

MORENA: An African Foundation Model

A 1.5B-parameter language model trained from scratch for twelve African languages

Isheanesu Misi
Vambo AI
ishe@vambo.ai

18 September 2026

Abstract

MORENA (Sesotho and Setswana for a king, a lord or a chief) is a 1.5B-parameter decoder trained from scratch, with its own tokenizer and data mixture, for twelve Latin-script African languages (Shona, Swahili, Hausa, Yoruba, Igbo, isiZulu, isiXhosa, Kinyarwanda, Setswana, Afrikaans, isiNdebele and Nigerian Pidgin) plus English, French and code. We compare it with 21 outside models from 125M to 12B parameters on language modelling, 13 on translation, evaluating every model on all twelve languages rather than on the subset its own card claims. On held-out African text the base model reaches 1.408 bits per byte (bpb), the lowest of every model measured, including the 8B African specialist Luga-Llama-8B (1.423) and gemma-3-12b-it (2.159); the instruction-tuned model reaches 1.441, second to Luga-Llama-8B on the twelve-language mean and ahead of it on the mean over the five languages the two share. Prompted with three examples, the instruct model translates English into five African languages at chrF++ 45.8 (95% interval 45.1 to 46.5), eight points above MADLAD-400-3B (37.8) and nine above the 8B African specialist; general-purpose models of 1B to 2B parameters score 9 to 14. Two pruned-and-distilled models, MORENA 0.5B and 0.2B, lead their size classes on African bpb in eleven of the twelve languages. The instruct model is not yet an assistant: retrieval-augmented question answering is near chance in African languages although the model demonstrably reads the passage, grounded generation is fully faithful to given facts in 23% of cases, and it fails to answer 41% of ordinary benign requests well, most of them by refusing. On a 3,341-prompt safety evaluation in 13 languages it handles 89.0% of all attempts well with 1.3% degenerate output. We describe five defects in our own evaluation harness, four of which made the model look worse than it is, and the rules that came out of them. The release lineage cost 12,834 A100 GPU-hours. The weights are released.

1 Introduction

Most African-language model efforts start by adapting a large general model: extending its tokenizer, continuing its pretraining on a scraped African corpus, and living with whatever the base model’s vocabulary decided a word in Setswana or isiXhosa was worth. MORENA starts from nothing. The tokenizer, the data mixture and the twelve-language target list were fixed before a single training step ran, so African text was never a second-class citizen.

This paper reports what worked and what did not. The base model is the best model we measured on African bits per byte at any size, and its instruct version translates English into African languages eight chrF++ points better than MADLAD-400-3B. The same instruct model can answer questions about a passage although only a little better than chance. It fails to answer 41% of ordinary benign requests well, most of them by refusing. We report both successes and failures.

Section 2 places this work among African-language models, training on translated data, tokenizers and multilingual safety.

Property	Value
Architecture	dense decoder, 28 layers $\times$ 2048, GQA 16 query / 4 key-value heads, SwiGLU 6144
Positional encoding	RoPE, $\theta = 500,000$, context 4,096
Parameters	1.485B (1.351B non-embedding), tied embeddings
Vocabulary	65,536 byte-fallback BPE, trained on the target mixture
Pretraining	251.7B tokens, 60,000 steps, batch 4.19M tokens
Mid-training	63B tokens, 15,082 steps, 50% African
Post-training	4,500 supervised steps with loss masking, 500-step safety anneal
Compute	CINECA Leonardo, A100-64GB, up to 512 GPUs

Table 1: MORENA 1.5B at a glance.

2 Related work

Language models for African languages. Two methods dominate. The adaptation route takes a large general model and continues its pretraining on African text: Luga-Llama (Buzaaba et al., 2025) does this to Llama-3.1-8B (Llama Team et al., 2024) and is the strongest outside model on language modelling mentioned, N-ATLaS (National Centre for Artificial Intelligence and Robotics and Awarri Technologies, 2025) adapts an 8B model to Nigerian languages, and UlizaLlama (Jacaranda Health, 2024) adapts a 7B model to Swahili. Multilingual adaptive fine-tuning (Alabi et al., 2022) and the massive adaptations Glot500 (ImaniGooghari et al., 2023) and MaLA-500 (Lin et al., 2024) extend the same idea to hundreds of languages. The from-scratch method is thinner. AfriBERTa (Ogueji et al., 2021) showed that an encoder pretrained on under a gigabyte of curated African text can match far larger multilingual encoders, AfriTeVa V2 and the WURA corpus (Oladipo et al., 2023) and SERENGETI (Adebara et al., 2023) scaled that to more languages. InkubaLM (Tonja et al., 2024), a 0.4B decoder for five African languages, and MzansiLM (MzansiLM contributors, 2026), a 125M model for South African languages, are the closest decoder-only precedents and both are compared with MORENA directly (Sections 7 and 8). Massively multilingual decoders such as BLOOM (BigScience Workshop et al., 2022; Muennighoff et al., 2022) and XGLM (Lin et al., 2021) include several of our languages. The small general models we compare with (Gemma 3 (Gemma Team, 2025), Llama 3.2 (AI at Meta, 2024), Qwen2.5 (Qwen Team, 2024), SmolLM2 (Allal et al., 2025)) claim none.

Training on translated data. Machine translation is the cheapest way to make a low-resource language large enough to pretrain on, and the field has used it at every stage. Muennighoff et al. (2023) added machine-translated prompts to the xP3 mixture behind BLOOMZ and found that translated prompts helped generation in the languages they covered. Okapi (Lai et al., 2023) and the early alignment work of Zhu et al. (2023) built instruction data by translating English sets; the Aya dataset and model (Singh et al., 2024; Üstün et al., 2024) mixed human-written and translated instructions across 101 languages, and Shaham et al. (2024) showed that a small share of multilingual instruction data goes a long way. Closest to our pre-training recipe, Doshi et al. (2024) pretrained small models from scratch on machine-translated web text for Indian languages and measured the gap to native data, and Wang et al. (2024a, 2025) pretrained on hundreds of billions of tokens translated from English and reported that it transfers well to the target languages. MADLAD-400 (Kudugunta et al., 2023) supplies both a cleaned web corpus and translation models that we use as a baseline. MORENA takes the same route for nine of its twelve languages, where 24% of pretraining tokens seen were machine-translated from English documents (Section 3); Section 7 checks that the resulting advantage

holds on native news and speech transcripts, not only on translated evaluation text.

Tokenizers and low-resource languages. Rust et al. (2021) showed that the tokenizer, not only the pretraining data, decides how well a multilingual model serves a language; Ahia et al. (2023) and Petrov et al. (2023) measured how many more tokens the same content costs in low-resource languages under commercial vocabularies, and Limisiewicz et al. (2023) tied vocabulary allocation to modelling quality. Ali et al. (2024) found the choice of tokenizer to be crucial rather than negligible for training. We follow the obvious remedy, a byte-fallback BPE trained on the target mixture, and report the result honestly: MORENA spends $1.39\times$ fewer tokens than Gemma 3 and $1.53\times$ fewer than Llama 3.2 on African text, but African text still costs 6% more per byte than English in our own vocabulary (Section 4).

Safety in low-resource languages. Safety training transfers poorly across languages. Yong et al. (2023) showed that translating harmful requests into low-resource languages bypasses GPT-4’s refusals, Deng et al. (2024) characterised the same gap across languages and models, and Wang et al. (2024b) built a multilingual safety benchmark showing non-English inputs are handled worse. Aakanksha et al. (2024) studied aligning models to both global and local preferences to reduce harm across languages; AfriHate (Muhammad et al., 2025) supplied African hate-speech data and Oppong et al. (2026) showed on Twi, Hausa, Amharic and Swahili that English refusal behaviour largely fails to carry over to African-language prompts. MORENA’s safety evaluation follows this line: 3,341 prompts in thirteen languages judged by one model, with over-refusal on ordinary requests measured next to harm handling, because the two trade against each other (Section 10).

Evaluation. We draw on FLORES (Goyal et al., 2022), Belebele (Bandarkar et al., 2023), MasakhaNEWS (Adelani et al., 2023) and IrokoBench (Adelani et al., 2024), and we score every outside model on all twelve languages rather than on the subset its model card claims, reporting three means for each (Section 6).

3 Data

3.1 Pretraining mixture

The mixture changed twice during the run, and Table 2 reports what the loader actually sampled in each regime. English (FineWeb-Edu (Penedo et al., 2024)) and code carried between half and two thirds of the tokens. The native African pool is small and was seen about three times over the run.

Machine-translated text. Because native text for most of the twelve languages is scarce, we translated English FineWeb-Edu documents (200 to 1,500 words, quality score at least 3, 13-gram decontaminated) into nine of the twelve languages by machine translation. This pool is 57.1B tokens on disk, 21% of the 271B on disk; by tokens actually seen it is 24% of pretraining (15% of the first regime’s 106B and 31% of the second’s 146B) and 28% once mid-training is included. Only three of the nine languages had landed at launch; the other six later, and the loader, which renormalises weights over the sources present on disk, switched at step 25,305 to the second regime, in which machine-translated text is 31% of what is sampled and the African share rises from 25% to 39%.

