99.50%	100.00%	100.00%	100.00%	100.00%	nko	100.00%	99.50%	100.00%	99.50%	100.00%	pny	99.01%	99.50%	100.00%	99.50%	100.00%
mua	100.00%	100.00%	100.00%	100.00%	100.00%	nku	100.00%	100.00%	11.00%	100.00%	100.00%	pnz	87.50%	96.97%	17.00%	96.97%	99.10%
mug	100.00%	100.00%	100.00%	100.00%	100.00%	nla	96.55%	98.99%	100.00%	98.99%	100.00%	pov	98.51%	98.48%	100.00%	98.48%	100.00%
muh	99.50%	100.00%	100.00%	100.00%	100.00%	nle	0.00%	0.00%	1.00%	0.00%	2.13%	poy	98.02%	99.50%	100.00%	99.50%	100.00%
mur	99.00%	100.00%	100.00%	100.00%	100.00%	nmz	100.00%	99.50%	100.00%	99.50%	100.00%	rag	99.50%	98.49%	100.00%	98.49%	100.00%
muy	99.50%	98.51%	100.00%	98.51%	100.00%	nnb	91.32%	91.32%	100.00%	91.32%	93.45%	rcf	99.50%	100.00%	100.00%	100.00%	100.00%
mwe	98.51%	100.00%	100.00%	100.00%	100.00%	nnh	99.50%	99.50%	100.00%	99.50%	100.00%	rel	98.51%	99.50%	100.00%	99.50%	100.00%
mwm	99.01%	100.00%	100.00%	100.00%	100.00%	nnq	99.50%	99.50%	100.00%	99.50%	100.00%	rif	100.00%	99.01%	100.00%	99.01%	100.00%
mwn	98.49%	98.02%	100.00%	98.02%	100.00%	nnw	100.00%	100.00%	100.00%	100.00%	100.00%	rim	100.00%	99.50%	100.00%	99.50%	100.00%
mws	98.49%	100.00%	100.00%	100.00%	100.00%	nqo	77.66%	79.57%	100.00%	79.57%	81.70%	rnd	100.00%	99.50%	100.00%	99.50%	100.00%
mxu	0.00%	0.00%	2.00%	0.00%	2.13%	nse	99.00%	96.15%	100.00%	96.15%	98.28%	rng	98.49%	99.00%	100.00%	99.00%	100.00%
myb	99.01%	99.50%	100.00%	99.50%	100.00%	nso	98.00%	97.51%	100.00%	97.51%	99.64%	rub	100.00%	100.00%	100.00%	100.00%	100.00%
myk	100.00%	100.00%	100.00%	100.00%	100.00%	ntr	100.00%	100.00%	100.00%	100.00%	100.00%	ruf	100.00%	100.00%	100.00%	100.00%	100.00%
myx	97.54%	97.54%	100.00%	97.54%	99.67%	nuj	96.45%	99.00%	100.00%	99.00%	100.00%	run	99.01%	99.50%	100.00%	99.50%	100.00%
mzk	100.00%	100.00%	100.00%	100.00%	100.00%	nup	96.77%	96.97%	16.00%	96.97%	99.10%	rwk	95.96%	95.96%	100.00%	95.96%	98.09%
mzm	100.00%	100.00%	100.00%	100.00%	100.00%	nus	99.50%	99.50%	100.00%	99.50%	100.00%	sag	99.50%	100.00%	100.00%	100.00%	100.00%
mzw	99.50%	100.00%	100.00%	100.00%	100.00%	nwb	100.00%	100.00%	100.00%	100.00%	100.00%	saq	98.99%	98.51%	100.00%	98.51%	100.00%
naq	96.62%	99.01%	100.00%	99.01%	100.00%	nwe	0.00%	0.00%	1.00%	0.00%	2.13%	say	99.39%	99.39%	82.00%	99.39%	100.00%
naw	99.50%	99.50%	100.00%	99.50%	100.00%	nxd	98.99%	100.00%	100.00%	100.00%	100.00%	sba	99.50%	100.00%	100.00%	100.00%	100.00%
