

When Multilingual Embeddings Collapse: Representation Failure and Metric (Un)reliability for Hausa and Fongbe Machine Translation

Mahounan Pericles Adjovi^{1*} Roald Eiselen² Prasenjit Mitra^{3*}

¹Blekinge Institute of Technology, Sweden

²Centre for Text Technology (CTeXT), North-West University, South Africa

³Victoria University of Wellington, New Zealand

`mahounanpericles.adjovi@bth.se`

`Roald.Eiselen@nwu.ac.za` `prasenjit.mitra@vuw.ac.nz`

Abstract

Neural evaluation metrics for machine translation inherit their behaviour from the multilingual encoders beneath them, yet how those encoders represent low-resource African languages is rarely examined. We evaluate four large language models on English-to-Hausa and English-to-Fongbe translation across scales from 500 to 10,000 sentences, validate automatic metrics against native-speaker judgment, and then probe the multilingual representation space that neural metrics depend on. The central finding is a representation failure: XLM-R assigns near-identical vectors to semantically different texts in both languages (within-language cosine similarity above 0.99), so BERTScore cannot separate systems of visibly different quality, and COMET produces a ranking anomaly for Fongbe. This representation collapse is a central reason embedding-based metrics are unreliable here, rather than any surface property of the languages. We further show that metric-human rank agreement is language-dependent, that in our samples system rankings stabilise by about 2,500 sentences, and that inclusion of a language in an encoder’s pretraining set does not guarantee usable representations. We frame these results as evidence that multilingual evaluation should be checked at the representation level before it is trusted.

1 Introduction

Machine translation for low-resource African languages is limited by data scarcity (Joshi et al., 2020; Nekoto et al., 2020), and large language models (LLMs) are increasingly proposed as a way to generate synthetic parallel data. Whether such translations are usable, and how to measure their quality, both depend on evaluation metrics that were developed and validated on high-resource European languages. The neural metrics in widest use,

COMET (Rei et al., 2020) and BERTScore (Zhang et al., 2020), do not measure quality directly; they read it off the geometry of a multilingual encoder, almost always XLM-R (Conneau et al., 2020). Their reliability therefore rests on an assumption that is rarely tested for low-resource languages: that the encoder represents the target language well enough for distances in its space to track meaning.

We test that assumption for **Hausa** (Afroasiatic, roughly 58 million speakers) and **Fongbe** (Niger-Congo, roughly 2.3 million speakers). We first run a controlled MT evaluation, four LLMs at four scales with human validation, to establish what good and bad translations look like for these pairs. We then use that study as an evidence base for the paper’s main contribution: a probe of the XLM-R representation space that neural metrics rely on. The probe shows that the space collapses for both languages, which is consistent with the metric behaviour we observe, most notably a COMET ranking anomaly for Fongbe and a near-total loss of discrimination in BERTScore.

This reframing matters because it moves the problem from “which metric scores highest” to “does the representation underneath the metric carry meaning at all.” We address two questions. RQ1 asks how well current LLMs translate into Hausa and Fongbe, and how quality varies by system and language. RQ2 asks whether automatic metrics, and the multilingual representations they depend on, reliably reflect human judgment for these languages.

Contributions. First, we provide a controlled LLM translation benchmark for Hausa and Fongbe with native-speaker validation across scales. Second, and centrally, we diagnose a representation failure in XLM-R for both languages and relate it to the unreliability of embedding-based metrics, while stating explicitly what it can and cannot explain. Third, we give practical, representation-aware guid-

*Work conducted while at Carnegie Mellon University Africa.

ance for evaluating MT in low-resource settings, including a simple pre-use probe of embedding quality.

2 Related Work

MT for African languages. Community efforts have widened coverage while leaving depth uneven. NLLB (NLLB Team, 2022) covers 200 languages including Hausa and Fongbe; CCMatrix (Schwenk et al., 2021) provides web-mined pairs mainly for Hausa; English2Gbe (Hacheme, 2021) built early NMT for Gbe languages; Hausa appears in the FLORES-101 benchmark (Goyal et al., 2022), and both languages in FLORES-200 (NLLB Team, 2022). AfriDoc-MT (Alabi et al., 2025) benchmarks document-level MT for several African languages but excludes Fongbe. These works report automatic scores; they do not validate whether the scores track human judgment for the specific languages evaluated.

