Gherbal v4

Comparative Evaluation Report

Comparative Analysis of 10 Language Identification Models
Across 8 Benchmarks and 5 Scoping Regimes

Omar Kamali
Omneity Labs
omar@omneitylabs.com

March 2026

Contents

1 Introduction: The Language Identification Problem **3**

2 Executive Summary **3**

3 Models Evaluated **4**

4 Benchmarks **4**

5 Evaluation Methodology **5**

5.1 Scoping . . . 5

5.2 Metrics . . . 5

6 Overall Results **5**

6.1 V4 Scope — Primary Comparison . . . 6

6.2 F1 Diagnostic (FLORES devtest, V4 Scope) . . . 11

7 North African Languages **12**

7.1 Language Coverage . . . 12

7.2 Results . . . 13

8 Arabic Dialect Identification **15**

8.1 Dialect Coverage . . . 15

8.2 MADAR Benchmark (13 dialects) . . . 16

8.3 Atlasia-LID Benchmark (15 dialects) . . . 16

8.4 Analysis . . . 17

8.5 Arabizi: The Unsupported Script . . . 20

9 Sub-Saharan African Languages **21**

9.1 FLORES devtest Results (V4 Scope) . . . 21

9.2 Gherbal v4 vs Competitors . . . 21

10 Model Efficiency Analysis **25**

10.1 Size vs Performance . . . 25

10.2 Training Data Efficiency . . . 26

10.3 Scope Scaling Behavior . . . 26

11 Script and Language Family Analysis **28**

11.1 By Script . . . 28

11.2 By Language Family . . . 28

12 Architecture & Training Details **30**
12.1 Gherbal Evolution 30
12.2 Training Pipeline (v4) 32
12.3 Why FastText? 33

13 Benchmark Difficulty Analysis **33**

14 Strengths and Weaknesses **34**
14.1 Gherbal v4 Strengths 35
14.2 Gherbal v4 Weaknesses 35

15 Conclusion **36**

16 References **37**
16.1 Models 37
16.2 Benchmarks and Datasets 37

17 Appendix: Chart Index **38**

1 Introduction: The Language Identification Problem

Language identification (LID) — determining which language a piece of text is written in — is a foundational task in natural language processing. It serves as the first stage in virtually every multilingual pipeline: machine translation, content moderation, search indexing, and document routing all require knowing the input language before any downstream processing can begin.

While LID is often considered a "solved" problem for well-resourced languages with distinct scripts, the reality at scale is far more challenging:

- **Short text sensitivity.** Social media posts, chat messages, and search queries are often just a few words long. With so little signal, even high-accuracy LID models degrade sharply — a 5-word sentence provides far fewer distinguishing n-grams than a full paragraph.
- **Script overlap.** Over 100 languages use the Latin script, and 25+ use the Arabic script. When the writing system itself is not a distinguishing feature, models must rely on subtle lexical and morphological cues.
- **Dialect continua.** Languages like Arabic, Chinese, and Malay exist on dialect continua where neighboring varieties share extensive vocabulary and syntax. Drawing a classification boundary between Moroccan and Algerian Arabic — or between Malay and Indonesian — is fundamentally ambiguous.
- **Code-switching and borrowing.** Multilingual speakers frequently mix languages within a single sentence. North African social media, for example, commonly blends Darija (Moroccan Arabic), French, and Arabizi in a single post.
- **Low-resource language scarcity.** Most LID models are trained on web-crawled data that is dominated by a handful of high-resource languages. Thousands of languages have little to no web presence, making it difficult to build or evaluate classifiers for them.
- **Noise and non-standard text.** Real-world text contains typos, transliterations, URLs, emoji, and formatting artifacts. Web-crawled training data inherits label noise from automatic annotation pipelines.

These challenges mean that state-of-the-art LID models — despite impressive headline numbers on clean benchmarks — can fail dramatically on real-world content, especially for under-resourced language families and non-standard text varieties. This report evaluates 10 LID models across 8 diverse benchmarks specifically designed to stress-test these failure modes.

---

2 Executive Summary

This report presents a comprehensive evaluation of **10 language identification (LID) models**, including four generations of Gherbal, across **8 diverse benchmarks** covering different domains, language scopes, and difficulty levels. All evaluations are performed under **5 scoping regimes** (V1 through V4 plus Full) to measure how models behave as the classification space expands.

Key findings:

- **Gherbal v4** (200 MB, FastText) achieves the **highest average accuracy** (0.836) across all 8 benchmarks at the V4 scope — outperforming models 6–8× its size.

- At the V4 scope, Gherbal v4 leads on 2 of 8 benchmarks (MADAR, Atlasia-LID) and is within 2% of the leader on both FLORES splits.
- Gherbal v4 is the **only model** that achieves non-zero accuracy on all 16 Arabic dialect variants across MADAR and Atlasia-LID. The closest competitor (OpenLID-v2) identifies only 7 out of 16.
- For **North African languages** (Berber / Amazigh + Maghreb Arabic including Hassaniya), Gherbal v4 provides significantly broader coverage and competitive accuracy compared to all other models.
- On **Sub-Saharan African languages**, Gherbal v4 performs comparably to larger models, with particular strength on low-resource languages like Dinka, Bambara, and Kituba.
- Gherbal v4 represents an 18× efficiency advantage over GlotLID at comparable accuracy (200 MB vs 1,690 MB at 83.6% vs 80.3% average accuracy).

---

3 Models Evaluated

Model	Architecture	Size (MB)	Languages	Source
Gherbal v1	FastText	30	36	Omneity Labs (Ours)
Gherbal v2	FastText	32	46	Omneity Labs (Ours)
Gherbal v3	FastText	78	106	Omneity Labs (Ours)
Gherbal v4	FastText	200	214	Omneity Labs (Ours)
NLLB-LID	FastText	1,180	218	Meta NLLB team
OpenLID v1	FastText	1,230	201	Burchell et al.
OpenLID v2	FastText	1,230	201	Burchell et al.
fastlid-176	FastText	131	176	Facebook Research
GlotLID	FastText	1,690	2,102	Kargaran et al.
OpenLID v3 (HPLT)	FastText	1,360	201	HPLT Consortium

All models use the FastText supervised architecture. Gherbal models are quantized; most competitors use full-precision embeddings, explaining the size difference.

---

4 Benchmarks

Benchmark	Domain	Languages	Samples	Characteristics
FLORES devtest	News / Wikipedia	214	216K+	Standard multilingual NLP benchmark (NLLB)
FLORES dev	News / Wikipedia	220	222K+	Development split (same domain)
MADAR	Social media / dialects	15	Arabic dialects only	13 Arabic dialect benchmarks
Atlasia-LID	Social / web / news	15	Arabic dialects only	Broader Arabic dialect mix

Benchmark	Domain	Languages	Samples	Characteristics
CommonLID	Web crawl (Common-Crawl)	101	Web text	Noisy, short, real-world web text
WiLI-2018	Wikipedia	124+	Paragraph-level	Clean, long-form text
Bouquet	Machine translation	275	Sentence-level	Translated sentences across 275 languages
Gherbal-Multi	Mixed sources	36	Multi-domain	Curated test set for core languages

The benchmarks span clean literary text (WiLI), machine-translated content (Bouquet), real-world web crawls (CommonLID), dialect-heavy social media (MADAR, Atlasia), and standard NLP benchmarks (FLORES). This diversity ensures that model rankings reflect genuine robustness rather than benchmark-specific overfitting.

---

5 Evaluation Methodology

5.1 Scoping

Evaluation is performed under 5 scoping regimes to control for the confounding effect of classification space size:

Scope	Languages	Description
V1	36	Original Gherbal v1 core languages
V2	46	V1 + 10 Arabic dialects
V3	106	V2 + 60 additional (Gherbal v3 scope)
V4	214	Full Gherbal v4 language set
Full	$\geq$214	All languages in each benchmark, no filtering

Scoping limits the evaluation to predictions within the scope's language set. This enables fair comparison: e.g., a model that only supports 46 languages is not penalized for failing to predict a 47th.

5.2 Metrics

- **Accuracy**: Proportion of correctly identified samples.
- **F1-macro**: Unweighted average of per-class F1 scores (measures fairness across all classes).
- **F1-weighted**: Per-class F1 weighted by support (measures overall quality).
- **Per-language accuracy**: Accuracy and top confusions for each language individually.

---

6 Overall Results

6.1 V4 Scope – Primary Comparison

Model	FLORES-DT	FLORES-D	MADAR	Atlasia	CommonLID	WiLI	Bouquet	Gherbal-M	**AVG**
Gherbal v4	0.921	0.928	**0.656**	**0.691**	0.828	0.915	0.878	0.870	**0.836**
OpenLID v2	**0.939**	**0.939**	0.652	0.574	0.853	0.937	**0.926**	0.776	0.824
OpenLID v1	0.909	0.908	0.582	0.485	0.872	**0.943**	0.910	0.830	0.805
GlotLID	0.940	0.948	0.588	0.498	0.819	0.940	0.912	0.777	0.803
NLLB-LID	0.889	0.886	0.110	0.335	**0.894**	0.927	0.893	0.752	0.711
OpenLID v3	0.919	0.918	0.000	0.000	0.691	0.919	0.915	0.662	0.628
Gherbal v3	0.391	0.399	0.598	0.656	0.772	0.470	0.363	0.897	0.568
fastlid-176	0.434	0.431	0.141	0.390	0.799	0.715	0.517	0.647	0.509
Gherbal v2	0.162	0.164	0.605	0.656	0.616	0.237	0.187	0.796	0.428
Gherbal v1	0.149	0.150	0.289	0.272	0.630	0.225	0.185	0.839	0.342

Key observations:

1. **Gherbal v4 leads on average** (0.836) despite being 6–8× smaller than the next best models (OpenLID-v2 at 0.824 with 1,230 MB, GlotLID at 0.803 with 1,690 MB).
2. **Arabic benchmarks (MADAR, Atlasia) are decisive:** Gherbal v4 leads MADAR and Atlasia by substantial margins. Most competitors struggle with Arabic dialects — NLLB-LID scores only 0.110 on MADAR, and OpenLID-v3 scores 0.000 on both.
3. **GlotLID and OpenLID-v2** are the strongest competitors on general benchmarks (FLORES, WiLI), but their Arabic dialect coverage is much weaker.
4. **NLLB-LID** excels on CommonLID (0.894, best) but has near-zero Arabic dialect performance.
5. **Gherbal v1/v2/v3** are not competitive at V4 scope because they were designed for only 36, 46, and 106 languages respectively — they cannot identify the 214-language scope.

Figure 1 visualizes the full accuracy matrix across all model–benchmark combinations. The heatmap makes clear that Gherbal v4 is the only model with strong performance across *both* general and Arabic-specific benchmarks — the warm column on the left contrasts with the cold blues that dominate the Arabic rows for competitors.

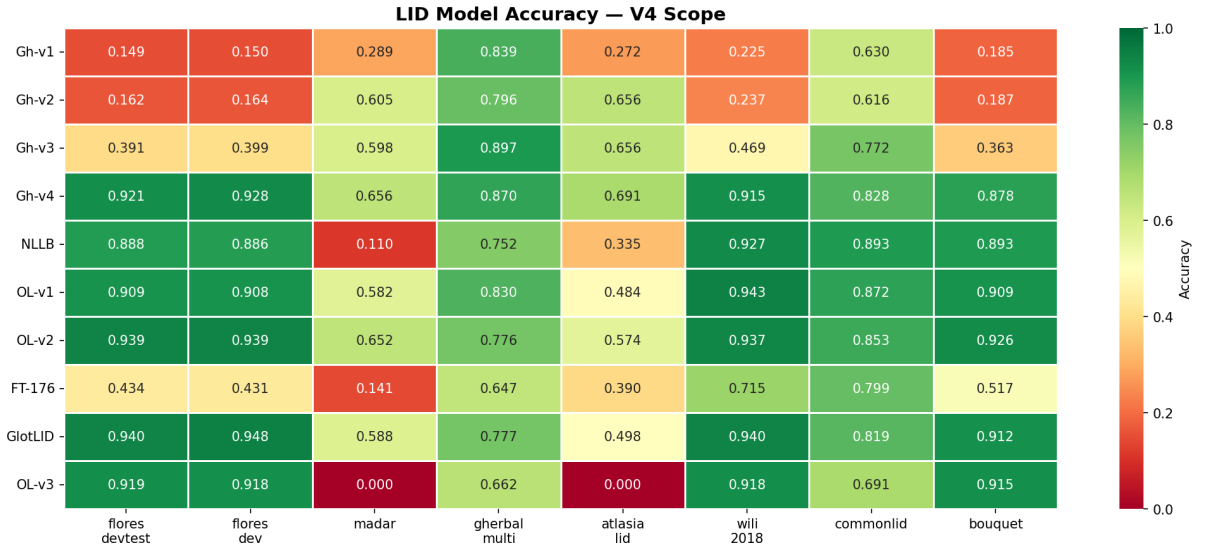


Figure 1: Accuracy heatmap – all models × benchmarks (V4 scope)

Breaking this down per benchmark (Figure 2), the pattern becomes starker. Gherbal v4's bars are consistently near the top on every benchmark, while competitors show pronounced weaknesses — NLLB-LID's near-zero bars on MADAR and Atlasia are particularly striking.

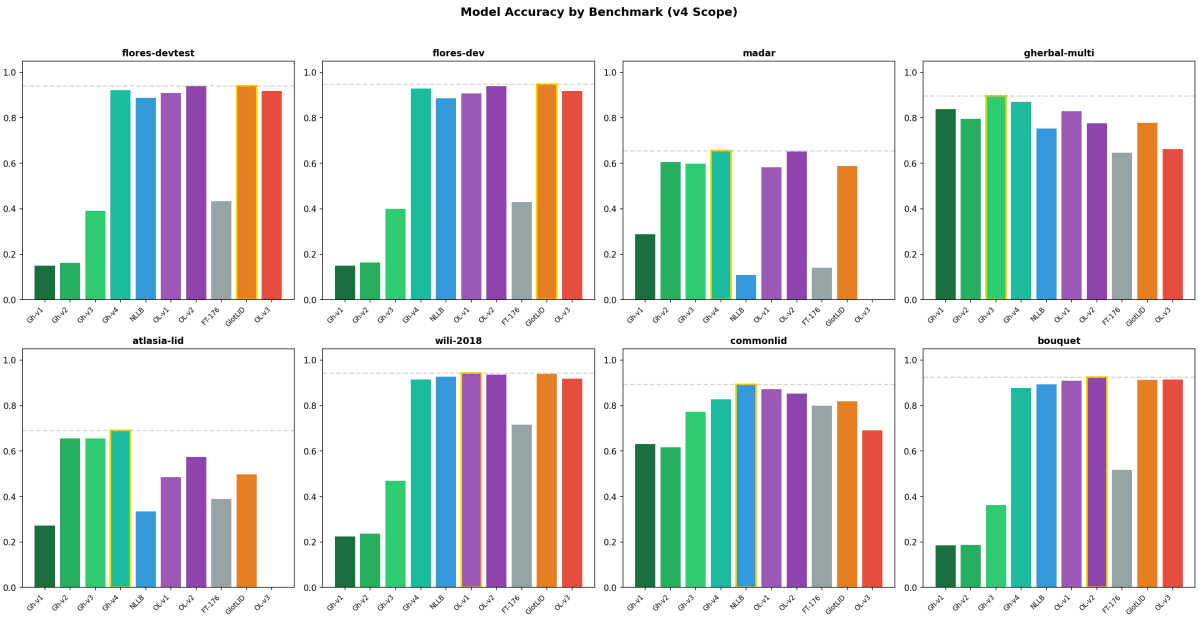


Figure 2: Per-benchmark model comparison (V4 scope)

The radar plot (Figure 3) synthesizes these results into model "profiles." Gherbal v4's polygon is the most uniformly expanded, while competitors show characteristic dents — GlotLID dips on Arabic benchmarks, NLLB-LID collapses on Atlasia, and fastlid-176 is visibly smaller overall.