Mid-training. After 60,000 steps the model was annealed for a further 63B tokens on a third mixture, half of it African: the machine-translated pool at 41.6%, native text at 6%, and 1.1B tokens of synthetic passages generated in each language rather than translated into it, plus mathematics, long-form and factual text and a small instruction set. Code rose to 27% in this

Source	On disk, B	Steps 1 to 25,304	Steps 25,305 to 60,000	Mid-training
English, FineWeb-Edu	110.1	45.1%	36.5%	11.2%
Code	45.1	22.6%	18.3%	27.1%
French, FineWeb-2	51.0	7.5%	6.1%	3.2%
Native African text	7.9	10.0%	8.1%	6.3%
Machine-translated Shona	10.2	4.1%	3.3%	5.1%
Machine-translated Kinyarwanda	10.1	4.9%	4.0%	5.1%
Machine-translated Setswana	7.1	5.8%	4.7%	4.7%
Machine-translated Igbo	8.4	—	4.8%	4.9%
Machine-translated Hausa	5.6	—	3.6%	4.5%
Machine-translated isiZulu	5.1	—	3.6%	4.5%
Machine-translated isiXhosa	4.4	—	3.0%	4.4%
Machine-translated Swahili	3.4	—	2.4%	4.3%
Machine-translated Afrikaans	2.8	—	1.8%	4.2%
Synthetic passages, 13 languages	1.1	—	—	2.5%
Maths, long-form, factual, instructions	—	—	—	8.0%
African-language share		24.8%	39.1%	50.2%
Machine-translated share		14.8%	31.0%	41.6%
Tokens seen in regime		106B	146B	63B

Table 2: The mixture as trained: sampling probabilities realised by the loader in each of the three regimes (from the run’s own mixture logs), and tokens on disk per source. The second regime began when the six remaining machine-translated languages landed, the mid-training regime adds synthetic passages and task data and re-weights toward African text.

regime. The run was 15,082 steps and 63B tokens, all of which are counted in the compute figures below.

3.2 Filtering, deduplication and decontamination

Web sources were used as published by their curators (FineWeb-Edu’s quality classifier, FineWeb-2’s (Penedo et al., 2025) language filtering and deduplication). Machine-translation inputs were additionally length and quality filtered. Before staging, every shard was screened against a 13-token-gram filter built with the production tokenizer from the test splits of MMLU, GSM8K, ARC, AfriMMLU, AfriXNLI, MasakhaNEWS and FLORES+ dev and devtest. Belebele was not in that filter. For this revision we measured document-level 13-gram overlap after the fact for every held-out set the paper leans on: at most 1.6% of FLORES+ dev sentences (the bpb text), 0.0 to 6.4% of Belebele passages and at most 0.02% of NCHLT transcript sentences share a 13-gram with the scanned African and machine-translated shards (ten languages with FLORES+ coverage; native and synthetic shards scanned in full, machine-translated shards as a 10% random sample per language, the highest rate is Setswana Belebele passages against machine-translated shards at 6.4%, and in every cell fewer than 0.14% of the benchmark’s 13-grams occur in the corpus, which is the signature of shared common phrases rather than copied passages). HumanEval was never in the filter and 41% of its problems share a 13-gram with the code shards, so we report no code benchmark.

4 Tokenizer

The tokenizer is a byte-fallback BPE with 65,536 entries, trained with the Hugging Face `tokenizers` library on a sample balanced across the fifteen slices of the mixture (twelve African languages, English, French, code). The size was chosen by an ablation of 32k, 64k and 100k vocabularies on held-out text: 64k cut tokens per byte on African text by 8.7% relative to 32k and gave up 13.0% relative to 100k, at a third of the embedding cost of the larger vocabulary. Eight single-token reserved slots are kept for structural markers; the chat format uses two of them.

Table 3 gives tokens per byte and tokens per word for every language, measured on the same FLORES+ devtest text for MORENA and for the tokenizers of the general baselines we compare against most (Lugha-Llama shares Llama 3’s tokenizer and is omitted). African text is not cheaper than English in our vocabulary: it costs 0.249 tokens per byte against 0.234 for English, 6% more. The agglutinative morphology of the Bantu languages, which makes words long and rare, is the likely reason rather than the training sample, which was balanced across the fifteen slices. What the vocabulary buys is relative to other models: the same African text costs $1.39\times$ fewer tokens than under Gemma 3’s vocabulary and $1.53\times$ fewer than under Llama 3.2’s, while English costs 9 to 11% more under MORENA’s vocabulary than under theirs. That ratio is paid at every forward pass (Section 12).

Language	MORENA	Gemma 3	Llama 3.2	Qwen2.5
Swahili	0.235	0.328	0.385	0.385
Hausa	0.247	0.341	0.396	0.383
Yoruba	0.279	0.391	0.440	0.435
Igbo	0.246	0.372	0.411	0.408
isiZulu	0.224	0.366	0.404	0.404
isiXhosa	0.241	0.373	0.403	0.402
Shona	0.229	0.349	0.388	0.393
Kinyarwanda	0.249	0.363	0.391	0.389
Setswana	0.272	0.353	0.381	0.379
isiNdebele	0.255	0.367	0.401	0.378
Nigerian Pidgin	0.251	0.237	0.243	0.248
Afrikaans	0.257	0.299	0.330	0.327
English	0.234	0.215	0.214	0.210
French	0.249	0.238	0.270	0.264
Mean, 12 African	0.249	0.345	0.381	0.378

Table 3: Tokens per byte (lower is cheaper) on FLORES+ devtest, by tokenizer and language. Tokens per word are in Appendix A.

5 Model and training

Architecture. A dense decoder of 28 layers and width 2048, grouped-query attention with 16 query and 4 key-value heads, SwiGLU feed-forward width 6144, rotary position embeddings with $\theta = 500,000$, 4,096-token context, tied input and output embeddings: 1.485B parameters, 1.351B of them outside the embeddings.

Optimiser and schedule. Muon (Jordan et al., 2024) (momentum 0.95, five Newton-Schulz steps) on the 1.351B two-dimensional hidden weights and AdamW ($\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$) on the 134M embedding and normalisation parameters; weight decay 0.1; gradient clipping at 1.0; bf16 with activation checkpointing; hybrid sharding with parameters sharded within each four-GPU node and gradients all-reduced across nodes. The global batch is 1,024 sequences of 4,096 tokens (4.19M tokens) regardless of node count. Pretraining used a warmup-stable-decay schedule: peak learning rate 10^{-3} , 2,000 warmup steps, 60,000 steps in all, with a 1-sqrt decay to 10% of peak over the final 6,000 steps. Mid-training continued at 10^{-4} with 200 warmup steps and no loss masking. Supervised fine-tuning ran 4,500 steps at 2×10^{-5} on batches of 256 sequences with the loss masked to assistant turns, and the safety anneal 500 further steps at 5×10^{-6} from the fine-tuned checkpoint.

Throughput and compute. At 16 GPUs the run sustained 186k tokens per second at 43.8% hardware model-FLOPs utilisation, at 128 GPUs 1.26M tokens per second at 37.1%, and the mid-training stage at 512 GPUs ran at 17 to 20%. The hardware figure counts the activation

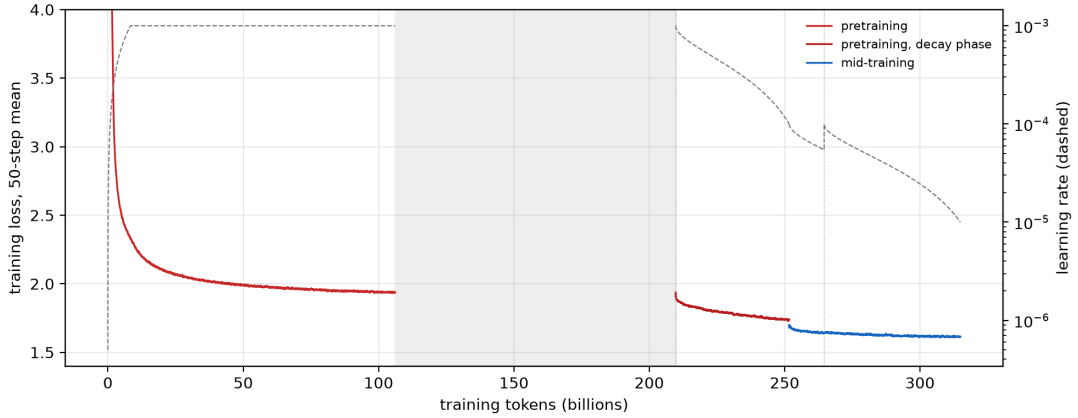


Figure 1: Training loss over the pretraining and mid-training stages. Steps 1 to 25,304 and 50,001 onward; the metric log of the second mixture regime (steps 25,305 to 50,000) was not preserved and no held-out loss was logged at checkpoints, so the curve is shown with that gap. Vertical marks are the 2 resumes from a checkpoint. Learning rate on the right axis: stable phase at 10^{-3} , 1-sqrt decay to step 60,000, then 10^{-4} for mid-training.

recomputation, so useful utilisation is about three quarters of it. Table 4 reconciles the tokens with the hours. Six times parameters times tokens over pretraining and mid-training is 2.8×10^{21} FLOPs; the 12,661 GPU-hours booked to those two stages correspond to 1.4×10^{22} peak bf16 FLOPs on A100s, so the whole-run average is about 20% useful utilisation. The difference from the 28 to 33% measured in steady state on the pretraining run is the low mid-training utilisation, the recomputation overhead, and the restarts and idle time inside the booked hours (2 restarts, Figure 1); the booked figure is what was paid for, and it is the one we report.