LLMs for low-resource translation. Commercial LLMs degrade sharply for languages with limited web presence (Robinson et al., 2023; Hendy et al., 2023), and instruction tuning does not uniformly help (Zhu et al., 2024). Africa-centric encoders such as AfroXLMR (Alabi et al., 2022) and AfriBERTa (Ogueji et al., 2021) improve understanding tasks but target comprehension rather than generation.

Metrics and multilingual representation. Lexical metrics (BLEU Papineni et al., 2002; chrF++ Popović, 2017; TER Snover et al., 2006) measure surface overlap; neural metrics predict human scores from encoder representations and correlate better on high-resource pairs (Freitag et al., 2022). Their dependence on the encoder is the crux: the geometry of pretrained representations is known to be anisotropic (Ethayarajh, 2019; Gao et al., 2019; Li et al., 2020), and multilingual encoders vary widely in how well they serve individual languages (Wu and Dredze, 2020; Pires et al., 2019). AfriMTE and AfriCOMET (Wang et al., 2024) already established, with substantially larger human annotation, that COMET is unreliable for many under-resourced African languages; we take that general finding as given. What remains open, and what we contribute, is a representation-level account of *why* embedding-based metrics fail for Hausa and Fongbe specifically, together with a cheap pre-use probe; we discuss the AfriCOMET

comparison itself as planned work (Section 8).

3 Linguistic Background

Hausa and Fongbe are both tonal, and differ mainly in whether tone is marked in the standard orthography. Hausa has grammatical gender, rich morphology, and complex tense-aspect-mood marking (Newman, 2000); tone is usually left unmarked in writing. Fongbe uses serial verb constructions and a three-tone system that is obligatorily marked with diacritics, where tone distinguishes lexical meaning (for example *kó* “harvest,” *kò* “build,” *kô* “neck”) (Lefebvre and Brousseau, 2002). Two consequences matter for evaluation. Morphological richness multiplies valid surface forms, so a correct translation can score poorly on exact-match metrics. And obligatory diacritics mean that correctness can hinge on marks that models frequently drop and that metrics may ignore, which places heavy weight on whether the underlying encoder represents these distinctions at all.

4 Methodology

Models. We evaluate four systems with temperature 0 for deterministic output: GPT-4o mini, Claude Sonnet 4, and Gemini 2.5 Flash (commercial), and Qwen2.5-7B (open-weight). The set pairs current commercial models with an open-weight baseline; broader open-weight coverage is left to future work (Section 8). Exact model snapshots are recorded in Appendix A.

Prompting. All systems receive the same instruction, which names the target language, requests only the translation, and provides five fixed in-context example pairs per language (5-shot). Outputs are decoded greedily (temperature 0, at most 128 new tokens) and reduced to a single line. The full template appears in Appendix A, and is held constant across models and scales so that differences reflect the systems rather than prompt tuning.

Data. We draw parallel sentences from OPUS (Tiedemann, 2012): CCMatrix (Schwenk et al., 2021) for English–Hausa (992,266 pairs) and NLLB (NLLB Team, 2022) for English–Fongbe (915,489 pairs), with Unicode NFC normalisation and deduplication. To study how evaluation stabilises with scale, we draw independent random samples with a fixed seed at 500, 1,000, 2,500, and 10,000 sentences, and use 50 sentences for human evaluation.

Reference quality. A post-hoc audit of 50 random Fongbe reference pairs found issues typical of web-mined low-resource corpora: domain skew (32% religious), wrong-language contamination, and encoding corruption, giving a mean quality of 2.94/5 with 6% severely corrupted. Three cases illustrate the corruption: a wrong-language pair whose Fongbe reference is in fact Chinese text (source “Speaking of demand, let’s get right to it.”); an encoding-corrupted pair that renders as mojibake (source “Looking forward to 3 days off!”); and, by contrast, a clean pair carrying full tone diacritics (source “Their territory was Jazer, and all the cities of Gilead. . .”). Because our contribution is about relative system behaviour and metric reliability rather than absolute quality, rankings remain informative, but absolute scores should be read with this caveat, and language-identification filtering is advisable before evaluation.