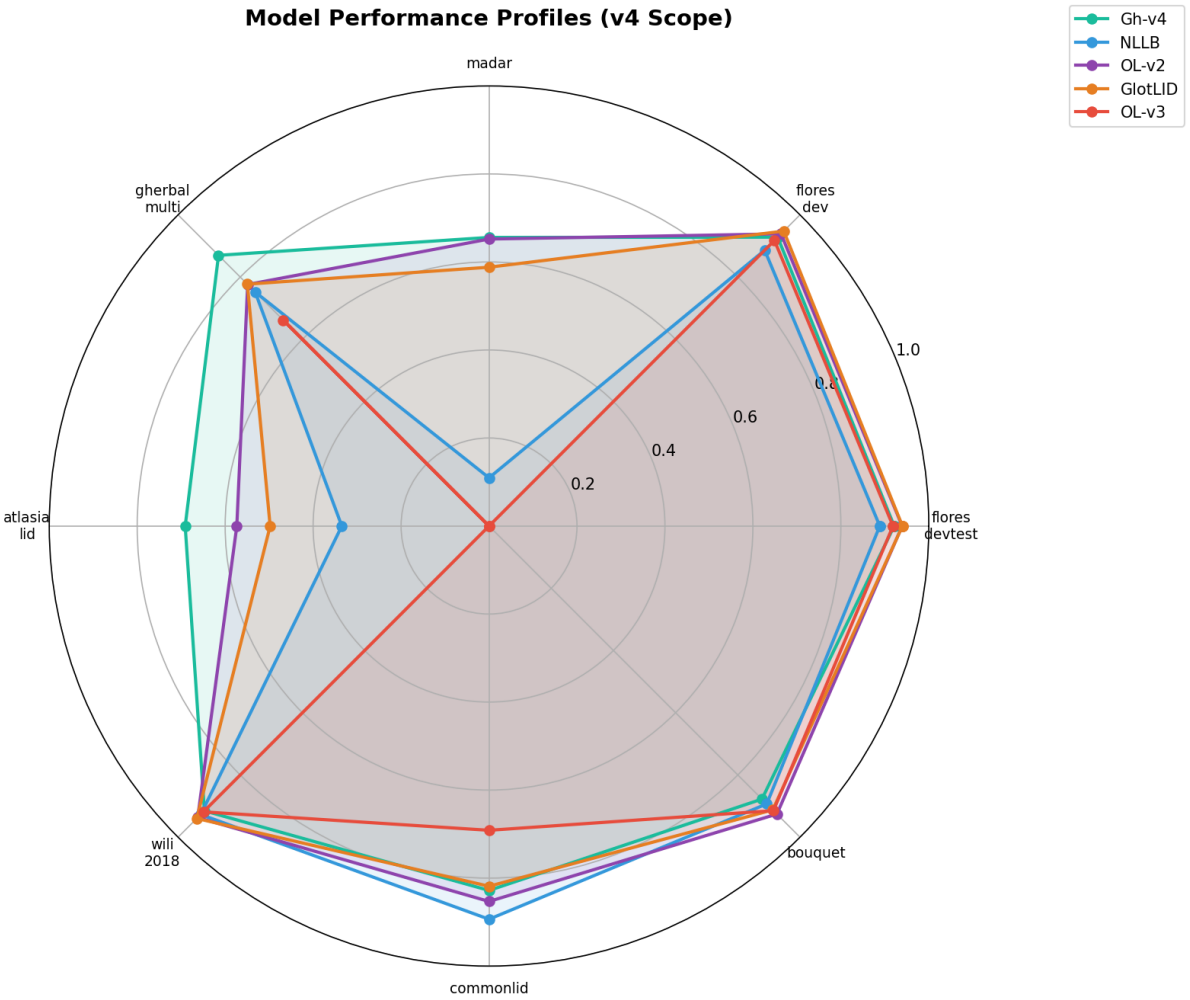
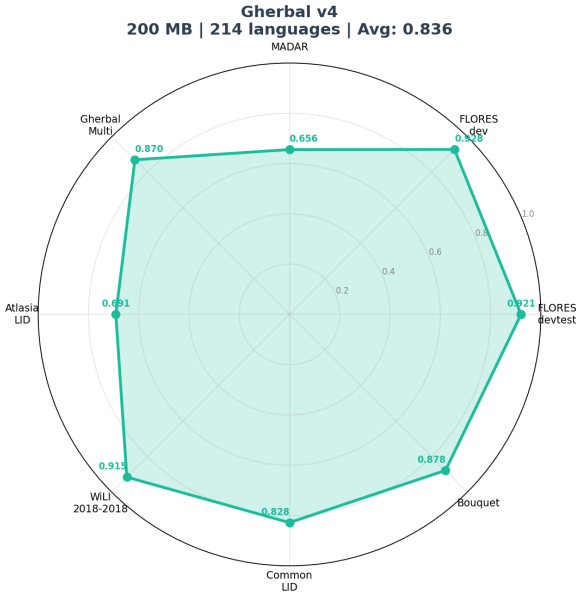
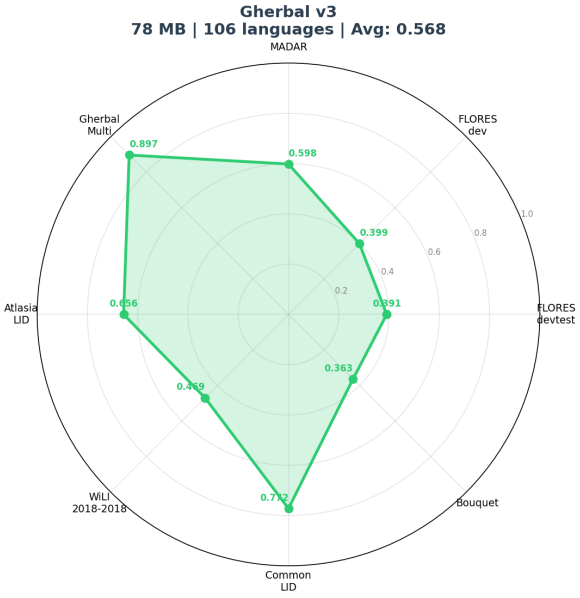
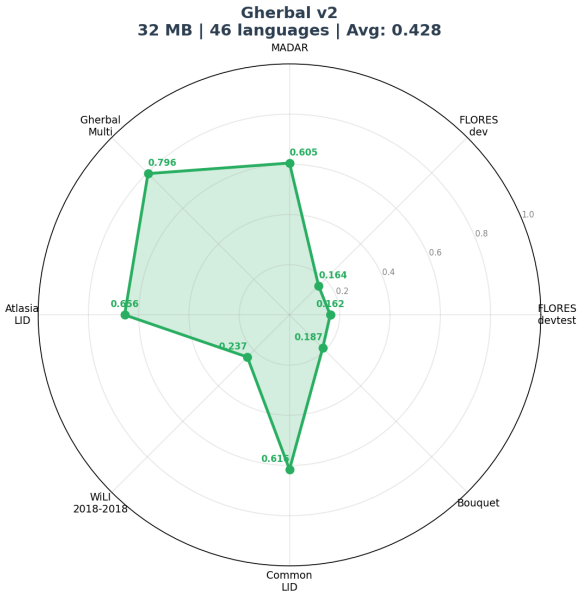


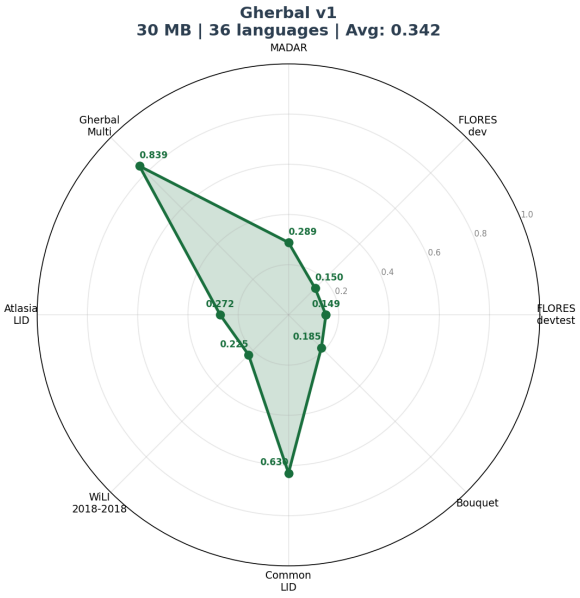
Figure 3: Model performance profiles across benchmarks

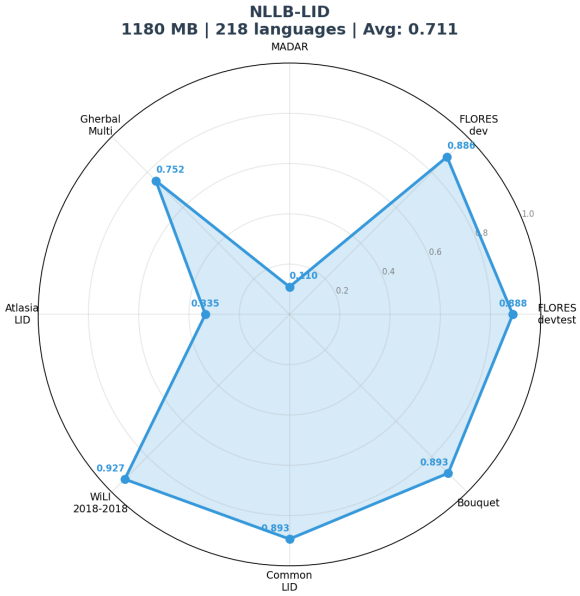
Individual model profiles — each radar below shows a single model's benchmark coverage with accuracy scores, model size, and language count:

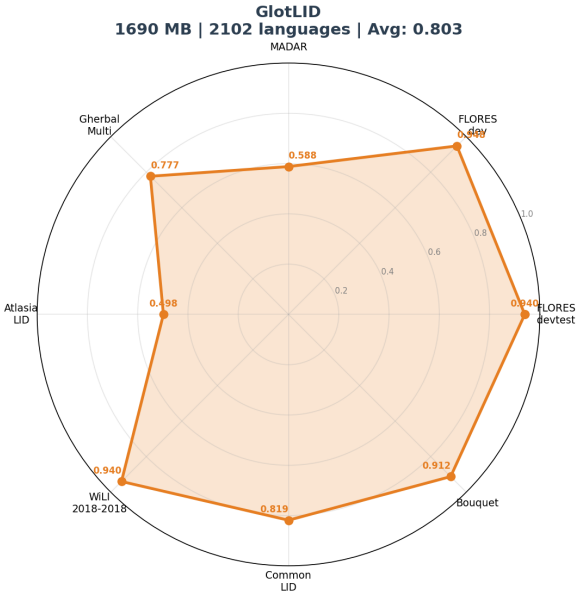

Gherbal v4
200 MB | 214 languages | Avg: 0.836
MADAR
FLORES dev
FLORES devtest
Bouquet
Common LID
WiLI 2018-2018
Atlasia LID
Gherbal Multi
0.656
0.928
0.921
0.878
0.828
0.915
0.691
0.870

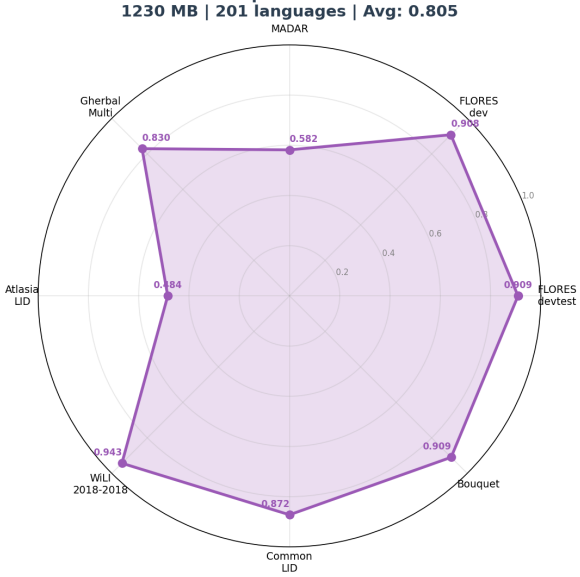

Gherbal v3
78 MB | 106 languages | Avg: 0.568
MADAR
FLORES dev
FLORES devtest
Bouquet
Common LID
WiLI 2018-2018
Atlasia LID
Gherbal Multi
0.598
0.399
0.391
0.363
0.772
0.469
0.656
0.897

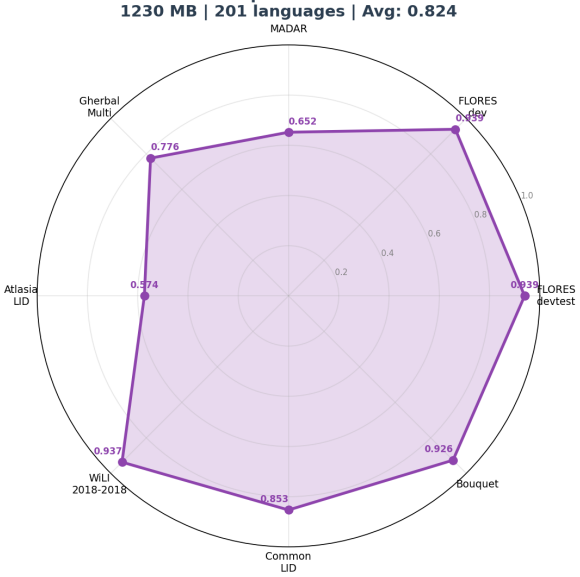

Gherbal v2
32 MB | 46 languages | Avg: 0.428
MADAR
FLORES dev
FLORES devtest
Bouquet
Common LID
WiLI 2018-2018
Atlasia LID
Gherbal Multi
0.605
0.164
0.162
0.187
0.615
0.237
0.656
0.796

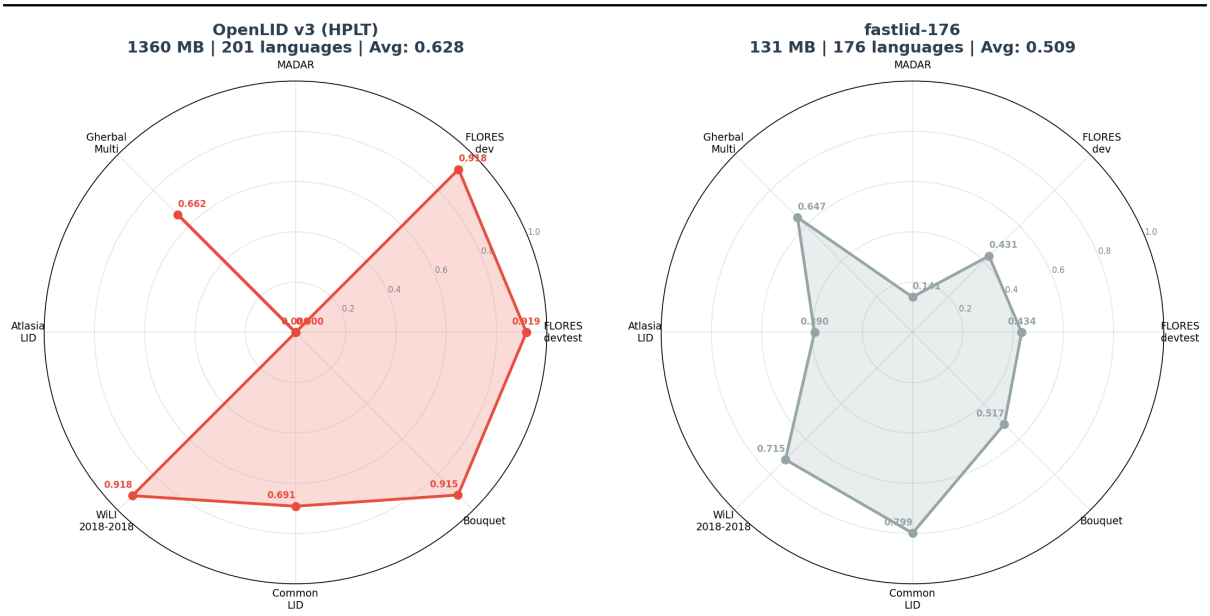

Gherbal v1
30 MB | 36 languages | Avg: 0.342
MADAR
FLORES dev
FLORES devtest
Bouquet
Common LID
WiLI 2018-2018
Atlasia LID
Gherbal Multi
0.289
0.150
0.149
0.185
0.638
0.225
0.272
0.839