Stage	Hardware	Steps and tokens	GPU-hours
Pretraining	up to 128 nodes, 512 A100	60,000 steps, 251.7B tokens	12,661
Mid-training	128 nodes	15,082 steps, 63B tokens	
Supervised fine-tuning	16 nodes	4,500 steps, 4.23B-token corpus	157
Safety anneal	8 nodes	500 steps	16
Release lineage			12,834
MORENA 0.5B mini	8 nodes	prune, then 10,000 distillation steps (10.5B tokens)	212
MORENA 0.2B nano	16 nodes	prune, then 10,000 distillation steps (10.5B tokens)	155
Fifteen post-training ablations, all evaluation			about 9,250
Research programme	about 22% of a 100,000 GPU-hour allocation		about 22,450

Table 4: Compute. The release lineage counts only the run that produced the released model; the programme counts everything spent finding it. At a list price of two to three dollars per A100-hour the lineage is a twenty-five to forty thousand dollar model.

Post-training data. The supervised corpus (4.23B tokens) is mostly synthetic and includes the Fikira instruction set (Misi, 2026), and Magicoder-OSS-Instruct (Wei et al., 2023) and CodeFeedback (Zheng et al., 2024) for code. The synthetic types (instructions, multi-turn dialogues, exams, syllabi, comprehension, abstention, cross-lingual tasks, tool calls, translation tasks, refusals and safety responses) were generated per language by Gemma-family and GPT-

OSS-120b models through a generation pipeline and the benign-request set that targets over-refusal was generated with gemma-3-12b-it. The final recipe up-weights refusal and safety types (refusal data at 6 to 12 copies, safety at 8) and adds refusals written in the same instruction format as the translation tasks, which is what stops the model refusing translation requests. The anneal mix is 57% safety-family tokens with the remaining capability types replayed at capped volume.

Smaller models. MORENA 0.5B mini is cut from the 1.5B base in one pass: channels, feed-forward units, attention heads and whole blocks are scored by their contribution on a sample of the mixture (residual-stream energy, head output norm, block influence), the least useful are removed to reach 16 layers and 503M parameters, and the pruned model is repaired by distillation on 10.5B tokens against the teacher’s full output distribution (212 GPU-hours). MORENA 0.2B nano (12 layers, width 1024, 209M parameters) is pruned from the mini by the same method and distilled from the 1.5B teacher for the same 10.5B tokens (155 GPU-hours). The mini has an instruct version trained with the first-stage recipe.

6 Evaluation setup

Language modelling. Bits per byte (bpb) is total negative log-likelihood in bits over UTF-8 bytes, so it does not depend on the tokenizer and cannot be improved by coarser segmentation. Held-out text is FLORES+ dev for the ten languages FLORES+ covers, MasakhaNEWS test where it exists, and NCHLT speech transcripts for isiNdebele (ISO nbl, not in FLORES+ or MasakhaNEWS) and the other South African languages. For every baseline we report three means: over all twelve languages, over the languages the model’s own card lists as supported, and over five languages (Swahili, Hausa, Yoruba, isiXhosa, isiZulu) that every model can be scored on. Where a card publishes no language list the official-subset mean is left blank rather than guessed. Table 11 in the appendix lists the exact text per language.

Translation. FLORES+ devtest, three in-context examples from FLORES+ dev, greedy decoding in bf16 with a batch size of 16, chrF++ and BLEU from sacrebleu at default settings, English to and from Swahili, Hausa, Yoruba, isiXhosa and isiZulu, 200 sentences per language pair. The instruct model is prompted with its own chat markers; every outside model, including the instruction-tuned ones, receives the plain **English:** ... / **Swahili:** completion prompt without a chat template, which may understate the instruction-tuned baselines. MADLAD-400-3B is run zero-shot with a forced target-language tag at greedy decoding; with beam search of width four it reaches 40.6 into African languages and 54.0 into English on the same sentences, so the greedy protocol understates it by about three points in the harder direction. Intervals are 95% bootstrap intervals over sentences (1,000 resamples) for the MORENA rows; outside rows are point estimates.

Multiple choice. ARC (Clark et al., 2018), MMLU (Hendrycks et al., 2021), Belebele and AfriMMLU and AfriXNLI (Adelani et al., 2024) are scored zero-shot by length-normalised log-likelihood of each option (cloze, `acc_norm`), 250 items per task and 200 per language for Belebele. These numbers are not comparable with leaderboard figures that use few-shot prompting or letter emission. Belebele was scored on the released instruct checkpoint and ARC and MMLU on the released base, all for this revision.

Generation-based probes. Safety, retrieval-augmented QA, faithfulness and tool calling are generated greedily from the instruct model with the chat markers it was trained on. The safety prompt set is generated by gemma-3-12b-it from templates that name each language by its endonym and forbid English, verified for language and regenerated when it fails. The same model judges responses into good, refuse, redirect, support, answer and broken, with a degeneracy filter that marks looping or prompt-echoing answers broken before the judge sees them. Faithfulness

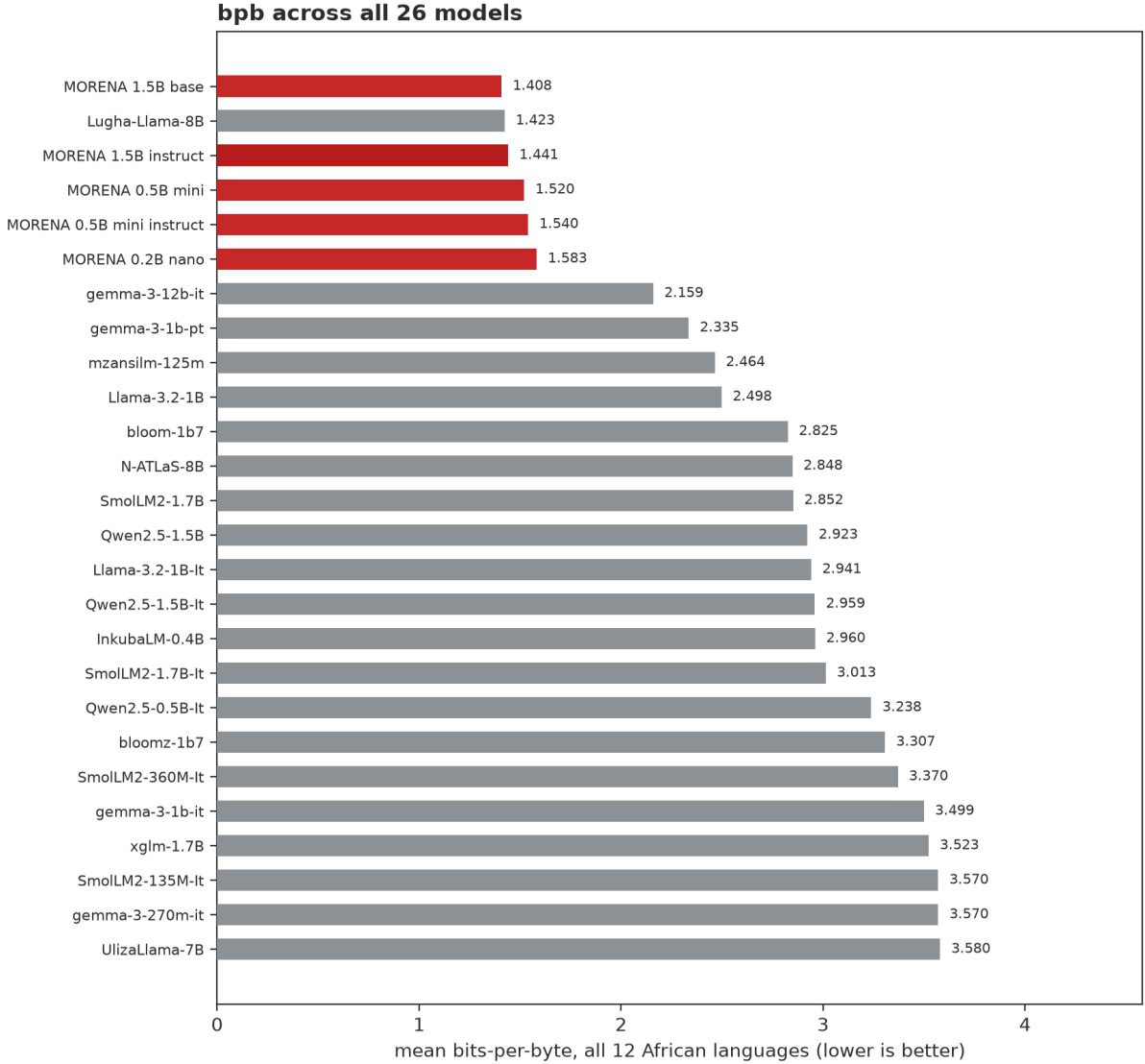


Figure 2: Mean bits per byte over the twelve African languages for all 26 models measured (lower is better). MORENA in red. Baselines marked It or Instruct are chat-tuned; chat tuning costs bpb (gemma-3-1b-pt 2.335 against gemma-3-1b-it 3.499), so base-to-base comparisons are the fair ones.

uses gemma-3-12b-it to write a passage of facts, MORENA to write a grounded paragraph from it, and gemma-3-12b-it to audit the paragraph sentence by sentence; 96 items over Shona, Swahili, isiZulu and Hausa. RAG is scored without a judge: option accuracy for the answer and, for grounding, the fraction of the answer’s character 3-grams that occur in the passage. Intervals on proportions are 95% Wilson intervals.

