

Tiny Aya: Bridging Scale and Multilingual Depth

Alejandro R. Salamanca^{**1} Diana Abagyan^{**2} Daniel D'souza¹
 Ammar Khairi² David Mora² Saurabh Dash¹ Viraat Aryabumi²
 Sara Rajae⁺² Mehrnaz Mofakhami¹ Ananya Sahu¹ Thomas Euyang¹
 Brittawnya Prince¹ Madeline Smith¹ Hangyu Lin² Acyr Locatelli²
 Sara Hooker^{†1} Tom Kocmi² Aidan Gomez² Ivan Zhang² Phil Blunsom²
 Nick Frosst² Joelle Pineau² Beyza Ermiş¹ Ahmet Üstün^{§2}
 Julia Kreutzer^{§1} Marzieh Fadaee^{§1}

¹Cohere Labs ²Cohere
 {alejandror, juliakreutzer, marzieh}@cohere.com *

Abstract

We introduce TINY AYA, a family of 3.35 billion parameters open-weight multilingual language models that achieve strong, balanced performance across 70 languages through deliberate data curation and training design. The release includes a pretrained base model, a globally balanced instruction-tuned variant, and three region-specialized models targeting Africa, South Asia, Europe, Asia-Pacific, and West Asia. Our approach combines three key elements: (1) a script-aware tokenizer weighting scheme that achieves competitive or superior compression across most writing systems, particularly for underrepresented scripts; (2) a posttraining pipeline combining FUSION, a multi-teacher synthetic data generation framework, with region-aware supervised finetuning and SIMMERGE, a predictive model merging method that preserves global consistency while retaining regional gains; (3) a comprehensive multilingual evaluation spanning translation, understanding, generation, safety, and cultural awareness across up to 66 languages. Despite its compact size, TINY AYA outperforms GEMMA3-4B on translation in 46 of 55 languages, is competitive with or outperforms similar-sized models on generative benchmarks with substantially lower cross-language variance, and achieves the highest mean safe response rate (91.1%) on multilingual safety benchmarks. Quantized variants run on smartphones and quantization degradation attenuates for lower-resource languages. We release all models to support research into efficient, inclusive multilingual AI.

1 Introduction

Multilingual language modeling has advanced rapidly, yet progress remains uneven across languages. Performance gains often track the distribution of available data, reinforcing

^{**}Equal contribution. [†]Work done during internship. [‡]Now at Adaption Labs. [§]Principal senior advisors.

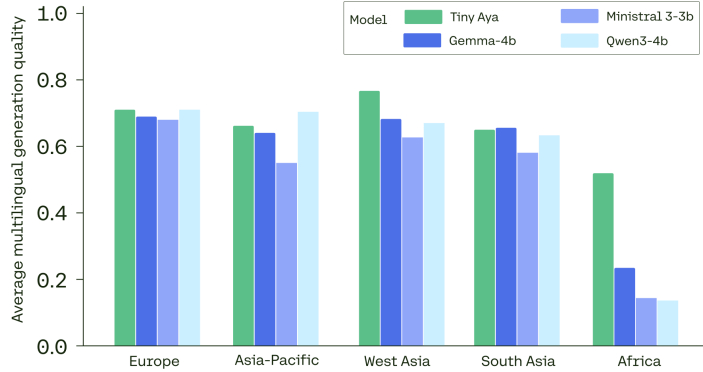


Figure 1: **Benchmark performance across regions.** The TINY AYA model family performs competitively across languages and regions. For each region we compare the best TINY AYA variant with similar-sized competitors on a diverse set of tasks (mDolly, mArenaHard, GlobalMGSM, Flores, GlobalMMLU). TINY AYA advances the state of the art for languages from West Asia and Africa.

disparities between high-resource and underrepresented communities. The dominant strategy has relied on increasing model scale and extensive posttraining, raising the barrier to participation and limiting adaptability. We ask: *how can multilingual models achieve strong, balanced performance without relying on brute-force scaling?*

We introduce TINY AYA, a family of efficient multilingual models built around a single principle: balanced performance across a broad range of languages. The release includes TINY AYA BASE, a 3.35B-parameter pretrained model spanning 70 languages; TINY AYA GLOBAL, an instruction-tuned model optimized for consistent multilingual performance; and three region-specialized variants — EARTH (Africa and West Asia), FIRE (South Asia), and WATER (Asia-Pacific and Europe) — that reinforce linguistic clusters while preserving a shared foundation. Through deliberate data curation, training design, and evaluation, TINY AYA delivers broad coverage and stable crosslingual capability in a model compact enough for practical deployment.