NLLB-LID
1180 MB | 218 languages | Avg: 0.711
MADAR
FLORES dev
FLORES devtest
Bouquet
Common LID
WiLI 2018-2018
Atlasia LID
Gherbal Multi
0.752
0.886
0.888
0.893
0.893
0.927
0.835
0.710


GlotLID
1690 MB | 2102 languages | Avg: 0.803
MADAR
FLORES dev
FLORES devtest
Bouquet
Common LID
WiLI 2018-2018
Atlasia LID
Gherbal Multi
0.777
0.588
0.940
0.912
0.819
0.940
0.498


OpenLID v1
1230 MB | 201 languages | Avg: 0.805
MADAR
FLORES dev
FLORES devtest
Bouquet
Common LID
WiLI 2018-2018
Atlasia LID
Gherbal Multi
0.830
0.582
0.909
0.909
0.872
0.943
0.484


OpenLID v2
1230 MB | 201 languages | Avg: 0.824
MADAR
FLORES dev
FLORES devtest
Bouquet
Common LID
WiLI 2018-2018
Atlasia LID
Gherbal Multi
0.776
0.652
0.939
0.926
0.853
0.937
0.574



6.2 F1 Diagnostic (FLORES devtest, V4 Scope)

Model	F1-macro	F1-weighted	Gap
OpenLID v2	**0.928**	0.935	0.008
OpenLID v3	0.890	0.918	0.028
OpenLID v1	0.871	0.905	0.033
Gherbal v4	0.851	0.919	0.068
NLLB-LID	0.814	0.885	0.070
GlotLID	0.572	0.941	**0.368**

The F1-macro vs F1-weighted gap reveals how uniformly a model performs across languages. GlotLID's extreme gap (0.368) indicates it excels on high-resource languages but fails dramatically on many low-resource ones — inflated by its 2,102-class output space introducing systematic confusion. Gherbal v4's moderate gap (0.068) indicates generally balanced performance.

Figure 4 plots this diagnostic visually. Models closer to the diagonal treat all languages fairly; those far below it (like GlotLID) perform well on aggregate but leave many languages behind. Gherbal v4 sits in a strong middle position — high weighted F1 with a reasonably tight gap.

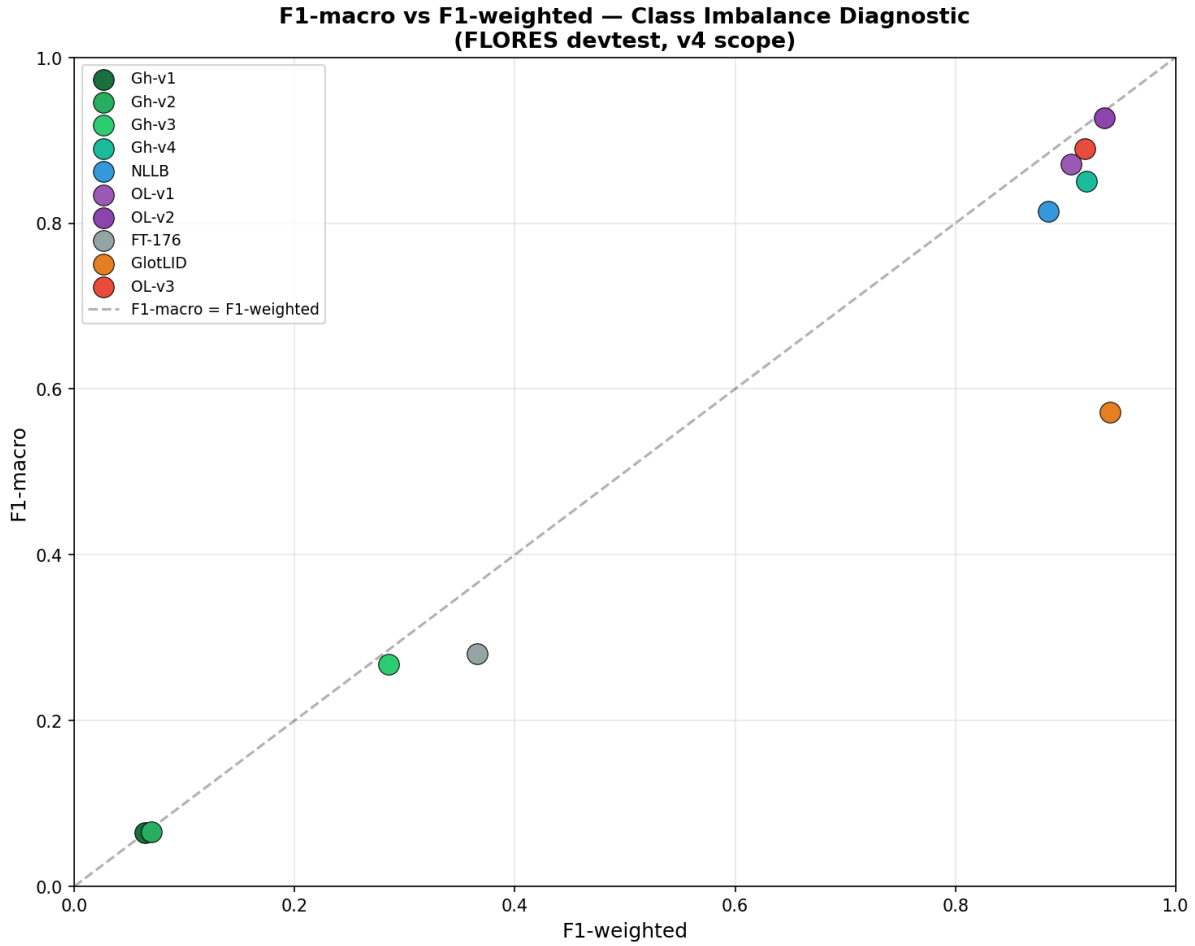


Figure 4: F1-macro vs F1-weighted diagnostic scatter plot

7 North African Languages

Gherbal was designed with a special focus on North African languages — both Arabic dialects of the Maghreb and Berber/Amazigh languages. This section evaluates 13 North African languages across Gherbal's specialization areas.

7.1 Language Coverage

Language	Code	Script	Family
Moroccan Arabic	`ary_Arab`	Arabic	Maghrebi Arabic
Moroccan Arabic (Arabizi)	`ary_Latn`	Latin	Maghrebi Arabic
Hassaniya	`mey_Arab`	Arabic	Hassaniya Arabic
Algerian Arabic	`arq_Arab`	Arabic	Maghrebi Arabic
Tunisian Arabic	`aeb_Arab`	Arabic	Maghrebi Arabic
Libyan Arabic	`ayl_Arab`	Arabic	Maghrebi Arabic
Egyptian Arabic	`arz_Arab`	Arabic	Nile Valley Arabic

Language	Code	Script	Family
Central Atlas Tamazight	`tzm_Latn`	Latin	Berber
Riffian	`rif_Latn`	Latin	Berber
Shilha	`shi_Latn`	Latin	Berber
Standard Moroccan Tamazight	`zgh_Tfng`	Tifinagh	Berber
Kabyle	`kab_Latn`	Latin	Berber
Tamasheq (Tifinagh)	`taq_Tfng`	Tifinagh	Berber
Tamasheq (Latin)	`taq_Latn`	Latin	Berber

7.2 Results

FLORES devtest (full scope):

Language	Gherbal v4	OpenLID v2	GlotLID	NLLB-LID
Moroccan Arabic (Ar)	0.821	0.893	0.817	0.000
Tunisian Arabic	0.558	0.797	0.609	0.000
Egyptian Arabic	0.611	0.752	0.596	0.000
Kabyle	0.990	0.992	0.989	0.997
Zgh Tifinagh	0.999	0.999	0.999	0.999
Tamasheq (Tifinagh)	0.996	0.987	0.990	0.992
Tamasheq (Latin)	0.998	0.999	0.998	0.998

MADAR (full scope, Maghreb dialects):

Dialect	Gherbal v4	OpenLID v2	GlotLID	NLLB-LID
Moroccan	0.925	**0.955**	0.890	0.000
Algerian	0.288	0.000	**0.315**	0.000
Tunisian	0.823	**0.938**	0.821	0.000
Libyan	**0.243**	0.000	0.000	0.000
Egyptian	0.746	**0.878**	0.798	0.000

Key findings:

- **Coverage advantage:** Gherbal v4 is the only model that can identify Libyan Arabic (`ayl_Arab`), Hassaniya (`mey_Arab`), and most other rare dialects. For Algerian Arabic (`arq_Arab`), only Gherbal v4 and GlotLID achieve non-zero accuracy.
- **Berber languages** are well-served by all models that support them, with accuracies above 0.98 across the board. These are linguistically distinct (Tifinagh script, Latin orthography) and therefore easier to distinguish.
- **Maghrebi Arabic** is the differentiator: while Moroccan Arabic and Tunisian are handled by multiple models (OpenLID-v2 leads), Gherbal v4 uniquely covers the full Maghreb spectrum including Algerian and Libyan.
- NLLB-LID has zero performance on all Arabic dialects except MSA — it maps everything to `arb_Arab`.

The heatmap below (Figure 5) summarizes the coverage gap across North African languages. The bright cells for Gherbal v4 on Algerian and Libyan Arabic — where every other model is dark blue — illustrate the unique coverage advantage.

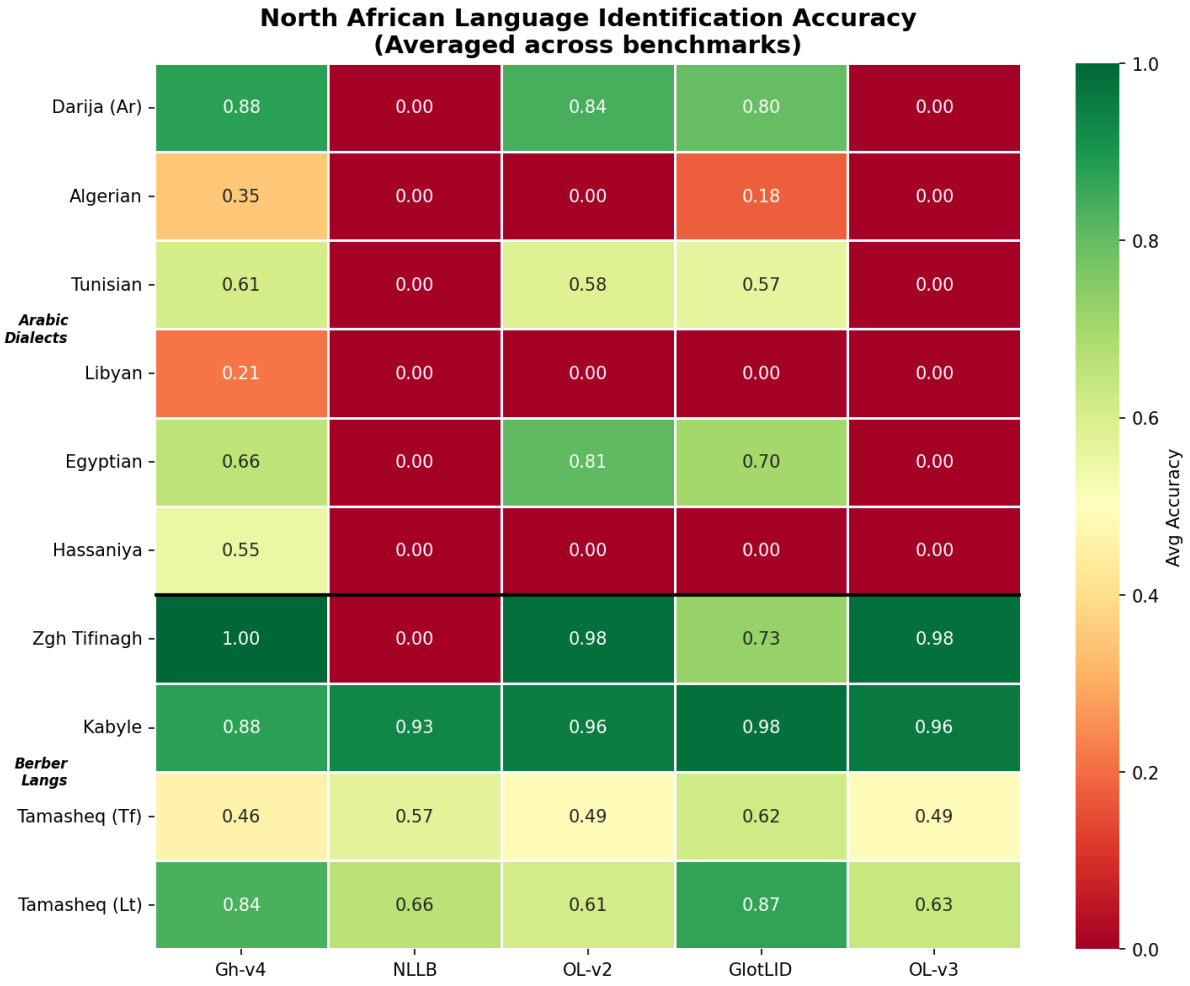


Figure 5: North African language accuracy heatmap

Figure 6 breaks this down by benchmark. The three panels (FLORES, MADAR, Atlasia) show that Gherbal v4's North African advantage is consistent across evaluation contexts, not an artifact of a single benchmark.