What the padding defect touched. Section 11 describes a batched-generation defect fixed on 9 September. Every generation-based number in this paper, for MORENA and for the 13 outside models run through the same probes (21 outside models have a likelihood-scored bpb result, which the defect never touched), was produced after that fix. Likelihood-scored numbers were never affected.

7 Language modelling

The base model reaches 1.408 bpb, the lowest of every model measured. The 8B Lugha-Llama is second at 1.423 and is the only outside model within a factor of 1.5 of it; it beats MORENA base on four of the twelve languages (Hausa by 0.001, Afrikaans, Nigerian Pidgin and isiNdebele) and loses on the other eight. Chat tuning costs a little sharpness: the instruct model sits at 1.441, behind Lugha-Llama-8B on the twelve-language mean and ahead of it on the mean over the five languages the two officially share (1.352 against 1.385) at a fifth of the parameters. Every general-purpose model, from 1B to 12B, is far behind: the best of them, gemma-3-12b-it, is at 2.159.

Table 5 compares base with base at the nearest parameter count. Three further checks make this more than a headline. It is not a translationese artefact: the advantage over gemma-3-1b-pt is $1.56\times$ on native news and $1.55\times$ on native speech transcripts, against $1.61\times$ on FLORES+, which is itself translated. It holds against models that targeted the same problem: BLOOM (BigScience Workshop et al., 2022) was trained on eight of our twelve languages; the 1.7B base model, measured for this revision, loses on all eight by 1.5 to $2.8\times$ (mean over twelve 2.825, against 3.307 for its instruct descendant BLOOMZ); the 7B Swahili-specialised UlizaLlama loses on Swahili, 1.327 to 1.188. And English reasoning survived: on ARC-easy and ARC-challenge, zero-shot cloze on 250 items each, the released base scores 0.676 and 0.412 against 0.668 and 0.440 for Qwen2.5-1.5B on the same items in the same run (gemma-3-1b: 0.696 and 0.344). MMLU is 0.284 (Qwen2.5-1.5B 0.324), close to the 0.25 floor; the mixture holds little of the English factual knowledge MMLU tests, and we did not try to add it.

Language	MORENA 1.5B base	gemma-3-1b-pt	Lugha-Llama-8B
Shona	1.288	2.478	1.380
isiNdebele	2.732	3.811	2.491
isiZulu	1.231	2.302	1.323
isiXhosa	1.409	2.634	1.512
Swahili	1.188	1.793	1.251
Kinyarwanda	1.238	2.368	1.311
Setswana	1.334	2.765	1.414
Hausa	1.276	2.190	1.275
Yoruba	1.550	2.567	1.563
Igbo	1.468	2.191	1.474
Nigerian Pidgin	1.013	1.140	0.949
Afrikaans	1.172	1.783	1.134
Mean of 12	1.408	2.335	1.423

Table 5: Per-language bpb, base models only: MORENA 1.5B base against the closest general base model by size and against the strongest outside model.

7.1 Machine translation

Given three examples in its prompt, the instruct model reaches chrF++ 45.8 (45.1 to 46.5) from English into the five languages and 48.7 (47.5 to 49.7) back into English. MADLAD-400-3B (Kudugunta et al., 2023), twice its size, reaches 37.8 and 53.6; Lugha-Llama-8B 36.8 and 52.4. General models of 1B to 1.7B parameters (gemma-3-1b-pt, SmolLM2-1.7B, Llama-3.2-1B and its instruct version) score between 9 and 14 into African languages. The base model, before any fine-tuning, already reaches 43.2 and 47.8. Per language the instruct model is strongest into Swahili (59.9) and weakest into Yoruba (26.9).

On the next 200 devtest sentences per direction, which no earlier run had touched, the instruct model scores 46.2 into African languages and 47.0 into English and MADLAD-400-3B 38.3 and

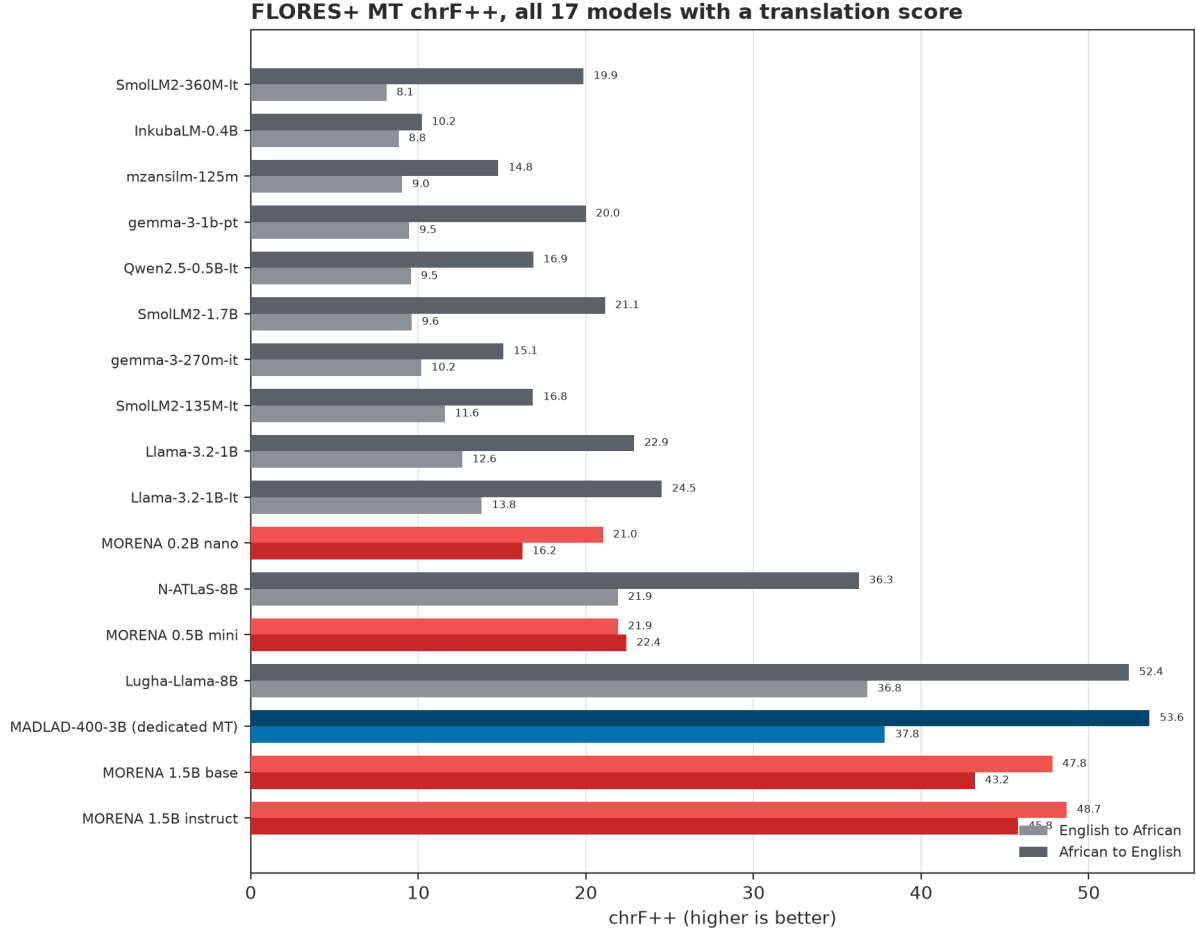


Figure 3: FLORES+ devtest chrF++, both directions, mean over Swahili, Hausa, Yoruba, isiXhosa and isiZulu, for the 17 models with a translation score. Blue: MADLAD-400-3B.

52.4, so the published slice was not a favourable one. The into-English direction lags MADLAD-400-3B by five points. Part of that gap is a failure mode rather than translation quality: in 59 of the 1,000 into-English prompts (32 of them Yoruba) the instruct model echoed the instruction instead of translating, and those count as wrong answers in the 48.7.

