1 Gherbal v4 – Language Identification at Scale

A 200 MB FastText model that outperforms competitors 6–8x its size on 214 languages

Omar Kamali, Omneity Labs — March 2026

1.1 The Challenge

Language identification (LID) is the first step in any multilingual NLP pipeline, yet state-of-the-art models struggle with Arabic dialects, low-resource African languages, and non-standard text like Arabizi (Latin-script Arabic). Most LID models exceed 1 GB and still fail on these critical languages.

1.2 Key Results

We evaluated **10 LID models** across **8 benchmarks** covering clean text, web crawls, Arabic dialects, and African languages:

Model	Size	Languages	Avg Accuracy (V4 scope)
Gherbal v4	**200 MB**	**214**	**0.836**
OpenLID v2	1,230 MB	201	0.824
GlotLID	1,690 MB	2,102	0.803
NLLB-LID	1,180 MB	218	0.711

Gherbal v4 achieves the highest average accuracy at one-sixth the size of its nearest competitors.

1.3 What Makes Gherbal Different

- **Arabic dialect coverage**: Only model to identify all 16 Arabic dialect variants. Competitors cover at most 8. NLLB-LID collapses all Arabic to MSA.
- **Arabizi support**: Only model that identifies Latin-script Moroccan Arabic (96.9% accuracy). All competitors score 0%.
- **North African specialization**: Unique coverage of Hassaniya, Libyan Arabic, and Berber / Amazigh languages.
- **African languages**: Best-in-class on Dyula, Kituba, Kamba, and Twi — languages where competitors score near zero.
- **Data efficiency**: Trained on <3 GB of curated data (vs 21 GB for OpenLID, 45 GB for GlotLID) using a 4-pass cleaning pipeline.

1.4 How It Works

Gherbal v4 uses the FastText supervised architecture with character n-grams (minn=2, maxn=5), which naturally captures the morphological patterns that distinguish Arabic dialects. The key innovation is a **4-pass training data cleaning pipeline**: script validation, cross-language deduplication, self-prediction disambiguation, and temperature resampling ($p^{0.3}$). This yields higher-quality training data that compensates for smaller model size.

1.5 Benchmarks Used

FLORES-200 (devtest + dev), MADAR (Arabic dialects), Atlasia-LID (Arabic dialects), CommonLID (web crawl), WiLI-2018 (Wikipedia), Bouquet (machine translation), and Gherbal-Multi (curated multi-domain).

1.6 Availability

The full evaluation report with 26 charts, per-language analysis, and methodology details is available at: *[link to report]*

Omneity Labs — Building language technology for underserved languages.