nba	98.99%	99.50%	100.00%	99.50%	100.00%	nya	90.50%	93.46%	100.00%	93.46%	95.59%	sbd	99.50%	99.50%	100.00%	99.50%	100.00%
nbl	98.49%	98.49%	100.00%	98.49%	100.00%	nyb	99.50%	98.48%	100.00%	98.48%	100.00%	sbp	98.49%	99.50%	100.00%	99.50%	100.00%
ncu	100.00%	100.00%	100.00%	100.00%	100.00%	nyd	93.19%	96.52%	100.00%	96.52%	98.65%	sbs	98.99%	99.00%	100.00%	99.00%	100.00%
ndc	97.09%	98.04%	100.00%	98.04%	100.00%	nyf	95.83%	94.95%	100.00%	94.95%	97.08%	sby	98.99%	98.49%	100.00%	98.49%	100.00%
nde	100.00%	99.00%	100.00%	99.00%	100.00%	nyk	98.99%	99.00%	100.00%	99.00%	100.00%	sef	100.00%	100.00%	100.00%	100.00%	100.00%
ndh	97.46%	97.49%	100.00%	97.49%	99.62%	nym	97.51%	100.00%	100.00%	100.00%	100.00%	seh	96.08%	99.50%	100.00%	99.50%	100.00%
ndi	100.00%	99.50%	100.00%	99.50%	100.00%	nyn	95.19%	95.61%	100.00%	95.61%	97.74%	ses	98.51%	98.02%	100.00%	98.02%	100.00%
ndj	100.00%	100.00%	100.00%	100.00%	100.00%	nyo	97.46%	97.46%	100.00%	97.46%	99.59%	sev	100.00%	100.00%	100.00%	100.00%	100.00%
ndo	95.38%	96.91%	100.00%	96.91%	99.04%	nyu	97.96%	99.01%	100.00%	99.01%	100.00%	sfw	99.00%	100.00%	100.00%	100.00%	100.00%
ndp	100.00%	99.50%	100.00%	99.50%	100.00%	nyy	97.56%	98.52%	100.00%	98.52%	100.00%	sgc	98.00%	99.50%	100.00%	99.50%	100.00%
ndv	98.99%	98.99%	100.00%	98.99%	100.00%	nza	99.00%	98.02%	100.00%	98.02%	100.00%	sgw	100.00%	100.00%	100.00%	100.00%	100.00%
ndy	99.50%	100.00%	100.00%	100.00%	100.00%	nzi	100.00%	99.50%	100.00%	99.50%	100.00%	shi	99.00%	98.49%	100.00%	98.49%	100.00%
ndz	100.00%	100.00%	100.00%	100.00%	100.00%	odu	100.00%	100.00%	100.00%	100.00%	100.00%	shj	99.50%	100.00%	100.00%	100.00%	100.00%
neb	99.50%	100.00%	100.00%	100.00%	100.00%	ogo	99.50%	99.50%	100.00%	99.50%	100.00%	shk	100.00%	99.50%	100.00%	99.50%	100.00%
nfr	100.00%	100.00%	100.00%	100.00%	100.00%	oke	98.51%	99.50%	100.00%	99.50%	100.00%	shr	97.03%	98.99%	100.00%	98.99%	100.00%
ngb	99.50%	99.50%	100.00%	99.50%	100.00%	oki	82.35%	90.00%	10.00%	90.00%	92.13%	shu	99.50%	98.99%	100.00%	98.99%	100.00%
ngc	98.99%	98.48%	100.00%	98.48%	100.00%	okr	98.49%	99.50%	100.00%	99.50%	100.00%	sid	99.01%	100.00%	100.00%	100.00%	100.00%
nge	0.00%	0.00%	1.00%	0.00%	2.13%	oku	99.50%	99.01%	100.00%	99.01%	100.00%	sig	100.00%	100.00%	100.00%	100.00%	100.00%