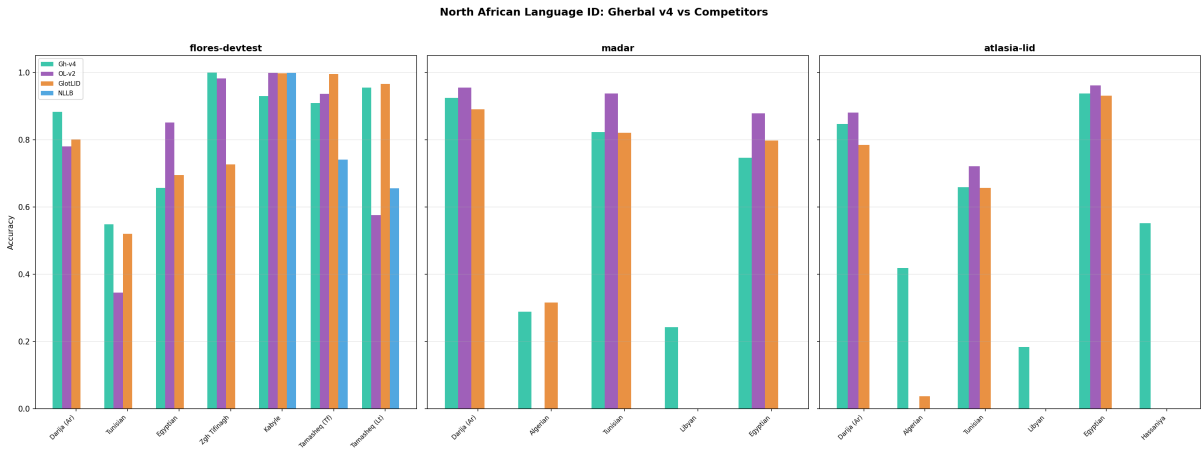


Figure 6: North African languages across FLORES, MADAR, and Atlasia

For Berber/Amazigh languages specifically (Figure 7), all models perform well on languages written in distinct scripts (Tifinagh, Latin). The real differentiation is on the Arabic-script Maghrebi varieties, where Gherbal v4 pulls ahead.

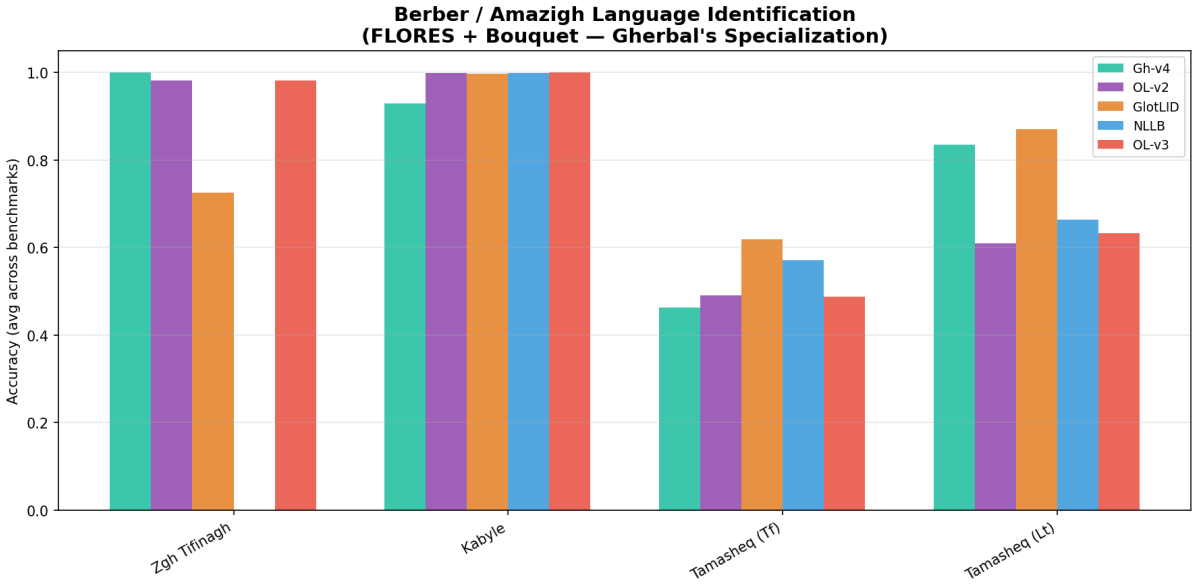


Figure 7: Berber/Amazigh language identification focus

---

8 Arabic Dialect Identification

Arabic dialect identification is one of the hardest problems in LID. Modern Standard Arabic (MSA, `arb_Arab`) is linguistically distinct from spoken dialects, but dialects share heavy lexical and morphological overlap. This section evaluates all 16 Arabic dialect variants.

8.1 Dialect Coverage

The 16 Arabic dialects evaluated span 5 geographic regions:

Region	Dialects
Maghreb	Moroccan (ary), Hassaniya (mey), Algerian (arq), Tunisian (aeb), Libyan (ayl)
Nile Valley	Egyptian (arz), Sudanese (apd)
Levant	Levantine (apc), Ta'izzi-Adeni (acq)
Gulf	Gulf Arabic (afb), Bahrani (abv), Omani (acx), Najdi (ars)
Others	Iraqi (acm), Yemeni (ayn), MSA (arb)

8.2 MADAR Benchmark (13 dialects)

Dialect	Gherbal v4	OpenLID v2	GlotLID	NLLB-LID
MSA	0.871	0.471	0.822	**0.998**
Moroccan	0.925	**0.955**	0.890	0.000
Egyptian	0.746	**0.878**	0.798	0.000
Iraqi	0.561	**0.888**	0.762	0.000
Levantine	0.777	**0.843**	0.522	0.000
Tunisian	0.823	**0.938**	0.821	0.000
Najdi	0.477	**0.807**	0.807	0.000
Algerian	0.288	0.000	**0.315**	0.000
Libyan	**0.243**	0.000	0.000	0.000
Gulf	**0.115**	0.000	0.000	0.000
Sudanese	**0.094**	0.000	0.000	0.000
Omani	**0.035**	0.000	0.000	0.000
Yemeni	**0.013**	0.000	0.000	0.000

8.3 Atlasia-LID Benchmark (15 dialects)

Dialect	Gherbal v4	OpenLID v2	GlotLID	NLLB-LID
MSA	0.969	0.714	0.674	**0.999**
Moroccan	0.847	**0.881**	0.784	0.000
Egyptian	0.937	**0.962**	0.931	0.000
Hassaniya	**0.552**	0.000	0.000	0.000
Tunisian	0.659	**0.721**	0.657	0.000
Levantine	**0.668**	0.651	0.349	0.000
Algerian	**0.418**	0.000	0.037	0.000
Gulf	**0.378**	0.000	0.000	0.000
Sudanese	**0.301**	0.000	0.000	0.000
Najdi	0.273	**0.642**	0.617	0.000
Iraqi	0.223	**0.354**	0.311	0.000
Libyan	**0.183**	0.000	0.000	0.000
Yemeni	**0.156**	0.000	0.000	0.000
Omani	**0.046**	0.000	0.000	0.000
Bahrani	**0.001**	0.000	0.000	0.000

8.4 Analysis

Coverage tiers:

1. **Tier 1 — All models** (OpenLID-v2, Gherbal v4, GlotLID): MSA, Moroccan, Egyptian, Tunisian, Levantine, Iraqi, Najdi. These are the highest-resource Arabic dialects with established web presence.
2. **Tier 2 — Gherbal v4 only** (except Algerian, also covered by GlotLID): Libyan, Gulf, Sudanese, Omani, Yemeni, Hassaniya, Bahrani. These are lower-resource dialects where only Gherbal v4 achieves non-zero accuracy.
3. **NLLB-LID** effectively collapses all Arabic to MSA (0.998 MSA, 0.000 everything else).

Confusion patterns (Gherbal v4): - Gulf → Najdi (lexically similar peninsular dialects) - Levantine → MSA (formal/written Levantine resembles MSA) - Bahrani → Gulf (geographic and linguistic proximity) - Omani → Najdi/Gulf (peninsular overlap)

The challenge with Tier 2 dialects is that training data is extremely scarce and often code-switched with MSA. Gherbal v4's ability to distinguish these at all — even at low accuracy — represents a genuine capability gap compared to every other evaluated model.

Figures 8 and 9 show per-dialect accuracy on MADAR and Atlasia-LID respectively. On MADAR, Gherbal v4 is the only model with non-zero bars for Gulf, Sudanese, Omani, and Yemeni. On Atlasia-LID, the same pattern holds — plus Hassaniya and Bahrani appear, which only Gherbal v4 attempts.

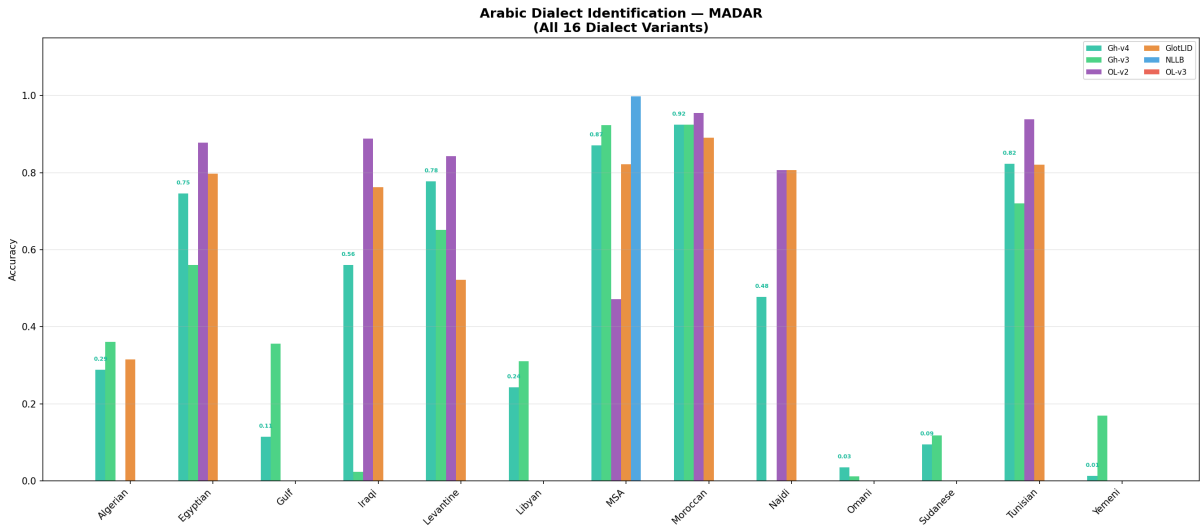


Figure 8: Detailed Arabic dialect accuracy – MADAR

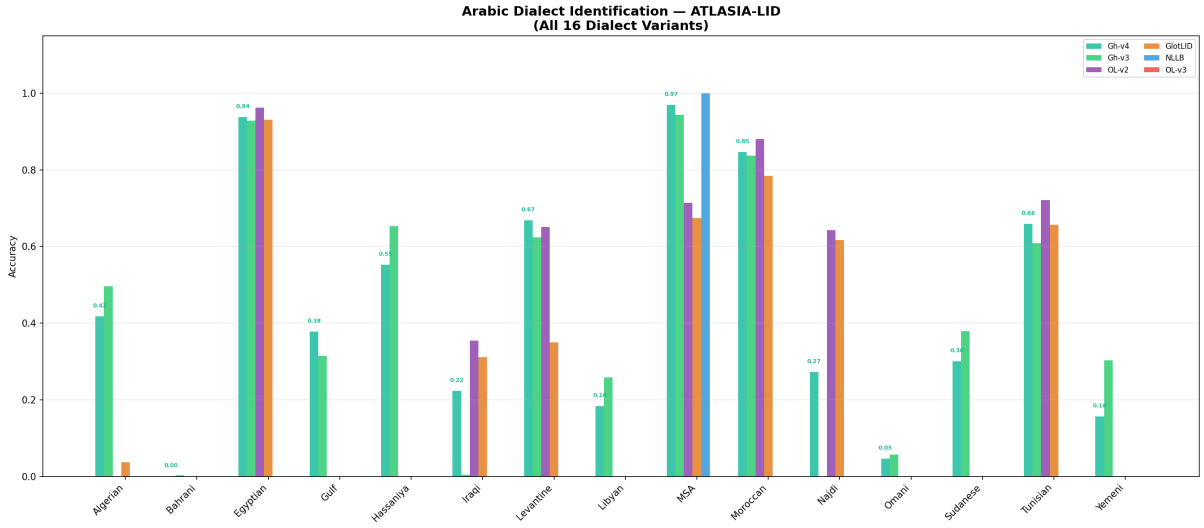


Figure 9: Detailed Arabic dialect accuracy – Atlasia-LID

The confusion matrix (Figure 10) reveals where Gherbal v4's Arabic predictions go wrong. The dominant off-diagonal patterns confirm the geographic confusion clusters described above: Gulf↔Najdi, Levantine→MSA, and Bahrani→Gulf. These are linguistically predictable confusions that reflect genuine dialect continua rather than model failures.

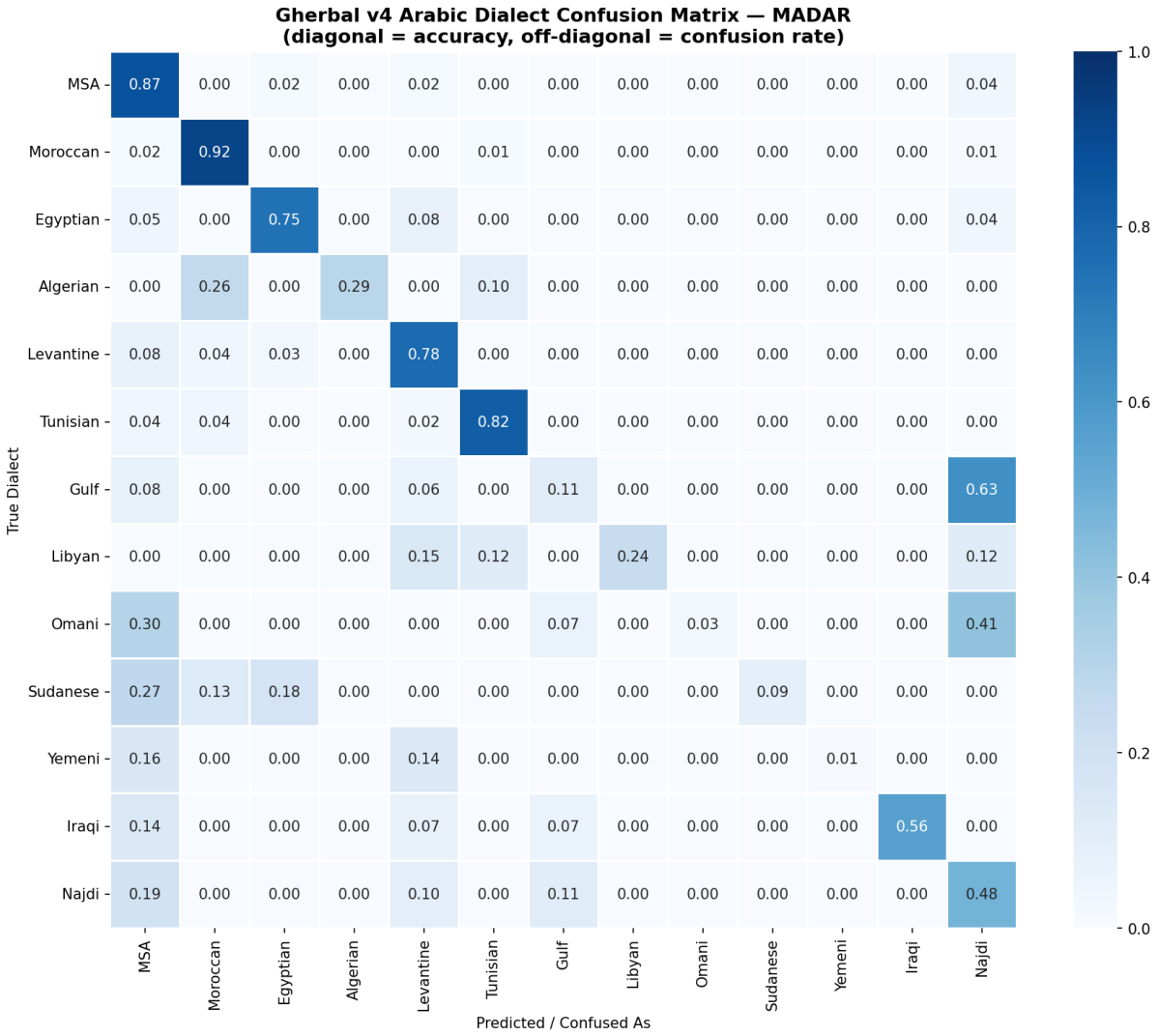


Figure 10: Gherbal v4 inter-dialect confusion matrix (MADAR)

Finally, Figure 11 quantifies the coverage gap directly: the number of Arabic dialects each model can identify at above-chance accuracy. Gherbal v4 covers all 16 variants; the next best (OpenLID-v2) covers only 8.