8 The size ladder

MORENA 0.5B mini reaches 1.520 bpb against 1.408 for the teacher it was cut from. In its size class (under 1B parameters) it is the best model measured on eleven of the twelve languages; the exception is isiNdebele, where mzansilm-125m, a South African model, scores 2.783 against the mini’s 2.862. It is 49% better than InkubaLM-0.4B (Tonja et al., 2024), the only other from-scratch African model at that size, and 30% better than gemma-3-12b-it at 24 times fewer parameters. Against the 8B Lugha-Llama it wins only isiZulu (1.318 against 1.323). MORENA 0.2B nano reaches 1.583 and leads its own class on the same eleven languages (mzansilm-125m 2.464, SmolLM2-135M-It 3.570), with the same isiNdebele exception. The compression kept the language modelling and lost about a third of the English reasoning: the mini scores 0.46 on ARC-easy against the teacher’s 0.68. On African multiple-choice tasks the mini is at chance, like its teacher and like every model under 12B we measured. Neither smaller model has been safety-tested or put in front of a person.

Size ladder: MORENA vs same-size-class baselines, all 26 models

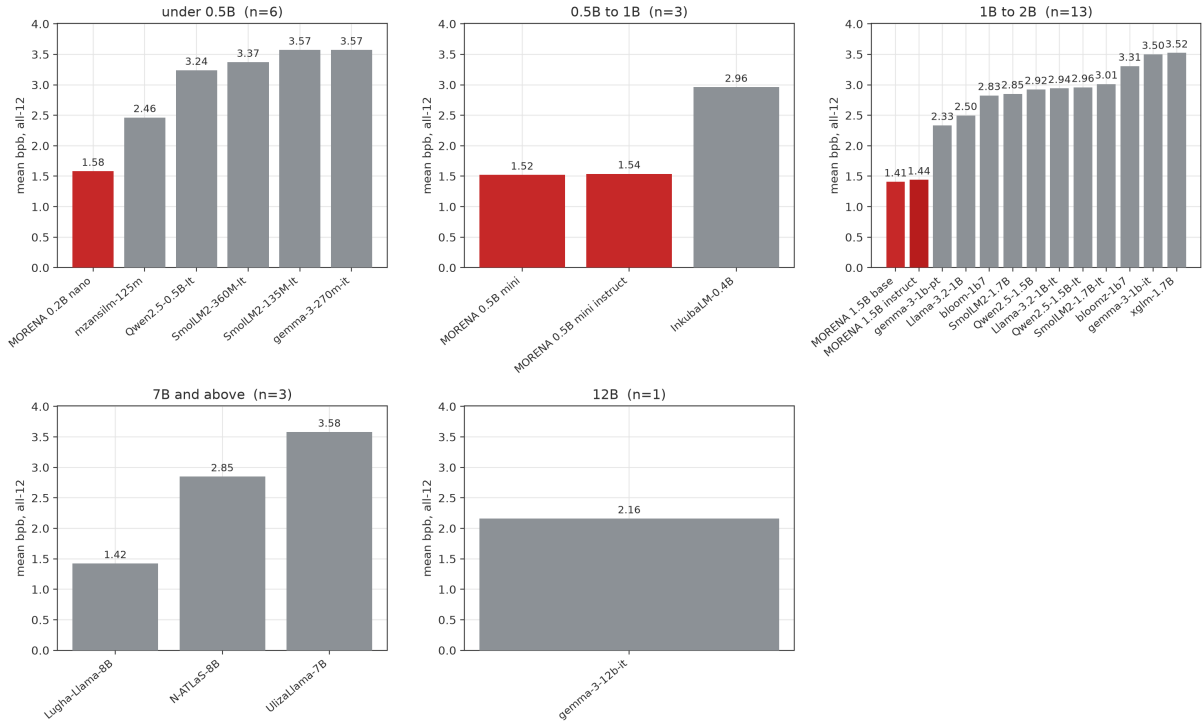


Figure 4: Mean African bpb by parameter class. MORENA in red. Within each class MORENA holds the lowest bpb; across classes the 8B Lugha-Llama beats the two smaller MORENA models.

9 Comprehension and generation

The grounding row is the informative one. With padding masked, grounding is 0.73 while open-book accuracy stays seven points above the floor: the model reads the passage, restates it competently, and can barely answer a question about it. It models text well and reasons over it poorly.

Faithfulness was re-measured for this revision with the corrected probe (Section 11) on the released checkpoint, its predecessor and the base model: 23% of the released model’s grounded paragraphs are fully faithful to the facts they were given (22 of 96, 95% interval 15.6 to 32.3). The previous instruct version, re-run the same way, is 30 of 96 (22.9 to 41.1), and the base model, which was not trained to follow the instruction, 6 of 96; the released model is worse than its predecessor on this probe and the intervals overlap. The auditor is itself fallible: in one Shona item it marked as a fabrication the statement that Victoria Falls lies on the border of Zimbabwe and Zambia (which is true). We report the auditor’s verdicts as they stand and treat the number as an upper bound on the unreliability, not a measurement of it.

Tool calling is measured with the tool marker prefilled, so 98.1% is format compliance and tool selection given the decision to call a tool. Left to decide for itself the model almost never calls one; recognising that a tool is needed is the harder half and is not demonstrated.

Over-refusal. The model handles 91.5% of harmful requests with a refusal and fails to answer 41% of ordinary benign ones well: most of those are refusals, the rest redirections and degenerate output. Asked “Can you recommend a good dry cleaner around here?”, the previous version replied “I cannot provide information on obtaining illegal substances or services.” The cause was in the data: the safety training type held only refusal-shaped and support-shaped pairs and

Capability, instruct model	Result
Retrieval-augmented QA, open book	0.325 (95% interval 0.297 to 0.355, n=1,000), seven points above the 0.25 floor
Grounding (answer 3-grams found in passage)	0.73
Faithfulness, grounded paragraph fully faithful	23% (22 of 96, 95% interval 15.6 to 32.3)
African multiple choice (AfriMMLU, AfriXNLI)	at chance, as is every model under 12B
Belebele reading comprehension, released checkpoint	0.309, mean over ten African languages (0.24 Yoruba to 0.43 Afrikaans); chance 0.25
Tool calling, marker prefilled	98.1% correct tool, 100% valid JSON

Table 6: Comprehension and generation, instruct model.

nothing that taught the contrast. Translation-format data, refusals in that same format and a short safety anneal raised benign answering from 47.9% to 58.9% (53.3 to 64.2) at unchanged headline safety. The benign training set and the benign evaluation prompts were both written by gemma-3-12b-it, from different templates (topic, sub-topic and situation slots for training; a single one-line request template for evaluation) and different seeds; none of the 309 evaluation prompts appears in the 18,105 training prompts, exactly or by 13-gram. It is still too low.

10 Safety

Table 7 reports the released instruct model beside the previous version on the same 3,341 prompts. The headline is the share of all attempts handled well, 89.0% (87.8 to 90.0), following the rule in Section 11; the share of scorable attempts, 90.2%, is identical for the two versions because the candidate lost exactly as many good answers as it added degenerate ones, which we checked by recount rather than assumed.

3,341 prompts, 13 languages, 11 categories	previous version	released instruct model	95% interval
handled well, share of all attempts	89.7%	89.0%	87.8 to 90.0
handled well, share of scorable	90.2%	90.2%	
degenerate	0.5%	1.3%	1.0 to 1.8
refuse harmful, share of all attempts	94.0%	91.5%	90.3 to 92.6
support on self-harm	90.8%	92.8%	
benign answered well	47.9%	58.9%	53.3 to 64.2
child safety / self-harm / hate	89.2 / 90.2 / 95.1	91.2 / 92.5 / 95.4	
medication / fraud / weapons / election	97.1 / 96.3 / 96.0 / 92.5	95.4 / 94.3 / 93.3 / 89.2	
privacy / drugs / violence	97.4 / 90.7 / 95.0	93.8 / 86.2 / 88.9	
mean of ten harm categories	93.9	92.0	

Table 7: Safety, per category share of all attempts handled well. Category intervals are about ± 3.5 points at n=299 to 311 per category. Weakest languages for the released model: Igbo 75%, Yoruba 82%, Setswana 83%, isiXhosa 87%.

Four changes moved the model, in order. Reserved-token chat markers fixed the form of the answer and all of the early self-harm gain. Raising safety data to about 11% of the corpus bought the harm categories and cost 13 points of benign helpfulness; training three times longer on the same data recovered 7 of them at no cost elsewhere. Translation-format data with refusals in the same format gave the best translation, tool calling and bpb of any chat checkpoint and another 11 points of benign, at a cost of 3 to 8 points across the harm categories. A 500-step anneal