ngl	95.65%	97.09%	100.00%	97.09%	99.22%	old	97.03%	100.00%	100.00%	100.00%	100.00%	sil	99.01%	98.52%	100.00%	98.52%	100.00%

Table C.4: Per-Language Results 4

ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid	ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid	ISO	Afrolid	Serengeti	Cheetah	FastText	Conlid
skg	93.90%	95.19%	100.00%	95.19%	97.32%	tir	95.05%	97.98%	100.00%	97.98%	100.00%	wec	94.47%	91.84%	100.00%	91.84%	93.97%
sld	99.50%	100.00%	100.00%	100.00%	100.00%	tiv	99.50%	99.50%	100.00%	99.50%	100.00%	wes	93.90%	96.15%	100.00%	96.15%	98.28%
sna	98.04%	98.52%	100.00%	98.52%	100.00%	tjo	40.00%	85.71%	4.00%	85.71%	87.84%	wib	100.00%	99.50%	100.00%	99.50%	100.00%
snf	100.00%	100.00%	100.00%	100.00%	100.00%	tke	99.50%	100.00%	100.00%	100.00%	100.00%	wlx	99.50%	99.50%	100.00%	99.50%	100.00%
sng	100.00%	99.50%	100.00%	99.50%	100.00%	tlj	100.00%	100.00%	100.00%	100.00%	100.00%	wmw	99.00%	98.51%	100.00%	98.51%	100.00%
snk	81.82%	96.00%	13.00%	96.00%	98.13%	tll	96.62%	98.52%	100.00%	98.52%	100.00%	wni	88.89%	75.00%	5.00%	75.00%	77.13%
snw	100.00%	99.00%	100.00%	99.00%	100.00%	tmc	100.00%	99.50%	100.00%	99.50%	100.00%	wob	100.00%	100.00%	100.00%	100.00%	100.00%
soe	96.41%	97.96%	100.00%	97.96%	100.00%	tnr	100.00%	99.50%	100.00%	99.50%	100.00%	wol	91.00%	94.79%	100.00%	94.79%	96.92%
som	100.00%	100.00%	100.00%	100.00%	100.00%	tod	99.00%	100.00%	100.00%	100.00%	100.00%	won	98.78%	98.18%	82.00%	98.18%	100.00%
sop	97.51%	97.56%	100.00%	97.56%	99.69%	tog	98.99%	98.99%	100.00%	98.99%	100.00%	wwa	98.02%	100.00%	100.00%	100.00%	100.00%
sor	99.50%	99.50%	100.00%	99.50%	100.00%	toh	100.00%	100.00%	100.00%	100.00%	100.00%	xan	100.00%	100.00%	100.00%	100.00%	100.00%
sot	99.50%	99.50%	100.00%	99.50%	100.00%	toi	98.02%	96.62%	100.00%	96.62%	98.75%	xed	99.50%	99.00%	100.00%	99.00%	100.00%
sox	0.00%	0.00%	1.00%	0.00%	2.13%	tpm	98.02%	99.00%	100.00%	99.00%	100.00%	xho	93.94%	94.85%	100.00%	94.85%	96.98%
soy	100.00%	99.50%	100.00%	99.50%	100.00%	tsb	0.00%	0.00%	2.00%	0.00%	2.13%	xkg	80.00%	100.00%	5.00%	100.00%	100.00%
spp	100.00%	99.50%	100.00%	99.50%	100.00%	tsc	99.50%	99.00%	100.00%	99.00%	100.00%	xmd	0.00%	0.00%	2.00%	0.00%	2.13%
spy	99.50%	100.00%	100.00%	100.00%	100.00%	tsn	98.02%	98.00%	100.00%	98.00%	100.00%	xmg	0.00%	0.00%	2.00%	0.00%	2.13%
srr	99.01%	96.52%	100.00%	96.52%	98.65%	tso	97.46%	97.49%	100.00%	97.49%	99.62%	xmv	99.50%	100.00%	100.00%	100.00%	100.00%