Metrics. We compute BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), and TER (Snover et al., 2006) with sacreBLEU (Post, 2018); COMET (Rei et al., 2020) (wmt22-comet-da); and BERTScore (Zhang et al., 2020) with XLM-R-large. Fongbe is absent from XLM-R’s 100 pre-training languages while Hausa is present, which we return to in the representation analysis.

Human evaluation. Thirteen native speakers were recruited (7 Fongbe, 6 Hausa); three native-speaker raters per language were used for the final analysis. Table 1 describes the full recruited pool, which is why it also lists the non-native and intermediate participants who were not part of the final set. Fongbe evaluators reported lower written literacy (reading 3.0, writing 2.0) than Hausa evaluators (4.0, 3.5), reflecting differences in formal orthographic education, which matters because they rated written output. Annotators rated Adequacy, Fluency, and Cultural Appropriateness on 5-point scales and gave an overall score; outputs were anonymised and order-randomised. Qwen2.5-7B was excluded from human rating given its very low automatic scores. Inter-annotator agreement, computed as pairwise linearly weighted Cohen’s κ (Cohen, 1960) averaged over evaluator pairs and interpreted following Landis and Koch (1977), was moderate for both languages (mean κ approximately 0.42 for Fongbe and 0.48 for Hausa) and highest on cultural appropriateness.

Table 1: Evaluator demographics and self-reported proficiency (1–5) for the full recruited pool. The three raters retained per language were all native speakers.

Characteristic	Fongbe ($n=7$)	Hausa ($n=6$)
Native speaker	6 (86%)	4 (67%)
Fluent (non-native)	1 (14%)	1 (17%)
Intermediate	0 (0%)	1 (17%)
<i>Proficiency (mean)</i>		
Speaking	4.57	4.50
Listening	4.71	4.67
Reading	3.00	4.00
Writing	2.00	3.50

Analysis. We report metric–human agreement with Spearman ρ at the system level. With only three human-rated systems per language, ρ takes values in $\{-1, -0.5, 0, 0.5, 1\}$ and should be read as rank agreement, not as a statistical correlation. We probe representation quality through out-of-vocabulary (OOV) rate and within-language cosine similarity: we embed the target-side reference sentences with XLM-R-large, mean-pool the final hidden layer to a sentence vector, sample up to 100 sentences per language, and compute the mean pairwise cosine similarity within each language (full procedure in Appendix B). Following Ethayarajh (2019), healthy spaces show within-language similarity near 0.5–0.7, whereas values above 0.95 indicate collapse (Gao et al., 2019; Li et al., 2020). Significance uses bootstrap resampling (Koehn, 2004) with 10,000 iterations.

5 Translation Quality (RQ1)

Table 2 reports automatic scores at $n=2,500$, our primary scale with complete metric coverage. Table 3 reports human scores.

Table 2: Automatic evaluation at $n=2,500$. Best per language in **bold**; $\downarrow$ lower is better. Models ordered alphabetically.

Lang	Model	BLEU	chrF++	TER $\downarrow$	COMET	BERT
Fon	Claude	4.97	15.60	97.28	0.365	0.884
	Gemini	7.18	18.63	118.01	0.356	0.880
	GPT-4o mini	1.87	9.00	105.80	0.375	0.860
	Qwen	0.53	5.30	111.03	0.432	0.832
Hau	Claude	15.75	40.39	75.98	0.597	0.878
	Gemini	13.51	38.50	93.89	0.584	0.877
	GPT-4o mini	14.19	38.37	77.66	0.587	0.876
	Qwen	2.20	20.73	104.36	0.450	0.845

A large cross-language gap. Hausa scores roughly three times higher than Fongbe across metrics and scales; the best Fongbe system (Gemini, BLEU 7.18) trails the best Hausa system (Claude,

Table 3: Human evaluation (1–5), three native speakers per language. Best per language in **bold**.