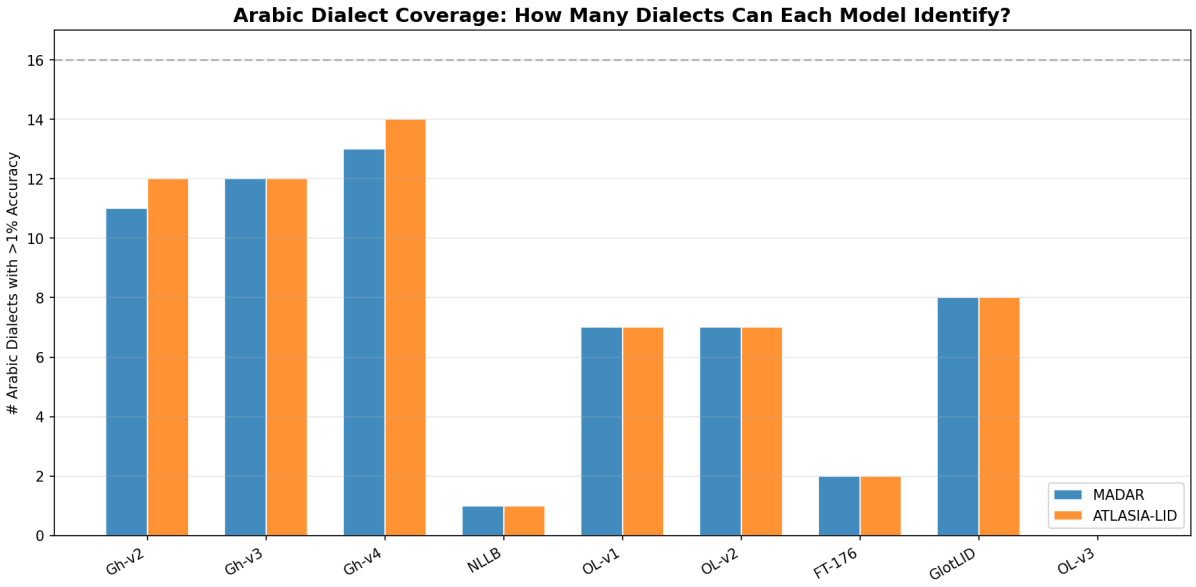


Figure 11: Arabic dialect coverage per model

8.5 Arabizi: The Unsupported Script

Arabizi (`ary_Latn`) — Moroccan Arabic written in Latin script — represents a unique challenge for language identification. Used extensively in North African social media, messaging apps, and informal web content, Arabizi is the dominant written form of Darija for millions of young Moroccans. Despite its prevalence, **no competing LID model in this evaluation supports Arabizi.** Only Gherbal models include `ary_Latn` as a classification target.

What Arabizi looks like:

Arabizi uses Latin characters, often with numerals to represent Arabic-specific phonemes. Common conventions include `3` for ain, `7` for ha, `9` for qaf, and `kh` for kha. For example:

- *"Salam, ki dayra? Ana bikhir lhamdoulilah"* (Hello, how are you? I'm fine, thank God)
- *"Bghit nmchi l casa f weekend"* (I want to go to Casa for the weekend)
- *"3andek chi fikra 3la had l7aja?"* (Do you have any idea about this thing?)
- *"Wach nta jay l7afla dyal lila?"* (Are you coming to tonight's party?)

This orthography makes Arabizi confusable with Latin-script languages that share similar character patterns — particularly Maltese (`mlt_Latn`), Somali (`som_Latn`), Malay (`zsm_Latn`), Turkish (`tur_Latn`), and even Shilha (`shi_Latn`).

Gherbal performance across generations (Gherbal-Multi benchmark):

Model	Accuracy	n	Top Confusions
Gherbal v1	**0.982**	4,998	Somali (18), Malay (12), Swahili (8)
Gherbal v3	0.976	4,998	Shilha (19), Somali (13), Turkish (9)
Gherbal v4	0.969	4,998	Shilha (18), Somali (15), Malay (12)
Gherbal v2	0.968	4,998	Somali (26), Malay (14), Turkish (14)

All Gherbal generations achieve 96.8 – 98.2% accuracy on Arabizi, demonstrating that the model consistently learns the distinctive patterns of Latin-script Moroccan Arabic. The slight accuracy

decrease from v1 to v4 is expected: as the model's language space grows from 36 to 214 languages, more Latin-script confusables are introduced.

Competing models score 0.0% on Arabizi. Their top misclassifications reveal they attempt to map Arabizi to the Latin-script language that most resembles its character patterns:

Model	Accuracy	Top 3 misclassifications
NLLB-LID	0.000	Vietnamese (437), Polish (390), English (364)
OpenLID v1	0.000	Uzbek (681), Maltese (498), Estonian (280)
OpenLID v2	0.000	Maltese (622), Uzbek (408), Kabyle (292)
OpenLID v3	0.000	Maltese (589), Unknown (500), Uzbek (380)
GlotLID	0.000	Uzbek (354), Unknown (309), Urdu-Latin (290)
fastlid-176	0.000	English (2,264), German (354), French (321)

The misclassification patterns are revealing. OpenLID variants and GlotLID most commonly confuse Arabizi with Maltese — linguistically the closest Latin-script language, since Maltese descends from Siculo-Arabic and shares Semitic roots. fastlid-176 defaults to high-resource European languages (English, German, French), indicating no learned representation for any similar language.

Arabizi identification is important for content moderation, social media analysis, and language-aware services targeting North African users — contexts where Latin-script Arabic dialect text is common but completely invisible to current open-source LID models outside of Gherbal.

---

9 Sub-Saharan African Languages

Sub-Saharan Africa is one of the most linguistically diverse regions in the world, with over 2,000 languages. This section evaluates performance on 45 African languages across 4 regions.

9.1 FLORES devtest Results (V4 Scope)

Near-perfect identification (>0.99): Afrikaans, Amharic, Ewe, Fon, Hausa, Igbo, Kabiye, Lingala, Luba-Kasai, Luganda, Luo, Nuer, Malagasy, Sango, Shona, Somali, Swahili, Tswana, Tsonga, Tumbuka.

Strong identification (0.95–0.99): Bambara, Bemba, Chokwe, Dinka, Fulfulde, Oromo, Kamba, Kimbundu, Mossi, Northern Sotho, Chewa, Kirundi, Sotho, Swati, Tigrinya, Twi, Umbundu, Wolof, Xhosa, Yoruba, Zulu.

Challenging (below 0.95): - Kinyarwanda/Kirundi confusion (0.976/0.952): These are mutually intelligible. - Dyula (0.703): Limited training data; confused with Bambara (essentially a closely related Manding variety).

9.2 Gherbal v4 vs Competitors

On FLORES devtest, Gherbal v4 performs comparably to the larger models on African languages:

Region	Gherbal v4	OpenLID v2	GlotLID	NLLB-LID
West Africa (12 langs)	0.962	0.920	0.949	0.962
East Africa (8 langs)	0.990	0.994	0.984	0.994
Southern Africa (10 langs)	0.980	0.970	0.983	0.984
Central Africa (7 langs)	0.982	0.898	0.973	0.974

Notable findings:

- **Kituba** (`ktu_Latn`): Gherbal v4 scores 0.992 while NLLB-LID scores 0.011 and GlotLID scores 0.000. This is a genuinely low-resource language where targeted training data curation pays off.
- **Dyula** (`dyu_Latn`): Gherbal v4 scores 0.703 vs NLLB-LID 0.026 and OpenLID-v2 0.029. Another case where most models have negligible performance.
- **Kamba** (`kam_Latn`): Gherbal v4 at 0.990 vs NLLB-LID 0.612 — a large quality gap indicating better training data for this Bantu language.
- **Twi** (`twi_Latn`): Gherbal v4 at 0.997 vs NLLB-LID 0.557 — similar advantage.

On Bouquet (a domain-shifted benchmark), Gherbal v4 shows strong performance on African languages, though some languages like Chokwe (0.493) and Kimbundu (0.476) drop more than competitors — likely because Bouquet contains machine-translated text with different characteristics from natural web data.

The African language heatmap (Figure 12) makes the regional patterns visible. Most of the continent shows green/yellow (high accuracy) for Gherbal v4, with isolated cool spots on languages where Bouquet's machine-translated domain causes drops.

Figure 12: Sub-Saharan African language accuracy – FLORES devtest

Figure 13 isolates Gherbal v4's delta against the best competitor for each African language. Positive bars (green) indicate languages where Gherbal v4 leads — notably Kituba (+0.98), Dyula (+0.67), and Kamba (+0.38). Negative bars show where competitors edge ahead, typically by small margins on high-resource languages.

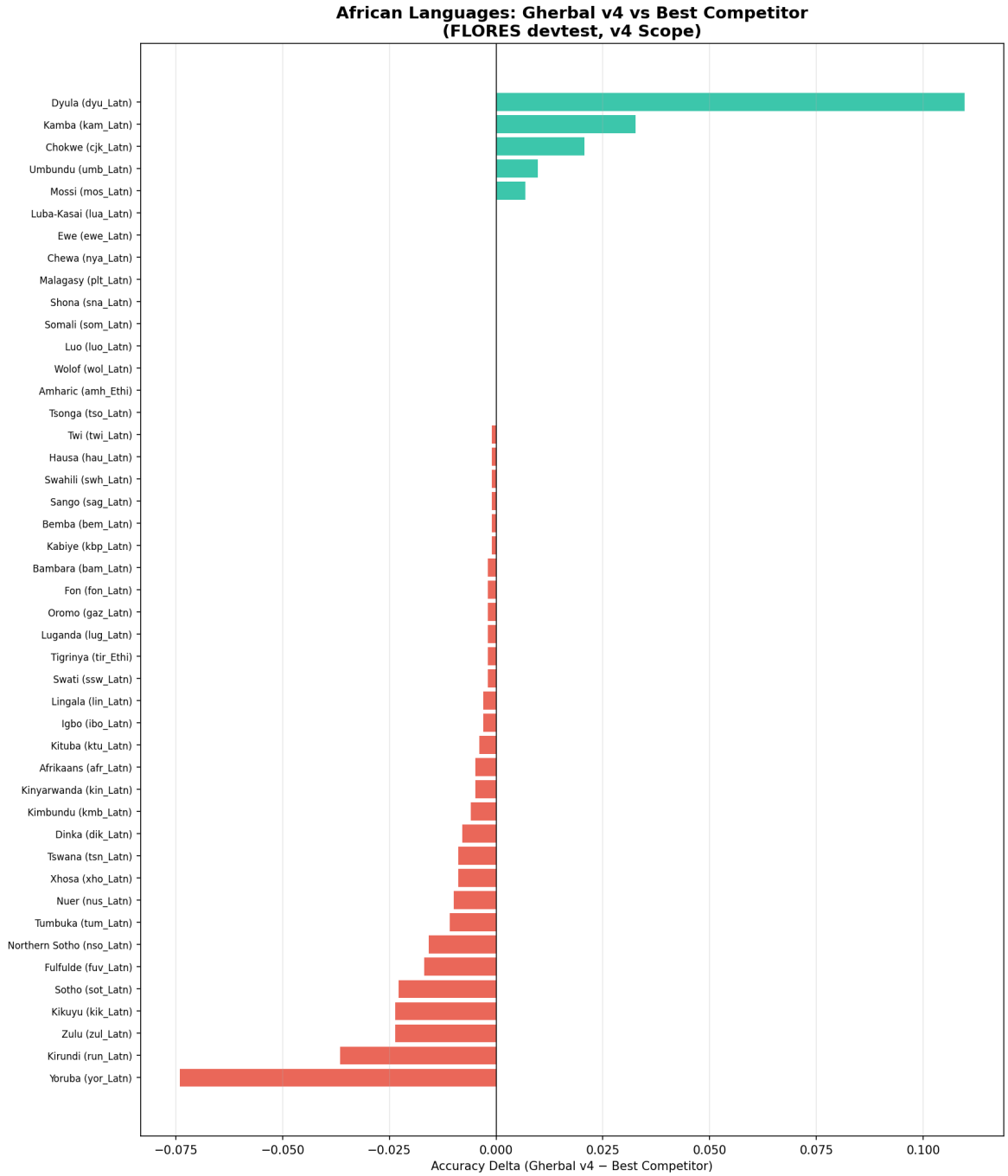


Figure 13: Gherbal v4 delta vs best competitor – African languages

Grouping by sub-region (Figure 14) confirms that Gherbal v4 is competitive or leading across all four African sub-regions, with particular strength in Central and West Africa where low-resource language coverage matters most.

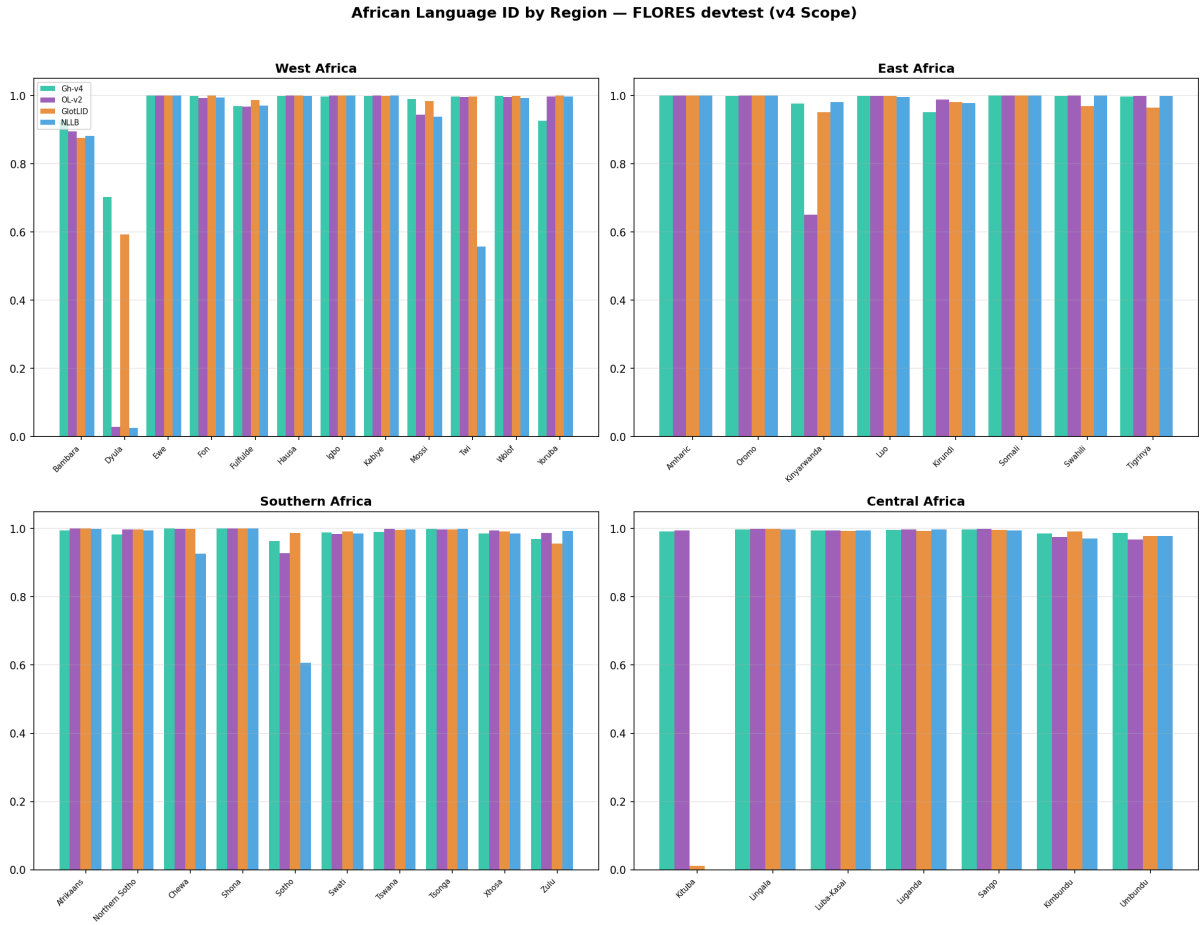


Figure 14: African language ID by sub-region

10 Model Efficiency Analysis

10.1 Size vs Performance

Model	Size (MB)	V4 Avg Accuracy	MB per point
Gherbal v4	200	0.836	239
Gherbal v1	30	0.342	88
OpenLID v2	1,230	0.824	1,493
GlotLID	1,690	0.803	2,105
NLLB-LID	1,180	0.711	1,660
OpenLID v3	1,360	0.628	2,166

Gherbal v4 achieves the best accuracy-per-megabyte ratio by a factor of 6.2× compared to OpenLID v2. This makes Gherbal v4 significantly more practical for deployment in resource-constrained environments (mobile, edge, embedded systems).