ssc	92.00%	98.04%	25.00%	98.04%	100.00%	tsw	99.01%	99.01%	100.00%	99.01%	100.00%	xnz	98.99%	100.00%	100.00%	100.00%	100.00%
ssn	0.00%	0.00%	2.00%	0.00%	2.13%	ttj	96.48%	98.02%	100.00%	98.02%	100.00%	xog	93.33%	97.98%	100.00%	97.98%	100.00%
ssw	100.00%	100.00%	100.00%	100.00%	100.00%	ttq	95.15%	95.61%	100.00%	95.61%	97.74%	xon	100.00%	99.50%	100.00%	99.50%	100.00%
stv	93.33%	100.00%	8.00%	100.00%	100.00%	ttr	98.99%	98.99%	100.00%	98.99%	100.00%	xpe	75.11%	75.83%	100.00%	75.83%	77.96%
suk	97.46%	99.00%	100.00%	99.00%	100.00%	tui	100.00%	100.00%	100.00%	100.00%	100.00%	xrb	100.00%	100.00%	100.00%	100.00%	100.00%
sur	100.00%	100.00%	100.00%	100.00%	100.00%	tul	99.50%	98.51%	100.00%	98.51%	100.00%	xsm	98.52%	99.50%	100.00%	99.50%	100.00%
sus	99.50%	100.00%	100.00%	100.00%	100.00%	tum	97.09%	99.01%	100.00%	99.01%	100.00%	xtc	99.50%	100.00%	100.00%	100.00%	100.00%
swa	72.51%	83.33%	100.00%	83.33%	85.46%	tuv	92.23%	97.03%	100.00%	97.03%	99.16%	xuo	100.00%	100.00%	100.00%	100.00%	100.00%
swb	97.49%	97.00%	100.00%	97.00%	99.13%	tuz	85.71%	100.00%	4.00%	100.00%	100.00%	yal	99.50%	100.00%	100.00%	100.00%	100.00%
swc	82.30%	84.39%	100.00%	84.39%	86.52%	tvs	66.67%	0.00%	2.00%	0.00%	2.13%	yam	100.00%	100.00%	100.00%	100.00%	100.00%
swh	82.00%	87.18%	100.00%	87.18%	89.31%	tvu	99.00%	100.00%	100.00%	100.00%	100.00%	yao	98.52%	99.01%	100.00%	99.01%	100.00%
swk	98.00%	99.00%	100.00%	99.00%	100.00%	twi	87.72%	90.50%	100.00%	90.50%	92.63%	yas	97.06%	98.52%	100.00%	98.52%	100.00%
sxb	100.00%	100.00%	100.00%	100.00%	100.00%	twx	96.37%	96.91%	100.00%	96.91%	99.04%	yat	99.00%	99.50%	100.00%	99.50%	100.00%
tap	94.36%	95.92%	100.00%	95.92%	98.05%	tzm	70.47%	76.24%	100.00%	76.24%	78.37%	yav	0.00%	0.00%	2.00%	0.00%	2.13%
taq	92.00%	92.46%	100.00%	92.46%	94.59%	udu	100.00%	100.00%	100.00%	100.00%	100.00%	yaz	100.00%	100.00%	100.00%	100.00%	100.00%
tbz	100.00%	100.00%	100.00%	100.00%	100.00%	umb	99.50%	100.00%	100.00%	100.00%	100.00%	yba	98.99%	98.99%	100.00%	98.99%	100.00%
tcc	99.50%	100.00%	100.00%	100.00%	100.00%	urh	99.50%	100.00%	100.00%	100.00%	100.00%	ybb	95.38%	93.94%	100.00%	93.94%	96.07%
tcd	97.98%	97.03%	100.00%	97.03%	99.16%	uth	100.00%	100.00%	100.00%	100.00%	100.00%	yom	97.06%	98.52%	100.00%	98.52%	100.00%
tdx	92.47%	94.79%	100.00%	94.79%	96.92%	vag	100.00%	100.00%	100.00%	100.00%	100.00%	yor	97.54%	98.52%	100.00%	98.52%	100.00%