Lang	Model	Adeq.	Flu.	Cult.	Overall
Fon	Claude	1.03	1.04	1.03	1.03
	Gemini	2.19	2.27	2.14	2.20
	GPT-4o mini	1.02	1.11	1.03	1.05
Hau	Claude	4.09	4.28	4.21	4.19
	Gemini	3.02	3.54	3.54	3.37
	GPT-4o mini	4.41	4.51	4.46	4.46

BLEU 15.75) at $n=2,500$. Human ratings confirm the divide: Hausa reaches 3.4–4.5/5 (usable), Fongbe only 1.0–2.2/5 (unusable). The relationship between data availability and quality is well known; what our probe adds later is a representation-level factor that plausibly deepens the gap.

Rankings are language-specific. No model wins both languages. Gemini leads for Fongbe, while for Hausa the automatic metrics rank Claude first but human raters prefer GPT-4o mini (4.46 vs 4.19). Performance on one low-resource African language does not predict performance on another, consistent with differences in pretraining representation rather than a general “low-resource capability.”

Scale matters. Rankings among the top three stabilise only at $n \geq 2,500$. At $n=1,000$, sampling variance produced misleading orders: for Hausa, Gemini briefly appeared to lead before falling to third by $n=10,000$ (BLEU 11.52), and Claude scored anomalously low (BLEU 2.84 versus 14.91 at $n=10,000$). Because each scale uses a single fixed-seed sample, we present $n \geq 2,500$ as a practical rule of thumb from our runs rather than a variance-backed guarantee; evaluations below 2,500 sentences risk artifact conclusions.

Large-scale validation. Extending to $n=10,000$ (Table 4) confirms the rankings: Gemini remains best for Fongbe and Claude for Hausa on surface metrics. Neural metrics were unavailable for Hausa at this scale (COMET returned zero for all systems), which is why Table 2 at $n=2,500$ remains our primary reference. Sentence-level bootstrap resampling over the 10,000-sentence set (Table 5) makes all reported pairwise differences significant at $p < 0.001$; the same comparisons were non-significant at $n=10$ ($p=0.70$), a caution against small-sample evaluation, though significance at large n reflects sampling power rather than the practical size of a gap.

Table 4: Large-scale validation at $n=10,000$ (surface metrics; neural metrics unavailable for Hausa at this scale). Best per language in **bold**.

Lang	Model	BLEU	chrF++	TER↓
Fon	Claude	2.51	11.11	106.44
	Gemini	4.66	15.46	105.22
	GPT-4o mini	1.29	6.30	119.20
	Qwen	0.68	5.32	110.32
Hau	Claude	14.91	39.55	76.88
	Gemini	11.52	36.14	81.82
	GPT-4o mini	14.50	39.08	77.66
	Qwen	2.23	20.56	105.07

Table 5: Selected pairwise significance from sentence-level bootstrap resampling (BLEU, $n=10,000$; 10,000 iterations). All comparisons are significant at $p < 0.001$.

Lang	Comparison	Δ	95% CI	p
Fon	Gemini vs Claude	+1.99	[1.88, 2.10]	<0.001
Fon	Claude vs GPT-4o mini	+0.34	[0.27, 0.41]	<0.001
Hau	Claude vs GPT-4o mini	+0.53	[0.36, 0.70]	<0.001
Hau	GPT-4o mini vs Gemini	+2.14	[1.96, 2.33]	<0.001

6 Representation Failure and Metric Reliability (RQ2)

This section is the core of the paper. We show that the neural metrics behave anomalously, relate the behaviour to the XLM-R representation space, and then interpret the surface-metric correlations in that light.

6.1 Neural metrics behave anomalously

COMET diverges from every other signal for Fongbe: Qwen2.5-7B receives the highest COMET score (0.432) despite ranking last on all lexical metrics (BLEU 0.53) and being excluded from human evaluation for very low quality. The anomaly persists at $n=10,000$, where Qwen again receives the highest Fongbe COMET score (0.434). BERTScore, meanwhile, is nearly flat across systems, 0.832–0.884 for Fongbe and 0.845–0.878 for Hausa, despite human overall scores that span more than a full point for Fongbe. A metric that cannot separate a 1.0/5 system from a 2.2/5 system is not measuring quality. Both symptoms point to the shared component beneath these metrics, the multilingual encoder. We cannot fully isolate the COMET anomaly from alternative causes without inspecting Qwen’s outputs by hand: TER above 100 for several Fongbe systems is consistent with over-generation, and the corrupted references compound this, so we report the result as *consistent*

with, rather than proof of, inadequate Fongbe representation.