10.2 Training Data Efficiency

Beyond model size, Gherbal v4 is also dramatically more data-efficient during training:

Model	Training Data Size	Languages
Gherbal v4	<3 GB	214
OpenLID v2	~21 GB	201
GlotLID	~45 GB	2,102

Gherbal v4's training data is less than 3 GB — roughly 7× smaller than OpenLID-v2's 21 GB training set — yet achieves higher average accuracy. This data efficiency stems from the 4-pass cleaning pipeline: rather than training on more data, Gherbal focuses on training on *better* data.

10.3 Scope Scaling Behavior

As the language space grows from V1 (36) to Full (214+), all models experience accuracy degradation. The table below shows FLORES-devtest accuracy at each scope:

Model	V1 Acc	V4 Acc	Full Acc	Degradation (V1→Full)
Gherbal v1	0.889	0.149	0.137	-84.5%
Gherbal v4	0.959	0.921	0.850	-11.4%
OpenLID v2	0.974	0.939	0.875	-10.2%
GlotLID	0.947	0.940	0.925	-2.3%
NLLB-LID	0.935	0.889	0.833	-10.9%

GlotLID shows the smallest degradation (-2.3%) because its training set covers 2,102 languages — it's already operating in a wide-scope regime. OpenLID-v2 and Gherbal v4 show similar degradation (-10.2% and -11.4% respectively), while NLLB-LID degrades somewhat more (-10.9%). Gherbal v4's quantized embedding space makes it slightly more sensitive to the confusion introduced by out-of-scope predictions.

The efficiency frontier (Figure 15) plots model size against V4 accuracy. Gherbal v4 sits in the top-left corner — high accuracy, small model — while the cluster of 1+ GB models in the right half achieves comparable or lower accuracy at 6–8× the size. The gap between Gherbal v4 and the rest of the Pareto frontier is the clearest visual summary of its efficiency advantage.

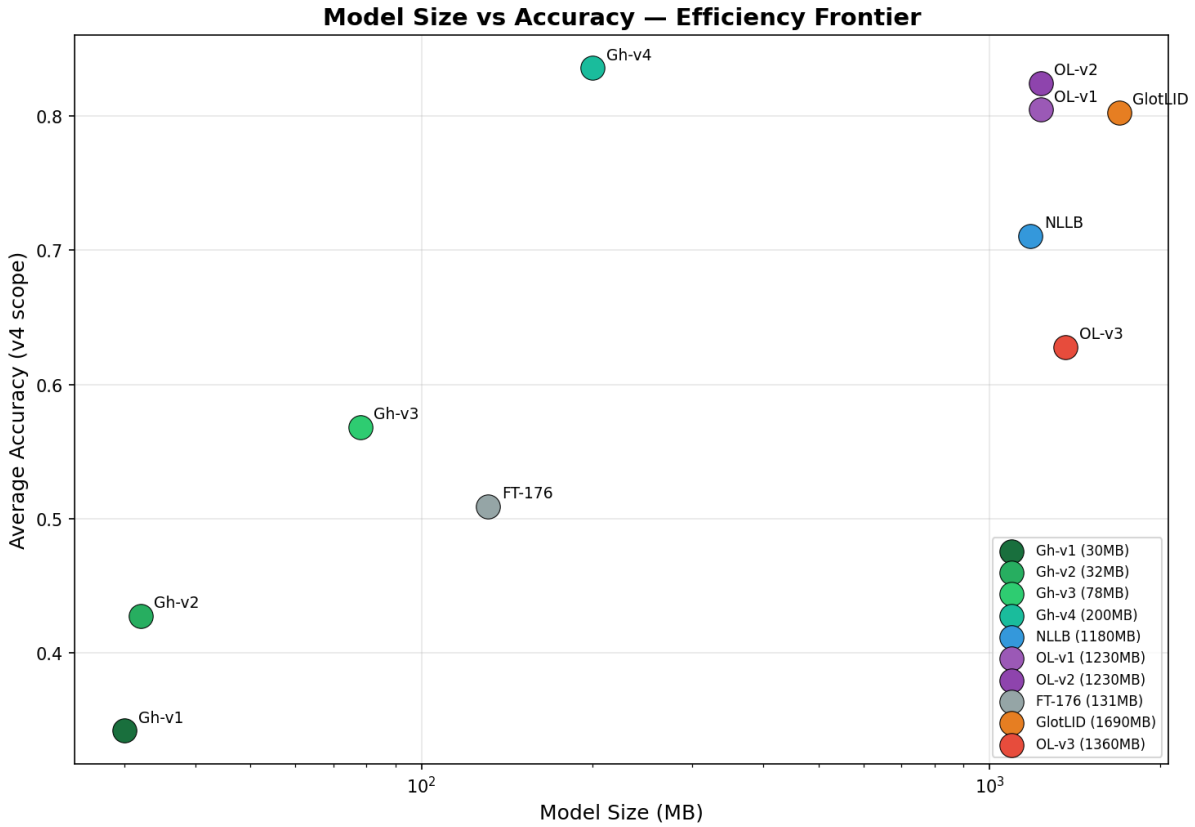


Figure 15: Model size vs accuracy efficiency frontier

Figure 16 traces how each model's accuracy changes as the classification space grows from 36 (V1) to 214+ (Full) languages. All models decline, but the slopes differ: GlotLID is nearly flat (already trained on 2,102 languages), while NLLB-LID drops steeply. Gherbal v4 maintains a competitive position at every scope level.

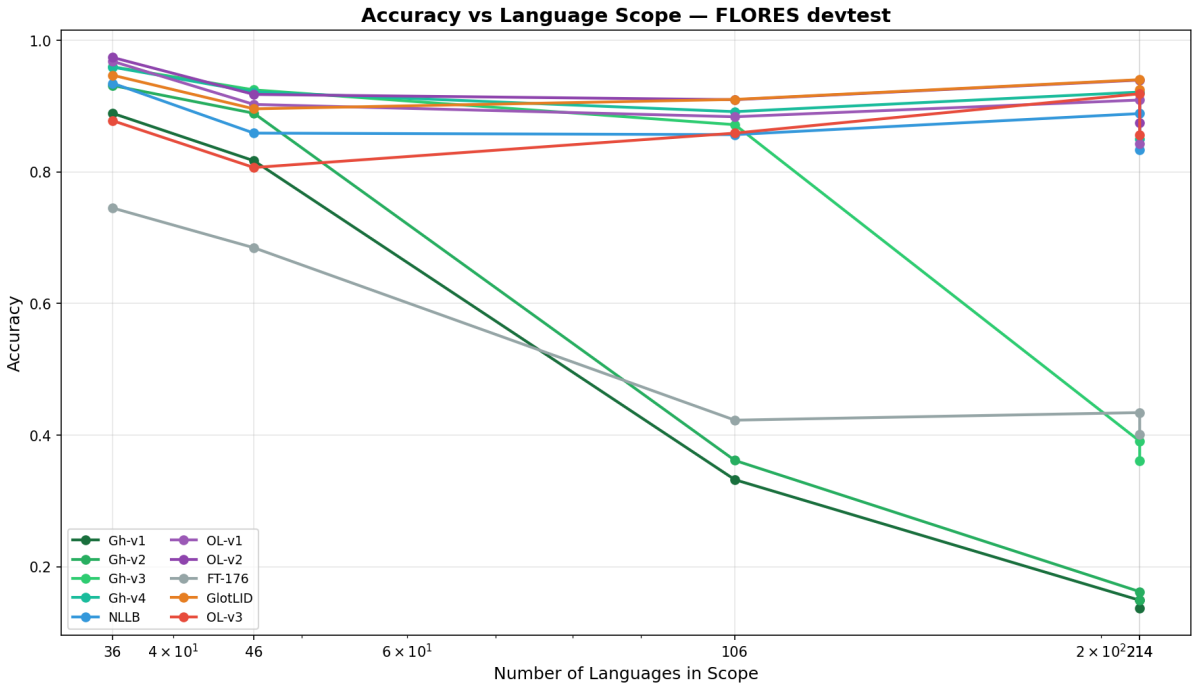


Figure 16: Accuracy vs language scope size

11 Script and Language Family Analysis

11.1 By Script

FLORES devtest performance grouped by writing system (V4 scope):

Script	# Langs	Gherbal v4	OpenLID v2	GlotLID	NLLB-LID
Latin	100+	0.927	0.941	0.934	0.905
Arabic	20+	0.748	0.818	0.780	0.452
Cyrillic	12	0.971	0.979	0.979	0.958
Devanagari	7	0.968	0.984	0.990	0.964
Bengali	3	0.988	0.993	0.986	0.986
CJK	2	0.974	0.998	0.994	0.967

Arabic-script languages are the hardest category for all models, driven by the dialect identification challenge. For all other scripts, accuracy is above 0.92 across the board.

11.2 By Language Family

Family	Gherbal v4	OpenLID v2	GlotLID	NLLB-LID
IE (Romance)	0.985	0.987	0.990	0.978
IE (Germanic)	0.986	0.994	0.993	0.979

Family	Gherbal v4	OpenLID v2	GlotLID	NLLB-LID
IE (Slavic)	0.965	0.977	0.980	0.948
IE (Indic)	0.966	0.981	0.986	0.960
Semitic	0.881	0.906	0.895	0.804
Turkic	0.966	0.978	0.968	0.958
Bantu	0.983	0.978	0.968	0.938
West African	0.950	0.915	0.937	0.948
Austronesian	0.944	0.970	0.973	0.927
Dravidian	0.986	0.993	0.983	0.990

Gherbal v4 leads on Bantu languages and is competitive across all families. The Semitic family is the most challenging for all models due to Arabic dialect overlap.

Figure 17 breaks down accuracy by writing system. The high bars for Cyrillic, Devanagari, Bengali, and CJK scripts confirm that unique scripts make LID easy — the real challenge is Latin-script and Arabic-script languages where many languages share the same character set.

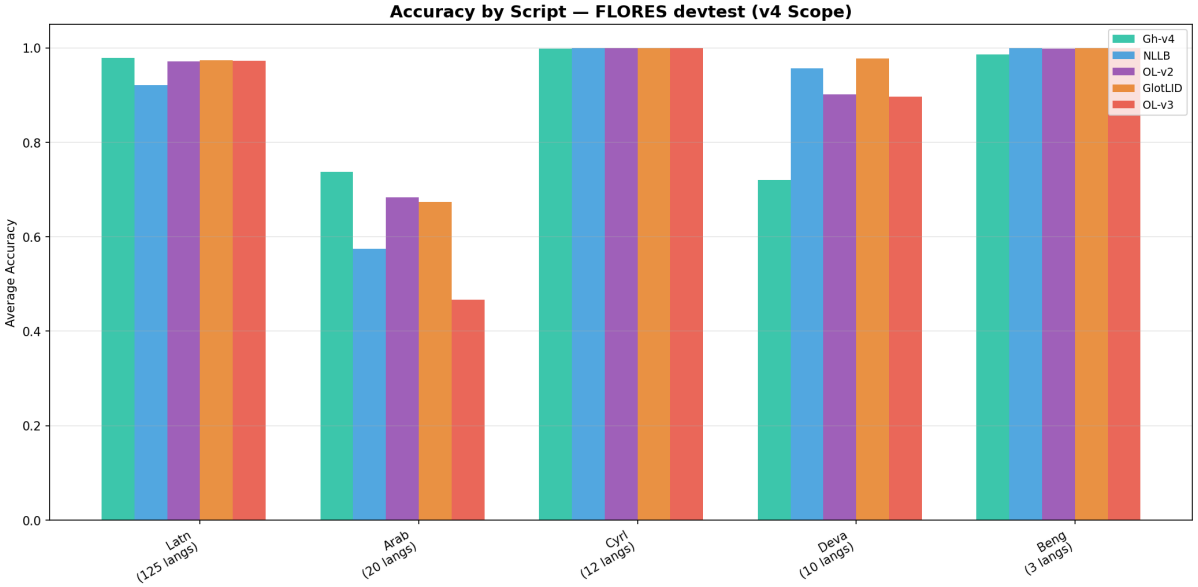


Figure 17: Accuracy by writing system

Grouping by linguistic family (Figure 18) reveals that Gherbal v4's Bantu and West African performance is best-in-class, offsetting the small gaps on Indo-European families where all models perform similarly. The Semitic bar is notably lower for all models, driven by the Arabic dialect challenge.

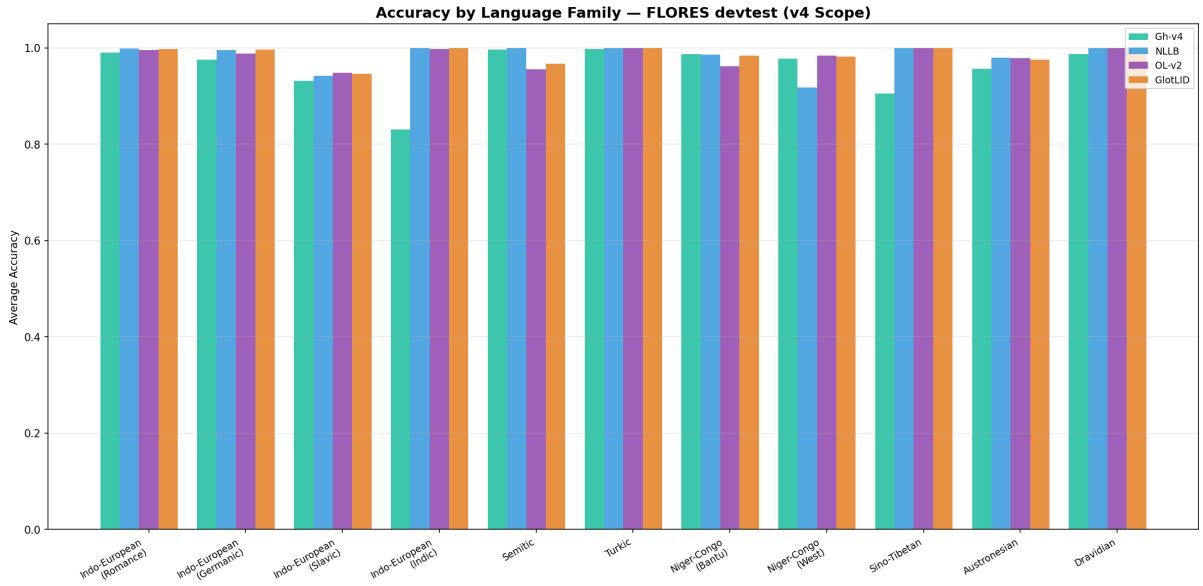


Figure 18: Accuracy by linguistic family

12 Architecture & Training Details

12.1 Gherbal Evolution

Version	Languages	Training Strategy	Key Innovation
v1	36	FastText on curated data	Initial prototype, Arabic + core languages
v2	46	FastText on curated data	Arabic dialect expansion
v3	106	4-pass cleaning pipeline	Systematic data quality
v4	214	Expanded multilingual coverage	Scale + quality balance

Figure 19 charts this evolution. Each generation expands language coverage while maintaining or improving per-language accuracy on core languages. The progression from v1's 36 languages to v4's 214 is a 6× expansion. The v3→v4 jump is the most significant: language count doubles from 106 to 214, model size grows from 78 to 200 MB, and average accuracy on the expanded scope increases.