ted	98.52%	98.51%	100.00%	98.51%	100.00%	vai	92.47%	92.47%	100.00%	92.47%	94.60%	yre	100.00%	100.00%	100.00%	100.00%	100.00%
tem	99.50%	100.00%	100.00%	100.00%	100.00%	ven	99.50%	100.00%	100.00%	100.00%	100.00%	zaj	88.52%	91.98%	100.00%	91.98%	94.11%
teo	98.00%	98.04%	100.00%	98.04%	100.00%	vid	98.51%	100.00%	100.00%	100.00%	100.00%	zdj	98.49%	98.51%	100.00%	98.51%	100.00%
tex	99.50%	100.00%	100.00%	100.00%	100.00%	vif	100.00%	100.00%	100.00%	100.00%	100.00%	zga	99.00%	98.99%	100.00%	98.99%	100.00%
tgw	99.50%	100.00%	100.00%	100.00%	100.00%	vmk	96.94%	96.97%	100.00%	96.97%	99.10%	zgh	71.13%	78.64%	100.00%	78.64%	80.77%
thk	99.00%	100.00%	100.00%	100.00%	100.00%	vmw	96.55%	97.51%	100.00%	97.51%	99.64%	ziw	97.49%	98.49%	100.00%	98.49%	100.00%
thv	90.43%	93.12%	100.00%	93.12%	95.25%	vun	99.50%	100.00%	100.00%	100.00%	100.00%	zne	99.01%	99.50%	100.00%	99.50%	100.00%
thy	0.00%	0.00%	1.00%	0.00%	2.13%	vut	100.00%	100.00%	100.00%	100.00%	100.00%	zul	96.59%	96.62%	100.00%	96.62%	98.75%
tig	94.42%	98.00%	100.00%	98.00%	100.00%	wal	99.01%	99.50%	100.00%	99.50%	100.00%						
tik	100.00%	100.00%	100.00%	100.00%	100.00%	wbi	98.48%	97.96%	100.00%	97.96%	100.00%						

Table C.5: Per-Language Results 5

Group	**Language**	**Baseline F_1**	**F_1_0.75**	**Δ_0.75**	**F_1_0.8**	**Δ_0.8**	**F_1_0.85**	**Δ_0.85**	**F1_0.9**	**Δ_0.9**	**F_1_0.95**	**Δ_0.95**
ara	gkn	99.00	99.50	+0.50	99.50	+0.50	99.50	+0.50	**100.00**	**+1.00**	**100.00**	**+1.00**
byf	byf	84.38	85.00	+0.62	85.00	+0.62	85.57	+1.20	85.15	+0.77	**85.71**	**+1.34**
	byv	91.24	92.52	+1.28	92.52	+1.28	92.09	+0.85	**92.96**	**+1.71**	92.09	+0.85
	ewo	91.18	**91.63**	**+0.45**	91.18	+0.00	91.18	+0.00	90.73	−0.45	90.20	−0.98
cop	cop	78.05	**86.96**	**+8.91**	**86.96**	**+8.91**	86.41	+8.36	86.41	+8.36	86.41	+8.36
ful	fub	90.38	91.79	+1.40	**92.23**	**+1.85**	**92.23**	**+1.85**	91.26	+0.88	**92.23**	**+1.85**
swa	swh	87.18	87.76	+0.58	87.76	+0.58	**88.32**	**+1.15**	**88.32**	**+1.15**	87.31	+0.13
gwr	gwl	32.79	74.18	+41.39	**76.50**	**+43.71**	**76.50**	**+43.71**	**76.50**	**+43.71**	**76.50**	**+43.71**
	gwr	70.50	70.59	+0.08	**72.13**	**+1.63**	**72.13**	**+1.63**	**72.13**	**+1.63**	70.97	+0.46
kau	kau	**87.05**	87.44	+0.39	86.29	−0.76	86.29	−0.76	84.69	−2.36	83.08	−3.97