6.2 Probing the XLM-R space

Table 6 reports representation diagnostics. Within-language cosine similarity exceeds 0.99 for *both* languages, far above the 0.5–0.7 range of a healthy space and past the 0.95 collapse threshold. In such a space the encoder assigns almost the same vector to different sentences, so any metric that reads meaning from cosine distance, BERTScore directly and COMET through its encoder, has little room left to discriminate.

Table 6: XLM-R representation diagnostics, computed over up to 100 target-side reference sentences per language. Within-language similarity above 0.95 indicates collapse.

Diagnostic	Fongbe	Hausa
OOV rate	0.33%	4.67%
Within-language similarity	0.996	0.995
Avg. embedding norm	30.61	30.11

Inclusion is not representation. The OOV pattern is counter-intuitive: Hausa, which *is* in XLM-R’s pretraining set, shows a much higher OOV rate (4.67%) than Fongbe (0.33%), which is not. Inspection shows Hausa OOV tokens are largely named entities, Arabic/English loanwords adapted to Hausa orthography, and agglutinated forms absent from the subword vocabulary. Fongbe’s low OOV is an artifact: the tokenizer, having no Fongbe subwords, falls back to character-level pieces it treats as “known” while carrying no learned meaning for them. Membership in the pretraining language list therefore does not guarantee usable representations; in particular, XLM-R-large is not a meaningful semantic encoder for Fongbe, which is reason enough to distrust BERTScore-XLM-R there.

6.3 Interpreting the correlations

Table 7 gives system-level metric–human rank agreement over the three rated systems. We report only the lexical metrics, because the collapsed neural metrics (Section 6.1) produce uninformative rank agreement: BERTScore’s 6% spread cannot be meaningfully ranked against a 113% spread in human scores.

For Fongbe, all lexical metrics identify Gemini as best (agreement 1.0), but this likely reflects

Table 7: Spearman ρ (rank agreement, $n=3$ systems) between lexical metrics and human overall scores. *Values indicate rank agreement, not statistical correlation.

Language	BLEU	chrF++	TER
Fongbe	1.00*	1.00*	−1.00*
Hausa	0.50*	0.50*	−0.50*

Gemini’s clear dominance (2.20 vs ~ 1.0) rather than metric reliability. For Hausa, metrics rank Claude first while humans prefer GPT-4o mini, a single rank swap ($\rho=0.5$) consistent with automatic metrics missing the naturalness and idiomatic quality Hausa speakers reward; with three systems, we treat this as indicative rather than statistically distinguishable. We report BERTScore without baseline rescaling, so its narrow range is suggestive rather than decisive. The lesson is not that one metric is best but that agreement is language-dependent and, for neural metrics, contingent on a representation space that may have collapsed.

6.4 Within-language metric (in)consistency

Beyond agreement with humans, we ask whether the metrics agree with *each other*. For Fongbe, human overall ranks Gemini > GPT-4o mini > Claude, whereas BLEU and chrF++ rank Gemini > Claude > GPT-4o mini; for Hausa, humans rank GPT-4o mini > Claude > Gemini while every lexical metric ranks Claude first. Tables 8 and 9 give the full picture. The sharpest inconsistency is COMET’s: for Fongbe it ranks Qwen2.5-7B *first* despite that system placing last on every lexical metric and being unfit for human rating, the symptom discussed in Section 6.1.

Table 8: Fongbe metric rankings at $n=2,500$.