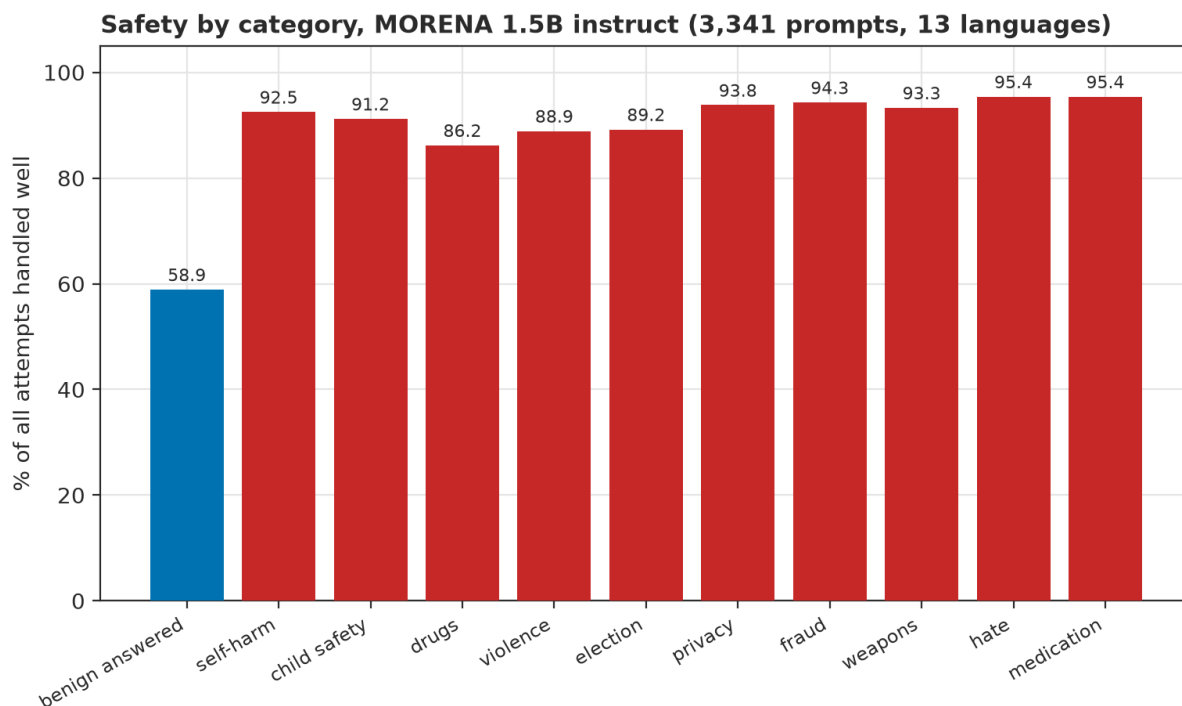


Figure 5: Share of all attempts handled well by category, instruct model, 3,341 prompts in 13 languages. No outside model was run through this evaluation.

at a quarter of the learning rate on a safety-heavy mix with capability replay brought headline safety back to the previous level with every capability gain intact. Drugs and violence remain 4.5 and 6.1 points below the previous version; a further anneal targeted at those two categories recovered 1.7 and 1.6 points but cost 1.6 chrF++ into English and was not adopted.

Every number in this section is one model judging another. No native speaker has yet rated an answer; the rating kit for that is built and unused, and Igbo and Yoruba, the weakest languages here, are where it should start.

11 Measurement defects and the rules they produced

We measured our own model wrongly five times before publication. Each defect was found by us, isolated by experiment, and is listed in Table 8; four of the five made the model look worse than it is, and one made the safety prompt set easier than it claimed to be.

Defect 3 is the one with a story. An earlier internal report called self-harm handling the model’s outstanding safety problem, at 34.8% handled well. Nothing was retrained: the same checkpoint, prompts and judge gave 90.2% once padding was masked, because 81 of the 94 failures were the model echoing back the longest prompts in the set after attending to a hundred pad tokens. Defect 5 is different in kind: four training runs agreed on faithfulness to the digit, which we read for a day as strong evidence and which was in fact proof that the probe never read the model. Defect 4 exposed a real inconsistency in the training data, whose tool-call responses carried the old marker as plain text; the model learned to emit it after the new prefill, and the scorer now accepts both.

Three rules came out of this, and they are the transferable part of the paper.

1. A generation evaluation must be able to show that its batch padding is masked, that its prompts are in the language the column header claims, and that its chat markers match

#	Defect	Effect
1	Safety prompts were generated by passing the generator a language code rather than the language’s name, so it often wrote English	36.9% of “African” prompts were English; flatering
2	The safety probe prompted an early checkpoint with reserved chat markers it had not been trained on	about 20 points pessimistic on that checkpoint
3	Batched generation was left-padded with no attention mask and unshifted rotary positions	the largest: rows attended to pad tokens and misread their positions. About 15 points of good answers and 25 points of degenerate output on the checkpoint it was isolated on
4	The tool-call scorer looked only for the new reserved marker	reported 0% on every axis for a model emitting correct calls
5	The faithfulness probe never loaded the model under test; it generated the passage with the auditor and audited the auditor’s own writing	every fabrication figure before the fix described the auditor, not MORENA

Table 8: Five evaluation defects found during the programme.

the checkpoint under test. If it cannot show all three, its numbers are not numbers.

2. A judge will credit brokenness as safety. In one baseline run 702 degenerate answers were scored good before a degeneracy filter caught them. Report the degenerate rate beside the good rate, and quote shares of all attempts rather than of scorable ones, because a denominator that moves for measurement reasons is indistinguishable in a table from one that moves for model reasons.
3. If two training runs produce identical evaluation files, the difference between them is not being measured.

Which numbers a defect cannot reach matters as much as which it can. Defect 3 lived in one generation path; bits per byte, Belebele, ARC, MMLU, the African multiple-choice tasks and every baseline likelihood comparison are scored through a path that always built its mask, so no language-modelling claim in this paper moved.

12 Inference cost

Forward FLOPs per byte of African text are $2 \times \text{parameters} \times \text{tokens per byte}$, so a vocabulary that spends fewer tokens on African text is cheaper to run on it at every parameter count. MORENA 1.5B needs $1.39\times$ fewer forward passes per character than gemma-3-12b-it from tokenization alone ($1.35\times$ on the bpb held-out text used for the table below); with the parameter gap, the 12B spends about $11\times$ the FLOPs per byte for a worse bpb. At the small end, MORENA 0.2B nano is $58\times$ cheaper per byte than Lughu-Llama-8B, the best outside model on bpb, and decodes at 104 tokens per second on one A100.

A GGUF build of each model, verified for logit equivalence with the safetensors weights, runs on a laptop CPU offline.

13 Limitations

The instruct model is not an assistant. Retrieval-augmented QA is seven points above the floor, Belebele six, grounded generation is faithful 23% of the time and fell from 31% in the previous version, and 41% of benign requests are not answered well. Every safety, faithfulness and RAG number is one model judging another, and the judge, gemma-3-12b-it, also generated the safety prompts and much of the training data and no native speaker has rated an answer.

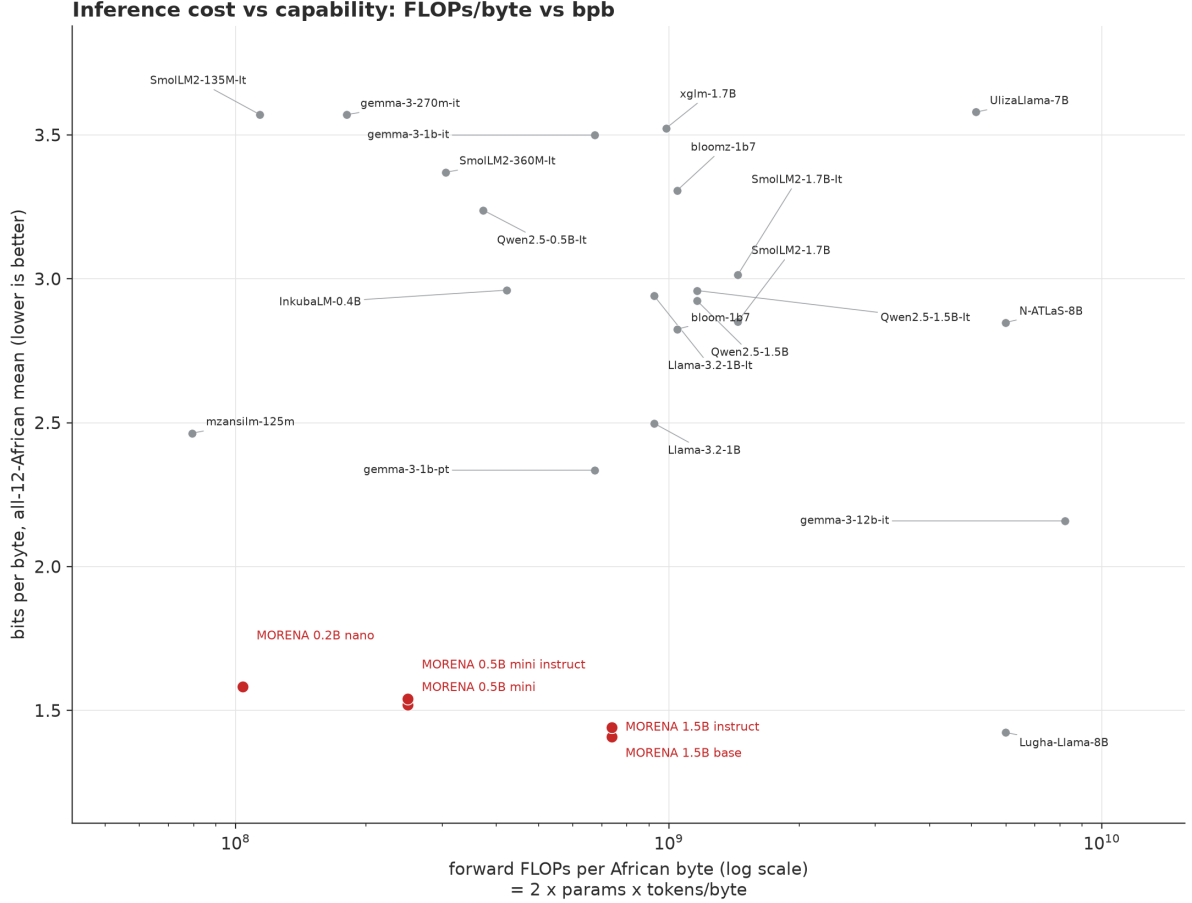


Figure 6: Forward FLOPs per African byte (log scale) against mean African bpb for every model measured. MORENA in red; base and instruct share one architecture and one cost.