	knc	**94.74**	95.15	+0.41	94.23	−0.51	94.23	−0.51	92.89	−1.85	91.51	−3.23
	lwg	**92.31**	92.45	+0.15	92.45	+0.15	91.59	−0.72	90.74	−1.57	90.74	−1.57
kma	kma	80.16	80.18	+0.01	**80.53**	**+0.37**	**80.53**	**+0.37**	**80.53**	**+0.37**	**80.53**	**+0.37**
	kmy	67.11	73.99	+6.88	**74.71**	**+7.61**	**74.71**	**+7.61**	**74.71**	**+7.61**	**74.71**	**+7.61**
	gur	99.50	**100.00**	**+0.50**	**100.00**	**+0.50**	**100.00**	**+0.50**	**100.00**	**+0.50**	**100.00**	**+0.50**
kon	kng	93.90	95.65	+1.76	96.12	+2.22	96.08	+2.18	**98.00**	**+4.10**	97.49	+3.59
	kon	80.00	84.44	+4.44	85.08	+5.08	86.49	+6.49	**88.89**	**+8.89**	88.42	+8.42
	ktu	**93.66**	94.58	+0.92	94.58	+0.92	94.58	+0.92	94.12	+0.46	**93.66**	**+0.00**
	kwy	96.15	96.62	+0.46	96.62	+0.46	**97.09**	**+0.93**	**97.09**	**+0.93**	**97.09**	**+0.93**
kzn	kzn	86.36	89.50	+3.14	90.11	+3.75	90.11	+3.75	**91.30**	**+4.94**	90.71	+4.35
	nse	96.15	**97.56**	**+1.41**	**97.56**	**+1.41**	**97.56**	**+1.41**	**97.56**	**+1.41**	**97.56**	**+1.41**
	nya	**93.46**	92.96	−0.50	92.96	−0.50	92.96	−0.50	93.84	+0.38	93.40	−0.06
ngn	bas	93.46	94.34	+0.88	93.84	+0.38	94.29	+0.83	94.29	+0.83	**94.74**	**+1.28**
	bsq	81.00	**82.54**	**+1.54**	81.48	+0.48	81.48	+0.48	81.48	+0.48	82.11	+1.11
	ngn	85.11	**88.56**	**+3.45**	88.12	+3.01	87.68	+2.58	87.25	+2.15	87.25	+2.15
nqo	nqo	75.98	75.86	−0.12	76.57	+0.59	77.27	+1.30	76.40	+0.43	**78.21**	**+2.23**
xpe	kea	**96.15**	96.62	+0.46	**96.15**	**+0.00**	**96.15**	**+0.00**	**96.15**	**+0.00**	**96.15**	**+0.00**
	xpe	75.83	79.66	+3.83	**80.17**	**+4.34**	**80.17**	**+4.34**	79.83	+4.00	79.83	+4.00
Average		–	–	+3.30	–	+3.93	–	+4.22	–	+4.20	–	**+4.55**

Table D.1: Hierarchical classification results using *Mirror-Serengeti* embeddings across confidence thresholds (75%, 80%, 85%, 90%, 95%). Baseline F_1 shows base Serengeti performance; columns show improvement over baseline. Bold indicates best performance per language.

argument	**description**	**value**
-epoch	number of training epochs	1
-train_batch_size	training batch size	200
-learning_rate	learning rate	2e-5
-max_length	max sequence length	50
-infoNCE_tau	InfoNCE temperature (τ)	0.04
-dropout_rate	dropout rate	0.0
-drophead_rate	drophead rate	0.05
-random_span_mask	length of random span mask	5
-agg_mode	aggregation mode	cls

Table D.2: Training hyperparameters for the Mirror-BERT model.