Metric	1	2	3	4
BLEU	Gemini	Claude	GPT-4o mini	Qwen
chrF++	Gemini	Claude	GPT-4o mini	Qwen
TER↓	Claude	GPT-4o mini	Qwen	Gemini
COMET	Qwen	GPT-4o mini	Claude	Gemini
BERTScore	Claude	Gemini	GPT-4o mini	Qwen
Human	Gemini	GPT-4o mini	Claude	—

For Hausa the automatic metrics are internally consistent (all rank Claude first) yet still disagree with speakers, who prefer GPT-4o mini; agreement among metrics is therefore not evidence of validity.

Table 9: Hausa metric rankings at n=2,500. $\approx$ denotes a BERTScore difference below 0.001 (Gemini 0.877 vs GPT-4o mini 0.876).

Metric	1	2	3	4
BLEU	Claude	GPT-4o mini	Gemini	Qwen
chrF++	Claude	Gemini	GPT-4o mini	Qwen
TER↓	Claude	GPT-4o mini	Gemini	Qwen
COMET	Claude	GPT-4o mini	Gemini	Qwen
BERTScore	Claude	Gemini $\approx$ GPT-4o mini	Qwen	—
Human	GPT-4o mini	Claude	Gemini	—

6.5 A paraphrastic-variation hypothesis, tested

We asked whether Fongbe’s low BLEU reflected greater paraphrastic variation rather than worse translation. Small samples (n=10) seemed to support this (higher Fongbe type-token ratio; higher Fongbe BERTScore), but the effect vanished at n=2,500 (type-token ratio 0.234 vs 0.238; BERTScore 0.864 vs 0.869, each averaged across systems). The apparent difference was a small-sample artifact, and the representation collapse, not paraphrase, is what flattens BERTScore. Table 10 shows the reversal directly.

Table 10: Small- vs large-sample comparison. Values are means across the four systems at each scale. Effects visible at n=10 disappear or reverse at n=2,500.

Finding	n=10	n=2,500	Status
BLEU: Hau > Fon	14.48 / 3.58	11.41 / 3.64	Confirmed
BERTScore: Fon > Hau	0.87 / 0.72	0.864 / 0.869	Reversed
Fongbe TTR	0.80	0.234	Changed
Hausa TTR	0.53	0.238	Changed

6.6 An illustrative failure

A short example makes the collapse concrete. For a Fongbe source whose reference carries full tone diacritics, a system output that drops the diacritics changes or destroys lexical meaning, yet BERTScore moves only within the 0.83–0.88 band, because XLM-R maps both the diacritized and undiacritized strings to nearly the same vector. The metric records almost no change where a native speaker records a large one. This single case is the whole paper in miniature: the failure is in the representation, and the metric merely inherits it.

7 Discussion

Validate the representation, not just the metric. The practical implication is a cheap pre-use check: before trusting an embedding-based metric

on a new low-resource language, compute within-language cosine similarity on a sample of target text. Values above roughly 0.9 are a warning that the metric will not discriminate quality, regardless of its reported correlations on other languages. In our data this check is coarse rather than graded: the two languages differ by only 0.001, so it flags fragility rather than measuring its degree (Section 8).

Metric guidance for Hausa and Fongbe. Given the collapse, we recommend chrF++ as the primary metric for these two languages, since it handles inflectional variation without depending on the encoder; COMET only with representation checks and, ideally, an African-adapted variant (Section 8); and human evaluation for any final judgment, especially for Hausa where metrics and speakers diverge.

Model and scale guidance. Model choice must be language-specific (Gemini for Fongbe, though quality remains poor; GPT-4o mini or Claude for Hausa depending on whether human or automatic criteria lead), and evaluations should use at least 2,500 sentences.

A metric hierarchy for these pairs. Consolidating the evidence, we suggest the following order of trust for Hausa and Fongbe; we do not claim it transfers to other languages without testing. chrF++ is the most dependable single automatic metric, because character-level matching tolerates inflectional variation and does not route through the collapsed encoder. BLEU is usable but language-dependent: it identified the best system for both languages yet showed different rank agreement with humans. COMET should be used only after an embedding check and, ideally, with an African-adapted variant, given its Fongbe anomaly. BERTScore should not be a primary metric here at all, since its 6% spread cannot separate systems that humans rate more than a full point apart. Human evaluation remains necessary for any final judgment, most acutely for Hausa, where all automatic metrics diverge from speakers.