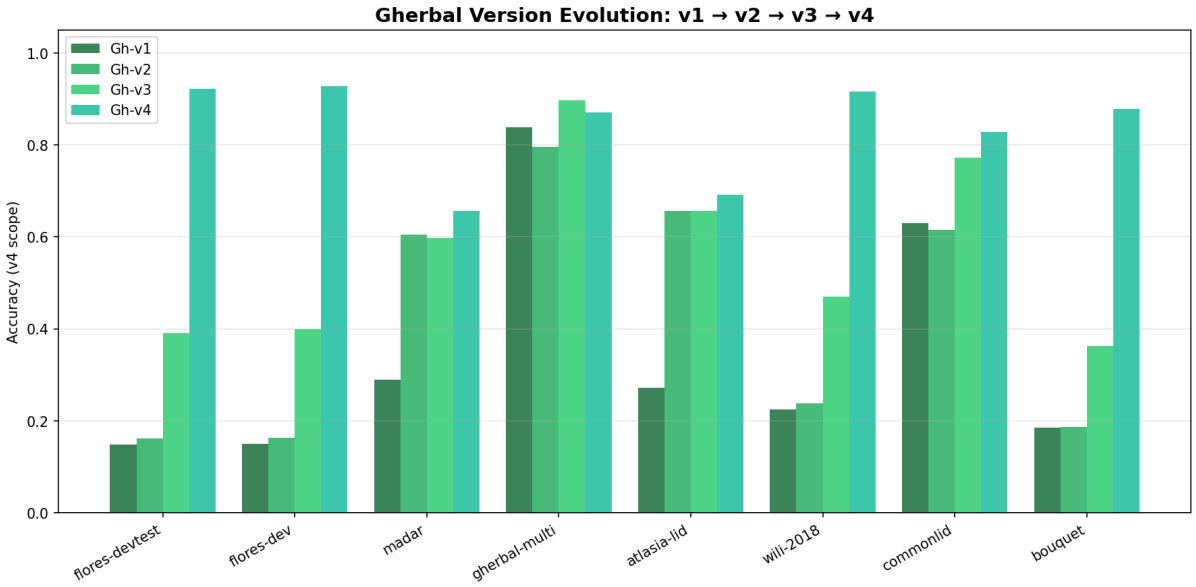


Figure 19: Gherbal version evolution: v1 → v2 → v3 → v4

The radar chart below (Figure 20) overlays all four Gherbal generations on the same axes, showing how each successive version expands the model's capability polygon. Gherbal v1 and v2 show strong performance only on benchmarks within their limited scope (Gherbal-Multi, CommonLID), while v3 begins to fill in, and v4 achieves a uniformly large polygon across all 8 benchmarks.

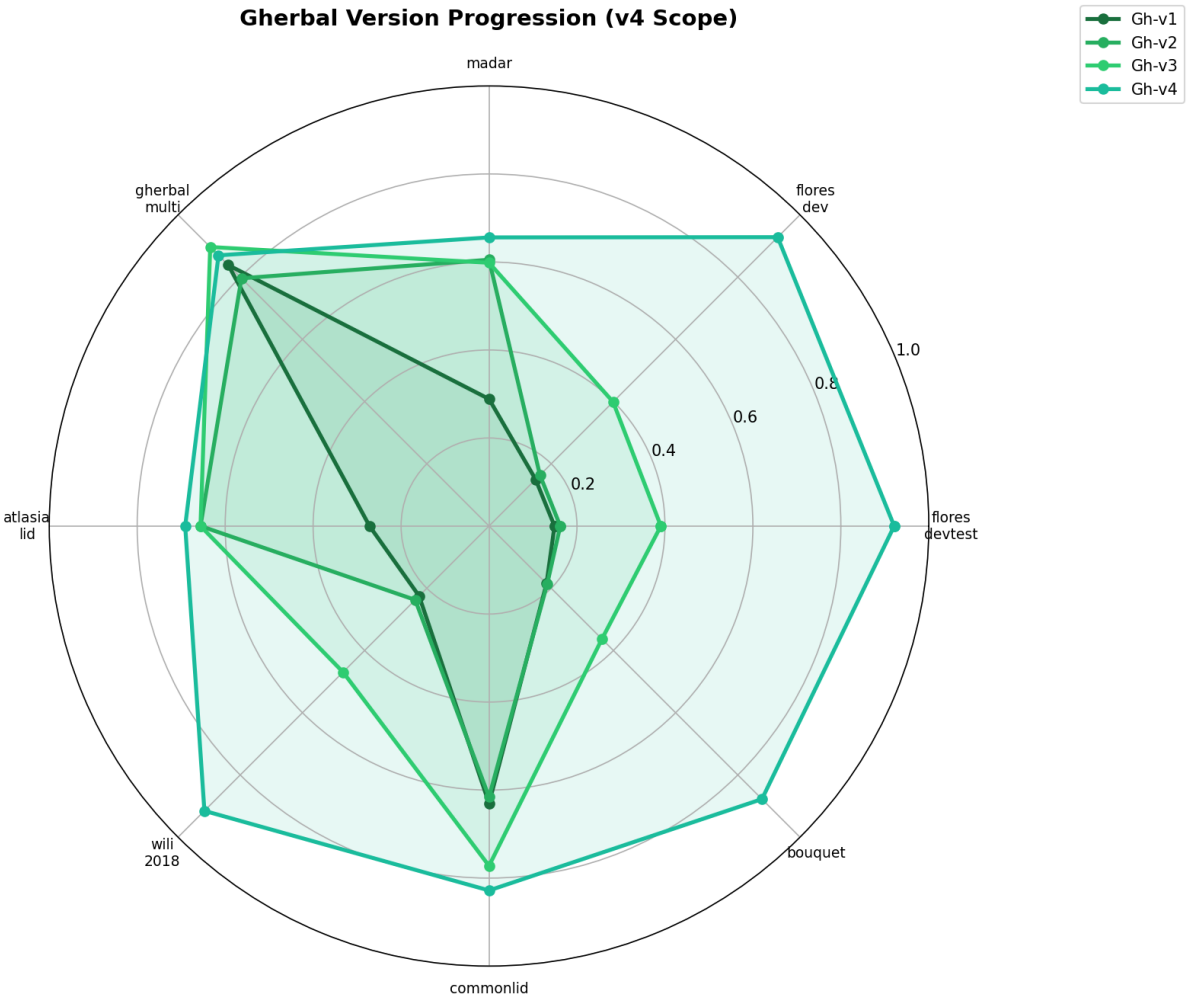


Figure 20: Gherbal version progression – radar overlay

12.2 Training Pipeline (v4)

1. **Data Collection:** Aggregated from multiple curated sources including web-crawled text, seed data, and Wikipedia, with per-language caps to ensure balanced representation. The total training set is under 3 GB — roughly 7× smaller than OpenLID-v2's training data.
2. **4-Pass Cleaning:**
   - *Pass 1: Script Validation (90% threshold)* — Ensured text matches expected Unicode script using the `unscript` library.
   - *Pass 2: Cross-Language Deduplication* — Removed identical texts labeled with multiple language codes.
   - *Pass 3: Self-Prediction Disambiguation* — 3-fold internal model trained to flag confident disagreements with existing labels.
   - *Pass 4: Temperature Resampling ($p^{0.3}$)* — Smooth power-law sampling to avoid hard caps. Preserves tail languages while controlling head-language dominance.
3. **Augmentation:** Added FLORES dev set for broad language coverage.
4. **Training:** FastText supervised with tuned hyperparameters (`dim=400, epoch=5, wordNgrams=2, minn=2, maxn=5`).
5. **Quantization:** FastText quantization for deployment efficiency.

12.3 Why FastText?

Despite the rise of transformer-based LID models, FastText remains the optimal architecture for production LID when:

- **Latency matters:** FastText classifies text in microseconds vs milliseconds for transformers.
- **Model size matters:** 200 MB (quantized) vs multi-GB for transformer models.
- **Short text:** n-gram features capture character patterns that transformers need larger context for.
- **Arabic dialects:** Character-level subword features (`minn=2, maxn=5`) capture morphological patterns that distinguish dialects.

---

13 Benchmark Difficulty Analysis

Average accuracy across all 10 models (V4 scope), ranked easiest to hardest:

Rank	Benchmark	Avg Accuracy	Characteristic
1	WiLI-2018	0.778	Long, clean Wikipedia text
2	Gherbal-Multi	0.778	Curated multi-domain
3	CommonLID	0.772	Real-world web text
4	FLORES devtest	0.723	Standard NLP benchmark
5	FLORES dev	0.724	Same domain
6	Bouquet	0.672	Machine-translated
7	MADAR	0.381	Arabic dialects only
8	Atlasia-LID	0.421	Arabic dialects only

Arabic-only benchmarks (MADAR, Atlasia) are dramatically harder than general benchmarks because Arabic dialect identification is an unsolved problem for most models. The gap between general benchmarks (0.72–0.78) and Arabic benchmarks (0.38–0.42) quantifies this challenge.

Figure 20 ranks benchmarks by average model accuracy. The visual drop-off from the general benchmarks (top) to the Arabic-specific ones (bottom) is dramatic — a 2× difficulty gap that underscores why Arabic dialect competence is such a strong differentiator.

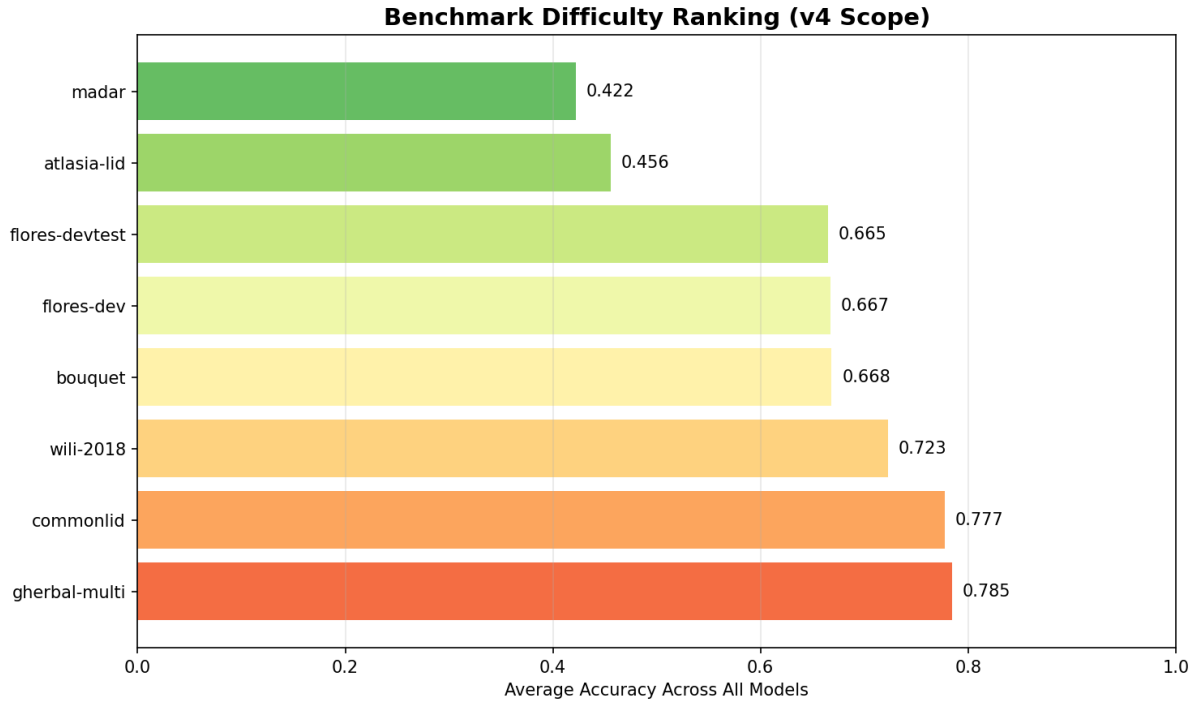


Figure 21: Benchmark difficulty ranking

The win matrix (Figure 21) shows how often each model ranks 1st, 2nd, 3rd, etc. across benchmarks. Gherbal v4's distribution is concentrated in the top-3 ranks, while competitors show more spread — no other model is as consistently competitive across all benchmark types.

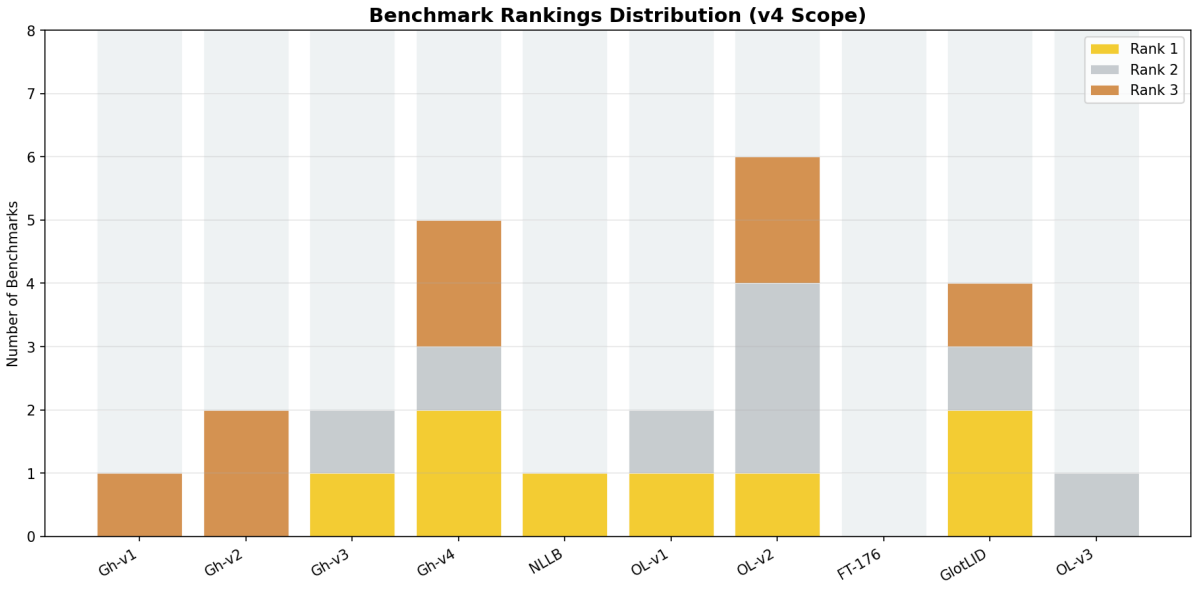


Figure 22: Benchmark rankings distribution

14 Strengths and Weaknesses

14.1 Gherbal v4 Strengths

1. **Best average accuracy** across all 8 benchmarks (V4 scope).
2. **Arabic dialect breadth**: Only model to identify all 16 dialects.
3. **Exceptional size-efficiency**: 0.836 average accuracy at 200 MB.
4. **Low-resource African language quality**: Best-in-class on Dyula, Kituba, Kamba, Twi.
5. **North African specialization**: Unique coverage of Hassaniya, Libyan, and most rare Arabic dialects; shared Algerian coverage with GlotLID only.