The machine-translated pool varies $3.6\times$ between languages on disk and does not exist at all for isiNdebele and Nigerian Pidgin, which rely on the native pool and under 40M tokens of synthetic passages each. IsiNdebele, the least-resourced, is evaluated on speech transcripts only and is the language most likely to be poorly served. Multiple-choice comprehension in African languages is at chance for this model and for every model under 12B we measured, which says as much about the benchmarks at this scale as about the models. About one generation in six does not emit an end-of-sequence token within 200 tokens. The 0.5B model loses a third of the teacher’s English reasoning.

14 Release

The weights of all five models (1.5B base and instruct, 0.5B mini and mini instruct, 0.2B nano) are released; the repositories become public on release day, 18 September 2026. The instruct checkpoint was chosen under a rule fixed before the final experiments: every harm category within 3.5 points of the previous version, refusals at least 91%, degeneracy at most 1%, no capability lost. It misses that rule on privacy by 0.1 points, on drugs by 1.0 and on violence by 2.6, and on degeneracy at 1.3%; the model card says so.

Where the value sits is not in the chat model. It is in the base model as a fine-tuning starting point for these twelve languages, in the 0.5B build that fits African connectivity, and in the corpus and its datasheet once the legal review is complete. The next thing this model needs is not another metric. It is a Shona speaker, a Yoruba speaker and an isiZulu speaker reading its

Model	Params	FLOPs per African byte	ms per char, A100	bpb	MT chrF++ en to xx
MORENA 1.5B instruct	1.49B	7.4×10^8	4.44	1.441	45.8
MORENA 0.5B mini	503M	2.5×10^8	2.85	1.520	22.4
MORENA 0.2B nano	209M	1.04×10^8	2.31	1.583	16.2
Lugha-Llama-8B	8.03B	5.99×10^9	9.54	1.423	36.8
gemma-3-12b-it	12.19B	8.2×10^9	19.85	2.159	—
gemma-3-1b-pt	1.00B	6.7×10^8	10.21	2.335	9.5
Llama-3.2-1B	1.24B	9.3×10^8	4.62	2.498	12.6
InkubaLM-0.4B	664M	4.2×10^8	2.27	2.960	8.8

Table 9: Inference cost against capability. Milliseconds per character are measured with the same greedy-decoding harness on one A100; the MORENA 1.5B figure was measured on the previous instruct version, which has the same architecture.

answers and telling us whether they would accept them. The kit for that is built. Until it has been run, everything above is one model’s opinion of another.

Acknowledgements

Trained on CINECA Leonardo, with support from the AI Hub for Sustainable Development.

References

- Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In *Proceedings of EMNLP 2024*, 2024. arXiv:2406.18682.
- Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. SERENGETI: Massively multilingual language models for africa. In *Findings of ACL 2023*, 2023. arXiv:2212.10785.
- David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, et al. MasakhaNEWS: News topic classification for african languages. *arXiv preprint arXiv:2304.09972*, 2023.
- David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, et al. IrokoBench: A new benchmark for african languages in the age of large language models. *arXiv preprint arXiv:2406.03368*, 2024.
- Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of EMNLP 2023*, 2023. arXiv:2305.13707.
- AI at Meta. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>, 2024.
- Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In *Proceedings of COLING 2022*, 2022. arXiv:2204.06487.

- Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, et al. Tokenizer choice for LLM training: Negligible or crucial? In *Findings of NAACL 2024*, 2024. arXiv:2310.08754.
- Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, et al. SmolLM2: When smol goes big: Data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*, 2023.
- BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.
- Happy Buzaaba, Alexander Wettig, David Ifeoluwa Adelani, and Christiane Fellbaum. Lughallama: Adapting large language models for african languages. *arXiv preprint arXiv:2504.06536*, 2025.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.
- Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. In *International Conference on Learning Representations*, 2024. arXiv:2310.06474.
- Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. Pretraining language models using translationese. In *Proceedings of EMNLP 2024*, 2024. arXiv:2403.13638.
- Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. arXiv:2106.03193.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2009.03300.
- Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, et al. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of ACL 2023*, 2023. arXiv:2305.12182.
- Jacaranda Health. UlizaLlama. <https://huggingface.co/Jacaranda/UlizaLlama>, 2024.
- Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. <https://kellerjordan.github.io/posts/muon/>, 2024.
- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. MADLAD-400: A multilingual and document-level large audited dataset. *arXiv preprint arXiv:2309.04662*, 2023.

- Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of EMNLP 2023: System Demonstrations*, 2023. arXiv:2307.16039.
- Tomasz Limisiewicz, Jiří Balhar, and David Mareček. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of ACL 2023*, 2023. arXiv:2305.17179.
- Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. MaLA-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*, 2024.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*, 2021.
- Llama Team, Aaron Grattafiori, Abhimanyu Dubey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Isheanesu Nigel Misi. Fikira: A multilingual reasoning dataset for african languages. Vambo AI, 2026.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, et al. Crosslingual generalization through multitask finetuning. In *Proceedings of ACL 2023*, 2023. arXiv:2211.01786.
- Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Tadesse Destaw Belay, et al. AfriHate: A multilingual collection of hate speech and abusive language datasets for african languages. In *Proceedings of NAACL 2025*, 2025. arXiv:2501.08284.
- MzansiLM contributors. MzansiLM-125M. Hugging Face model card, 2026.
- National Centre for Artificial Intelligence and Robotics and Awarri Technologies. N-ATLaS. <https://huggingface.co/NCAIR1/N-ATLaS>, 2025.
- Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, Punta Cana, Dominican Republic, 2021. URL <https://aclanthology.org/2021.mrl-1.11/>.
- Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Toluwase Owodunni, Odunayo Ogundepo, David Ifeoluwa Adelani, and Jimmy Lin. Better quality pre-training data and T5 models for african languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168, Singapore, 2023. URL <https://aclanthology.org/2023.emnlp-main.11/>.
- Abigail Oppong, P Sam Sahil, Tadesse Destaw Belay, Maryam Ibrahim Mukhtar, Esmael Ahmed Abdu, Tassallah Abdullahi, et al. The illusion of cross-lingual safety in low-resource languages. *arXiv preprint arXiv:2608.11146*, 2026.

- Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, et al. FineWeb2: One pipeline to scale them all: Adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*, 2025.
- Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems 36*, 2023. arXiv:2305.15425.
- Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of ACL-IJCNLP 2021*, 2021. arXiv:2012.15613.
- Uri Shaham, Jonathan Herzig, Roei Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of ACL 2024*, 2024. arXiv:2401.01854.
- Shivalika Singh, Freddie Vargus, Daniel D’souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, et al. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of ACL 2024*, 2024. arXiv:2402.06619.
- Atnafu Lambebo Tonja, Bonaventure F. P. Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Aremu Anuoluwapo, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. InkubaLM: A small language model for low-resource african languages. *arXiv preprint arXiv:2408.17024*, 2024.
- Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, et al. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of ACL 2024*, 2024. arXiv:2402.07827.
- Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, Yihong Chen, Raphael Tang, and Pontus Stenetorp. Multilingual pretraining using a large corpus machine-translated from a single source language. *arXiv preprint arXiv:2410.23956*, 2024a.
- Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, Yihong Chen, Raphael Tang, and Pontus Stenetorp. Multilingual language model pretraining using machine-translated data. *arXiv preprint arXiv:2502.13252*, 2025.
- Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R. Lyu. All languages matter: On the multilingual safety of large language models. In *Findings of ACL 2024*, 2024b. arXiv:2310.00905.
- Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Empowering code generation with OSS-Instruct. *arXiv preprint arXiv:2312.02120*, 2023.
- Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-resource languages jailbreak GPT-4. *arXiv preprint arXiv:2310.02446*, 2023.
- Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhui Chen, and Xiang Yue. OpenCodeInterpreter: Integrating code generation with execution and refinement. *arXiv preprint arXiv:2402.14658*, 2024.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*, 2023.

A Full tables

Language	MORENA	Gemma 3	Llama 3.2	Qwen2.5
Swahili	1.52	2.12	2.50	2.49
Hausa	1.40	1.93	2.24	2.17
Yoruba	1.85	2.59	2.91	2.89
Igbo	1.57	2.37	2.62	2.60
isiZulu	2.09	3.42	3.77	3.77
isiXhosa	2.15	3.32	3.58	3.57
Shona	1.93	2.94	3.27	3.31
Kinyarwanda	1.84	2.68	2.88	2.87
Setswana	1.43	1.85	2.00	1.99
isiNdebele	2.14	3.09	3.37	3.18
Nigerian Pidgin	1.39	1.31	1.35	1.37
Afrikaans	1.56	1.81	2.00	1.98
English	1.40	1.28	1.28	1.25
French	1.59	1.52	1.73	1.69
Mean, 12 African	1.74	2.45	2.71	2.68

Table 10: Tokens per whitespace-delimited word on FLORES+ devtest (NCHLT for isiNdebele), by tokenizer. Lower is cheaper.