Why neural metrics need validation, not just reporting. The Fongbe COMET result is the clearest warning. A metric trained to predict human quality on high-resource pairs assigned its highest Fongbe score to the system every other signal rejected. Reported uncritically, this would have inverted the conclusion. The lesson generalises:

neural metrics import the coverage and geometry of their encoder, so for a new low-resource language they should be validated against speakers before they are trusted, not adopted because they correlate well elsewhere.

Practical takeaways. For practitioners the study yields four concrete rules: probe within-language similarity before using an embedding-based metric; report multiple metrics and treat disagreement as signal rather than noise; use at least 2,500 sentences for stable rankings; and reserve a human-evaluation pass for final decisions, especially where metric and speaker rankings diverge.

8 Limitations

AfriCOMET not yet run. The clearest next step is to re-score with AfriCOMET (Wang et al., 2024), which adapts COMET to under-resourced African languages and directly targets the coverage gap we identify. We were unable to run AfriCOMET for this version, so we do not report its scores and do not claim the COMET anomaly is definitively a coverage artifact; we frame it as *consistent with* inadequate Fongbe representation and treat the AfriCOMET comparison as planned work. We will not substitute estimated numbers for a measurement we have not made.

Interpreting the collapse diagnostic. Two caveats bound Table 6. We did not measure a high-resource control language with the same pipeline, so we cannot fully separate a Hausa/Fongbe-specific failure from general encoder anisotropy or our pooling choice; a same-pipeline control is the clearest next test. And the diagnostic is nearly identical for the two languages (0.995 versus 0.996) even though their metric behaviour differs, so collapse is best read as a shared background condition that makes embedding-based metrics fragile for both, rather than as the full explanation of the language-specific differences. Because the probe runs on references, of which 32% are religious-domain, topical homogeneity may also inflate the similarity.

Other limitations. Human evaluation uses 50 sentences and three raters per language, so the correlations are rank-agreement indicators rather than statistical relationships, and Fongbe annotators had lower written literacy. Each scale uses a single fixed-seed sample, so the 2,500-sentence guidance would be strengthened by repeated subsampling.

BERTScore is reported without baseline rescaling. We evaluate one open-weight model; broader open-weight baselines (for example Llama-3, Mistral, and Africa-centric LLMs) are left to future work. We cover two languages and the English-to-target direction only, and reference-corpus contamination means absolute scores should be read cautiously.

Future work. We plan a same-pipeline control-language probe, BERTScore with baseline rescaling reported against human judgment, repeated subsampling to back the scale guidance, direct inspection of Qwen’s Fongbe outputs, and the AfriCOMET comparison above, which together would turn these diagnoses into tested claims.

9 Conclusion

Neural MT metrics for Hausa and Fongbe are unreliable, and a collapsed multilingual representation, which maps different texts to nearly identical vectors, is a central reason that a leaderboard alone would hide. Establishing this required a controlled translation benchmark, but the benchmark’s role is to expose a representation-level factor rather than to crown a system. For low-resource multilingual evaluation, the representation should be probed before the metric is trusted; inclusion in a pretraining set is not enough, and a simple similarity check can flag fragility before it corrupts conclusions.

Acknowledgements

We thank the native-speaker evaluators who participated in our translation quality study. This publication was developed as part of the Center for Inclusive Digital Transformation of Africa (CIDTA) and the Afretec Network, managed by Carnegie Mellon University Africa with financial support from the Mastercard Foundation, where the research was conducted. The presentation of this work was supported by Blekinge Institute of Technology through the Cybercampus Sweden project, Department of Computer Science. The views expressed in this document are solely those of the authors and do not necessarily reflect those of Carnegie Mellon University, the Mastercard Foundation, or Blekinge Institute of Technology.

Ethics Statement

All corpora (CCMatrix, NLLB) are publicly available under permissive licenses. Human evaluators were volunteers who gave informed consent and

could withdraw; only language proficiency was collected. Documenting weak MT quality for a language risks discouraging investment; we report it to set realistic expectations and to motivate representation-aware evaluation rather than to characterise any language as intractable.