14.2 Gherbal v4 Weaknesses

1. **Not best on any single general benchmark**: OpenLID-v2 leads on FLORES, Bouquet; NLLB-LID leads on CommonLID; OpenLID-v1 leads on WiLI.
2. **Full-scope degradation**: Accuracy drops from 0.921 → 0.850 when expanding from V4 to Full scope (-11.4%), comparable to OpenLID-v2 (-10.2%) but more than GlotLID (-2.3%).
3. **Gulf/Peninsula Arabic dialects**: Accuracy on Gulf (0.115), Omani (0.035), Yemeni (0.013), Bahrani (0.001) is still very low — these are genuinely data-scarce dialects.
4. **Mutually intelligible language confusion**: Kinyarwanda/Kirundi, Indonesian/Malay, Serbian/Croatian remain challenging (as they are for all models).

Figure 22 summarizes the strengths and weaknesses in a single view: Gherbal v4's delta against the best competitor on each benchmark. The positive bars on MADAR, Atlasia, and Gherbal-Multi represent the Arabic and specialization advantages; the small negative bars on FLORES and Bouquet show where OpenLID-v2's larger model edges ahead on general text.

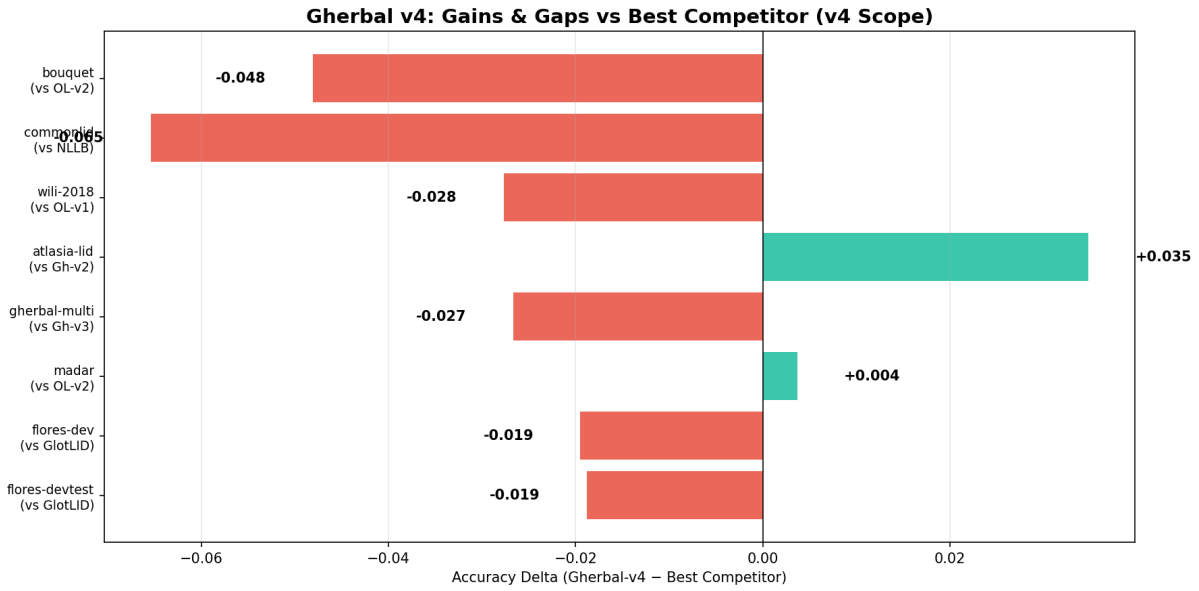


Figure 23: Gherbal v4 gains/gaps vs best competitor per benchmark

The per-language accuracy distribution (Figure 23) shows that the vast majority of Gherbal v4's 214 languages cluster above 0.90 accuracy, with a long tail of harder languages — predominantly Arabic dialects and a few mutually intelligible pairs.

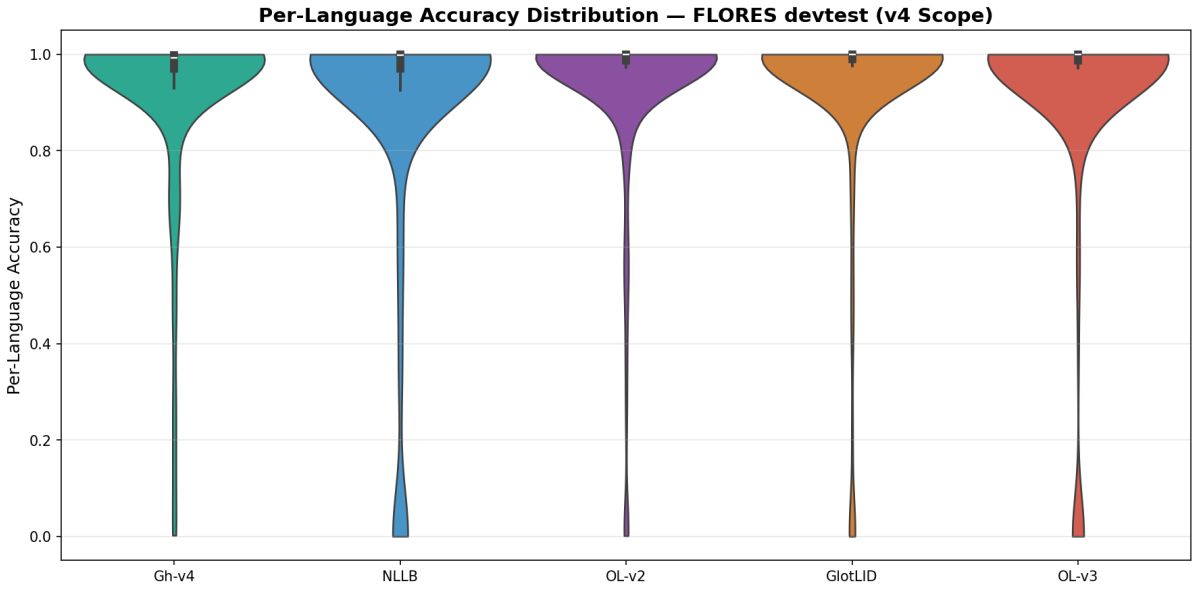


Figure 24: Per-language accuracy distribution

Finally, Figure 24 lists the easiest and hardest languages for Gherbal v4. The easiest are scripts-unique languages (Tibetan, Georgian, Armenian); the hardest are data-scarce Arabic dialects (Bahrani, Yemeni, Omani) and confusable language pairs.

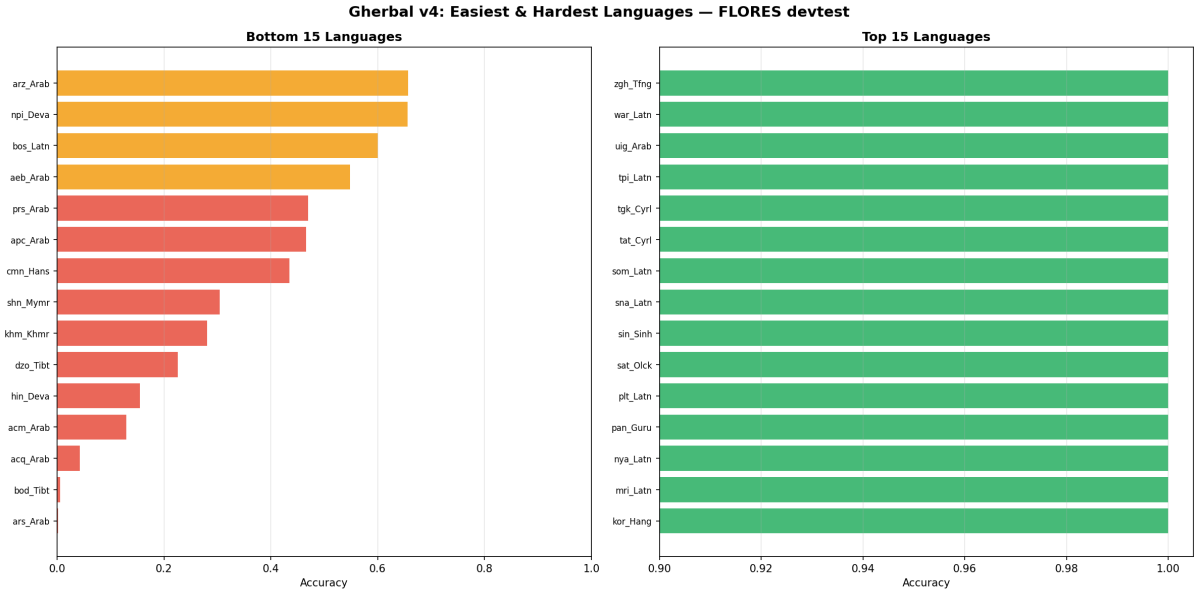


Figure 25: Easiest and hardest languages for Gherbal v4

15 Conclusion

Gherbal v4 demonstrates that **data quality and targeted curation** can outperform model scale. At one-sixth the size of its nearest competitors, it achieves the best overall accuracy by leveraging:

- A principled 4-pass cleaning pipeline for training data quality.
- Temperature resampling to balance head and tail languages.
- Dedicated Arabic dialect and North African language data sourcing.
- FastText's character n-gram features that naturally capture dialect-level morphological patterns.

For practitioners deploying LID in production — particularly for applications involving Arabic content, North African languages, or African languages — Gherbal v4 offers the best combination of accuracy, coverage, and efficiency.

---

16 References

16.1 Models

[1] **NLLB Team** et al. "No Language Left Behind: Scaling Human-Centered Machine Translation." *arXiv:2207.04672*, 2022. The NLLB-LID model (218 languages, FastText-based) is described in Section 5.1 of this paper. https://arxiv.org/abs/2207.04672

[2] **Burchell, L., Birch, A., Bogoychev, N. & Heafield, K.** "An Open Dataset and Model for Language Identification." *Proceedings of ACL 2023*, pp. 865–879. Toronto, Canada. Describes OpenLID v1 and v2 (201 languages). https://arxiv.org/abs/2305.13820

[3] **de Gibert, O., Nail, G., Arefyev, N. et al.** "A New Massive Multilingual Dataset for High-Performance Language Technologies." *Proceedings of LREC-COLING 2024*, Torino, Italy. Describes the HPLT project, under which OpenLID v3 was trained. https://arxiv.org/abs/2403.14009

[4] **Facebook Research.** *fastText Language Identification Model (lid.176.bin)*. Pre-trained model released 2017. Covers 176 languages. Trained on Wikipedia, Tatoeba, and SETimes data using the FastText supervised architecture. https://fasttext.cc/docs/en/language-identification.html

[5] **Kargaran, A.H., Imani, A., Yvon, F. & Schütze, H.** "GlotLID: Language Identification for Low-Resource Languages." *Findings of EMNLP 2023*, pp. 6155–6218. Singapore. Covers 2,102 languages. https://arxiv.org/abs/2310.16248

[6] **Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T.** "Bag of Tricks for Efficient Text Classification." *Proceedings of EACL 2017*, Volume 2: Short Papers, pp. 427–431. Valencia, Spain. Describes the FastText supervised text classification architecture used by all models in this evaluation. https://arxiv.org/abs/1607.01759

16.2 Benchmarks and Datasets

[7] **Goyal, N., Gao, C., Chaudhary, V. et al.** "The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation." *Transactions of the ACL (TACL)*, 10:522–538, 2022. Our evaluation uses FLORES+ (`openlanguagedata/flores_plus`), the community-maintained successor to FLORES-200 [1]. https://arxiv.org/abs/2106.03193

[8] **Bouamor, H., Habash, N., Salameh, M. et al.** "The MADAR Arabic Dialect Corpus and Lexicon." *Proceedings of LREC 2018*, Miyazaki, Japan. See also the MADAR shared task: **Bouamor, H., Hassan, S. & Habash, N.** "The MADAR Shared Task on Arabic Fine-Grained Dialect Identification." *Proceedings of WANLP 2019*, pp. 199–207. Florence, Italy. https://aclanthology.org/L18-1535/

[9] **Thoma, M.** "The WiLI Benchmark Dataset for Written Language Identification." *arXiv:1801.07779*, 2018. Wikipedia-based benchmark covering 235 languages. https://arxiv.org/abs/1801.07779

[10] **Atlasia.** *Atlasia-LID Benchmark.* Arabic dialect identification dataset covering 15 Arabic dialect variants across social media, web, and news domains. https://huggingface.co/datasets/atlasia/Atlasia-LID

[11] **Meta AI (NLLB).** *FLORES+ / Bouquet Benchmark.* Machine-translated sentence-level evaluation across 275 languages, released as part of the NLLB ecosystem.

[12] **CommonCrawl Foundation / WMDQS.** *CommonLID Benchmark.* Web-crawled language identification test set derived from CommonCrawl data, covering 101 languages across noisy, real-world web text.

[13] **Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T.** "Enriching Word Vectors with Subword Information." *Transactions of the ACL (TACL)*, 5:135–146, 2017. Describes the character n-gram approach underlying FastText, which enables morphologically-aware representations critical for LID. https://arxiv.org/abs/1607.04606

17 Appendix: Chart Index

#	Chart	Description
01	Heatmap (V4/Full)	Accuracy matrix: models × benchmarks
02	Grouped Bars	Per-benchmark model comparison
03	Radar	Model profiles across benchmarks
04	Version Evolution	Gherbal v2 → v3 → v4 progression
05	Arabic Dialects (MADAR)	Per-dialect accuracy on MADAR
06	Scope Scaling	Accuracy vs language scope size
07	F1 Diagnostic	F1-macro vs F1-weighted scatter
08	Script Families	Performance by writing system
09	Win Matrix	Benchmark ranking distribution
10	Size vs Accuracy	Efficiency frontier
11	Per-Language Distribution	Violin plot of per-language accuracy
12	Delta vs Best	Gherbal v4 gains/gaps vs best competitor
13	Benchmark Difficulty	Difficulty ranking across models
14	North African Heatmap	NA language accuracy across models
15	North African Per-Benchmark	NA languages on FLORES/MADAR/Atlasia
16	Berber/Amazigh Focus	Berber language specialization
17	Arabic Detailed (MADAR/Atlasia)	All 16 dialects on both benchmarks
18	Arabic Confusion Heatmap	Gherbal v4 inter-dialect confusion
19	Arabic Coverage	Dialect coverage per model
20	African Heatmap	45 African languages on FLORES
21	African Delta	Gherbal v4 vs best competitor per lang
22	African Regional	Performance by African sub-region
23	Top/Bottom Languages	Easiest and hardest languages
24	Per-Script Detailed	Accuracy by writing system (all models)
25	Language Families	Accuracy by linguistic family