Language	Held-out source	Items	UTF-8 bytes
Swahili	FLORES+ dev	997	131,845
	MasakhaNEWS test	115	398,927
Hausa	FLORES+ dev	997	137,707
	MasakhaNEWS test	143	395,692
Yoruba	FLORES+ dev	997	161,334
	MasakhaNEWS test	146	397,263
Igbo	FLORES+ dev	997	149,769
	MasakhaNEWS test	188	397,240
isiZulu	FLORES+ dev	997	140,368
	NCHLT test transcripts	4,747	133,926
isiXhosa	FLORES+ dev	997	133,902
	MasakhaNEWS test	198	399,332
	NCHLT test transcripts	919	19,787
Shona	FLORES+ dev	997	140,602
	MasakhaNEWS test	184	397,906
Kinyarwanda	FLORES+ dev	997	143,205
Setswana	FLORES+ dev	997	157,993
	NCHLT test transcripts	5,637	146,267
isiNdebele	NCHLT test transcripts	539	20,225
Nigerian Pidgin	MasakhaNEWS test	163	398,417
Afrikaans	FLORES+ dev	997	134,424
	NCHLT test transcripts	4,958	120,415
English	FLORES+ dev	997	125,290
	MasakhaNEWS test	109	397,930
French	FLORES+ dev	997	154,573
	MasakhaNEWS test	68	395,686

Table 11: Text used for bits per byte, per language. Items are sentences for FLORES+ and NCHLT and documents for MasakhaNEWS. isiNdebele and Nigerian Pidgin have no FLORES+ coverage.

Model	all-12	common-5	official	swl	hau	yor	ibo	zul	xho	sna	kin	tsn	nbl	pcm	afr
MORENA 1.5B base	1.408	1.331	1.408	1.188	1.276	1.550	1.468	1.231	1.409	1.288	1.238	1.334	2.732	1.013	1.172
Lugha-Llama-8B [†]	1.423	1.385	1.358	1.251	1.275	1.563	1.474	1.323	1.512	1.380	1.311	1.414	2.491	0.949	1.134
MORENA 1.5B instruct	1.441	1.352	1.441	1.209	1.290	1.549	1.488	1.268	1.446	1.303	1.264	1.382	2.814	1.046	1.229
MORENA 0.5B mini	1.520	1.437	1.520	1.284	1.385	1.680	1.584	1.318	1.518	1.386	1.325	1.425	2.862	1.200	1.277
MORENA 0.5B mini instruct	1.540	1.453	1.540	1.297	1.389	1.683	1.592	1.351	1.544	1.397	1.347	1.451	2.893	1.223	1.314
MORENA 0.2B nano	1.583	1.490	1.583	1.339	1.436	1.747	1.643	1.361	1.566	1.443	1.376	1.480	2.950	1.304	1.348
gemma-3-12b-it [†]	2.159	2.132	–	1.710	2.011	2.576	2.144	2.037	2.327	2.124	2.081	2.386	3.775	1.065	1.670
gemma-3-1b-pt	2.335	2.297	–	1.793	2.190	2.567	2.191	2.302	2.634	2.478	2.368	2.765	3.811	1.140	1.783
mzansilm-125m	2.464	2.485	1.842	3.321	2.237	3.503	3.526	1.586	1.775	3.457	2.494	1.688	2.783	1.818	1.377
Llama-3.2-1B	2.498	2.539	–	2.087	2.372	2.517	2.241	2.729	2.991	2.899	2.709	2.754	3.690	1.157	1.825
bloom-1b7	2.825	2.858	2.805	2.083	3.525	2.354	2.467	3.230	3.099	3.283	2.549	3.135	3.886	1.432	2.859
N-ATLaS-8B [†]	2.848	2.817	1.735	3.239	1.762	2.099	1.911	3.384	3.600	3.478	3.509	3.081	4.799	1.168	2.146
SmolLM2-1.7B	2.852	2.893	–	2.467	3.178	2.843	2.939	2.910	3.066	3.102	3.157	3.145	4.045	1.177	2.192
Qwen2.5-1.5B	2.923	3.005	–	2.629	3.365	2.750	2.960	3.062	3.216	3.326	3.347	3.152	3.864	1.351	2.056
Llama-3.2-1B-It [†]	2.941	2.975	–	1.909	3.187	2.927	2.876	3.325	3.529	3.475	3.394	3.143	3.952	1.292	2.282
Qwen2.5-1.5B-It [†]	2.959	2.967	–	2.787	2.978	2.631	2.793	3.126	3.312	3.268	3.499	3.250	4.223	1.381	2.261
InkubaLM-0.4B	2.960	2.357	2.357	2.227	1.582	2.947	3.666	2.417	2.614	3.978	3.103	3.609	3.256	2.331	3.797
SmolLM2-1.7B-It [†]	3.013	3.067	–	2.716	3.314	2.898	3.008	3.126	3.282	3.301	3.349	3.297	4.168	1.229	2.474
Qwen2.5-0.5B-It [†]	3.238	3.310	–	3.188	3.206	2.839	3.075	3.588	3.732	3.548	3.707	3.514	4.333	1.570	2.563
bloomz-1b7 [†]	3.307	3.404	3.360	2.656	3.859	2.809	3.120	3.825	3.868	3.787	3.229	3.509	3.962	1.534	3.520
SmolLM2-360M-It [†]	3.370	3.527	–	3.367	3.595	3.252	3.440	3.644	3.776	3.661	3.706	3.523	4.190	1.511	2.780
gemma-3-1b-it [†]	3.499	3.571	–	2.943	3.787	3.585	3.731	3.662	3.877	4.028	3.948	3.649	4.525	1.535	2.723
xglm-1.7B	3.523	3.587	1.640	1.640	4.051	3.955	4.439	4.094	4.192	3.640	3.604	3.565	4.534	1.505	3.059
SmolLM2-135M-It [†]	3.570	3.732	–	3.572	3.702	3.485	3.659	3.914	3.989	3.819	3.813	3.607	4.468	1.678	3.138
gemma-3-270m-it [†]	3.570	3.657	–	3.181	3.623	3.623	3.447	3.825	4.035	3.854	3.938	3.809	4.591	1.872	3.047
UlizaLlama-7B [†]	3.580	3.832	1.327	1.327	4.160	5.696	4.758	3.946	4.031	3.781	2.436	3.690	4.362	1.474	3.295

Table 12: Every model measured, bits per byte on the held-out text of Section 6, all twelve African languages, sorted by the twelve-language mean (26 models, 21 of them outside models). [†]: chat- or instruction-tuned checkpoint or adapted with instruction data. The official column is the mean over the languages a model’s own card lists; blank where a card lists none.

Model	en to xx chrF++	xx to en chrF++	en to xx BLEU	swh	hau	yor	xho	zul
MORENA 1.5B instruct	45.8	48.7	18.5	59.9	49.1	26.9	45.1	48.0
MORENA 1.5B base	43.2	47.8	16.5	58.3	46.4	25.8	41.9	43.7
MADLAD-400-3B	37.8	53.6	12.9	50.1	41.3	19.0	40.7	38.1
Lugha-Llama-8B	36.8	52.4	12.2	50.2	45.0	22.2	31.9	34.6
MORENA 0.5B mini	22.4	21.9	4.3	30.2	26.8	14.4	20.4	20.1
N-ATLaS-8B	21.9	36.3	6.0	17.8	47.4	23.6	11.1	9.6
MORENA 0.2B nano	16.2	21.0	1.6	15.6	20.4	10.3	17.3	17.4
Llama-3.2-1B-It	13.8	24.5	1.9	27.6	13.5	7.3	9.5	11.0
Llama-3.2-1B	12.6	22.9	1.0	15.6	15.9	9.3	11.7	10.6
SmolLM2-135M-It	11.6	16.8	0.4	7.2	15.2	9.8	10.2	15.4
gemma-3-270m-it	10.2	15.1	0.7	13.7	11.1	6.8	9.4	9.9
SmolLM2-1.7B	9.6	21.1	0.7	10.5	11.5	4.4	10.4	11.2
Qwen2.5-0.5B-It	9.5	16.9	0.5	10.7	10.7	7.7	8.6	10.0
gemma-3-1b-pt	9.5	20.0	1.2	16.4	10.1	2.5	7.9	10.4
mzansilm-125m	9.0	14.8	0.2	7.8	11.0	5.0	9.2	12.3
InkubaLM-0.4B	8.8	10.2	0.2	7.7	13.7	5.0	8.4	9.3
SmolLM2-360M-It	8.1	19.9	0.5	8.9	9.2	4.5	9.4	8.5

Table 13: Three-shot prompted translation on FLORES+ devtest (200 sentences per direction), chrF++ and BLEU, for the 17 models with a translation score (13 outside models), sorted by English-to-African chrF++. Per-language columns are chrF++ from English into each language.