References

- Jesujoba Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In *Proceedings of COLING*.
- Jesujoba Oluwadara Alabi, Israel Abebe Azime, Miaoran Zhang, Cristina España-Bonet, Rachel Bawden, Dawei Zhu, David Ifeoluwa Adelani, Clement Oyeleke Odoje, Idris Akinade, Iffat Maab, Davis David, Shamsuddeen Hassan Muhammad, Neo Putini, David O. Ademuyiwa, Andrew Caines, and Dietrich Klakow. 2025. AFRIDOC-MT: Document-level MT corpus for African languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 27770–27806, Suzhou, China. Association for Computational Linguistics.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*.
- Alexis Conneau et al. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.
- Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of EMNLP*.
- Markus Freitag et al. 2022. Results of WMT22 metrics shared task. In *Proceedings of WMT*.
- Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. In *Proceedings of ICLR*.
- Naman Goyal et al. 2022. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the ACL*.
- Gilles Quentin Hacheme. 2021. English2Gbe: A multilingual machine translation model for Fon/Ewe GBE. *arXiv preprint arXiv:2112.11482*.
- Amr Hendy et al. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of ACL*.
- Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*.
- Claire Lefebvre and Anne-Marie Brousseau. 2002. *A Grammar of Fongbe*. Mouton de Gruyter.
- Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of EMNLP*.
- Wilhelmina Nekoto et al. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of EMNLP*.
- Paul Newman. 2000. *The Hausa Language: An Encyclopedic Reference Grammar*. Yale University Press.
- NLLB Team. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the MRL Workshop*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of ACL*.
- Maja Popović. 2017. chrF++: Words helping character n-grams. In *Proceedings of WMT*.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of WMT*.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of EMNLP*.
- Nathaniel Robinson, Perez Ogayo, David R Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of WMT*.
- Holger Schwenk et al. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of ACL*.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of LREC*.

Jiayi Wang et al. 2024. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced african languages. In *Proceedings of NAACL*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the RepL4NLP Workshop*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of ICLR*.

Wenhao Zhu et al. 2024. Multilingual machine translation with large language models. *arXiv preprint*.

A Prompt Template and Model Snapshots

Every system received the same 5-shot instruction. For English-to-Fongbe the instruction was: “*You are an expert English-to-Fon translator. Use standard Fon Latin orthography with special characters and tone marks. Output only the Fon translation on one line, no quotes or extra text.*” It was followed by five fixed English–Fongbe example pairs and the cue “*Now translate: English: {source} Fon:*”. The English-to-Hausa instruction was identical except that it named Hausa and requested standard Boko (Latin) orthography with proper diacritics. The five in-context example pairs per language were fixed across all models and scales and were drawn from religious-domain parallel text.

For reproducibility we record the exact model snapshots used: GPT-4o mini (gpt-4o-mini-2024-07-18), Claude Sonnet 4 (claude-sonnet-4-20250514), and Gemini 2.5 Flash (gemini-2.5-flash), all accessed through the CMU AI Gateway (a LiteLLM proxy). Decoding used temperature 0 with a 128-token limit for all systems. COMET used Unbabel/wmt22-comet-da and BERTScore used xlm-roberta-large. Sampling used a fixed random seed of 42 at each scale.

B Representation-Probe Procedure

For the diagnostics in Table 6 we take a sample of up to 100 target-side reference sentences per language. Each sentence is encoded with XLM-R-large, and we form a sentence vector by mean-pooling the final hidden layer over the sentence’s tokens. Within-language cosine similarity is the mean cosine similarity over all off-diagonal pairs of these sentence vectors within a language, and

the average embedding norm is the mean L2 norm of the sentence vectors. The OOV rate is a heuristic proxy: the fraction of tokenizer pieces that are not marked as sub-word continuations, which approximates how often the tokenizer fragments a word rather than matching a known sub-word, and we report it as an indicator rather than a strict vocabulary-coverage figure.

C Data and Code Availability

The parallel corpora are redistributed under their original licenses (CCMatrix and NLLB via OPUS). Our evaluation code, prompts, and the sampling seeds accompany this paper so that the benchmark and the representation probe can be reproduced.