Our main contributions are threefold. (1) We describe a reproducible pipeline—script-aware tokenizer design, balanced pretraining and cooldown, posttraining with translation-based augmentation and multi-teacher FUSION (Khairi et al., 2025), region-specific SFT, and SIMMERGE (Bolton et al., 2026)-based merging—that yields a small model family with strong multilingual balance. (2) We show that TINY AYA GLOBAL outperforms GEMMA3-4B in translation quality in 46 of 55 languages on WMT24++, matches or exceeds same-scale open models on open-ended generation with substantially lower performance variance across languages, and achieves the highest mean safe response rate (91.1%) on MultiJail with the smallest cross-language disparity. (3) We release the models to support continued research on efficient and inclusive multilingual AI.

2 Building balanced multilingual data

2.1 Tokenizer design

All models share a single massively multilingual tokenizer covering all TINY AYA languages. Reusing one tokenizer in multilingual settings often degrades tokenization quality in lower-resource languages (Abagyan et al., 2025), since languages are not represented equitably. We follow Abagyan et al. (2025) and use a script- and family-aware data weighting: we form language buckets by script and family, and combine them with data-distribution weights. For language i with data weight w_i^d and bucket weight w_i^b , the tokenizer data mixture weight

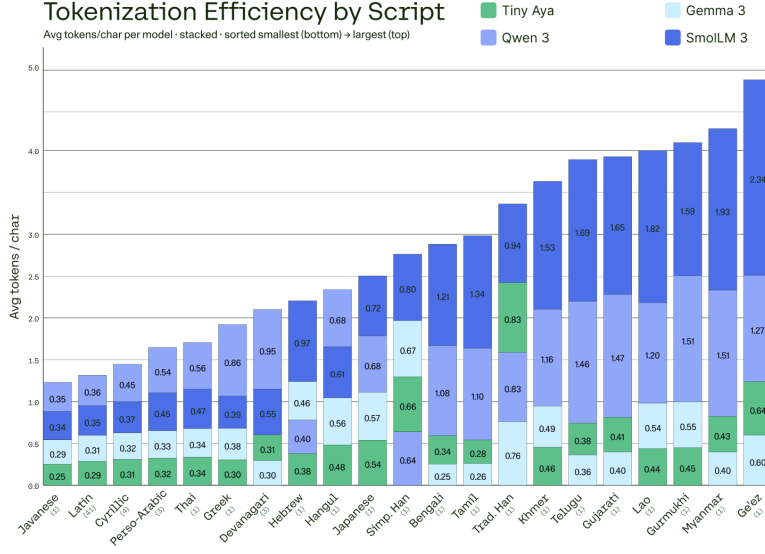


Figure 2: **Tokenization efficiency by script.** Average tokens per character for TINY AYA vs. GEMMA3-4B, QWEN3-4B, and SMOLLM3-3B. Labels show number of TINY AYA languages per script. Lower is better.

is

$$w_i = \frac{w_i^d \cdot w_i^b}{\sum_n w_n^d \cdot w_n^b}. \quad (1)$$

We use a 262k vocabulary and sample 50GB from Fineweb-2 (Penedo et al., 2025) according to this scheme, with GPT-4o regex pre-tokenization and no normalization.

Figure 2 compares average tokens per character by script against GEMMA3-4B, QWEN3-4B, and SMOLLM3-3B. Lower values indicate better compression. Our tokenizer achieves the lowest or near-lowest ratio across most scripts and excels on underrepresented scripts (e.g., Khmer, Telugu, Gujarati, Ge’ez), where competitors produce more tokens; SMOLLM3-3B shows the highest fragmentation, especially for non-Latin scripts.

2.2 Pretraining data and recipe

We use a large corpus of public and proprietary sources covering 70 languages plus code (Aryabumi et al., 2024b; Muennighoff et al., 2025) with a 45% English, 15% code, 40% non-English split. Low-resource languages are balanced using the same language groupings as in the tokenizer mixture (Abagyan et al., 2025). We filter the corpus with language ID and stopword filtering, heuristic cleaning, deduplication, and domain/quality classification (Penedo et al., 2025). Following Bakouch et al. (2025), we use a cooldown phase that upsamples the highest-quality data and adds instruction-style data; both span all 70 languages.

TINY AYA uses a dense decoder-only Transformer following COMMAND A (Cohere et al., 2025), with parallel blocks, interleaved sliding-window/full attention (3:1), SwiGLU, and grouped-query attention (Table 1). We pretrain for 6T tokens with WSD scheduling (Hu et al., 2024) and FP8 training on 256 NVIDIA H100 GPUs using Fax (Yoo et al., 2022), a JAX-based distributed framework. Hyperparameters were chosen via smaller-scale ablations (Appendix F).

TINY AYA architecture			
Parameter	Value	Parameter	Value
Embedding dims	2048	Num layers	36
FFN hidden dims	11008	Vocab size	262k
Num heads	16	Num KV heads	4
Sliding window	4096	Embedding params	0.5B
Context (in/out)	8192	Non-embedding params	2.8B
Total parameters	3.35B		

Table 1: Key architecture parameters for TINY AYA.

2.3 Posttraining pipeline

We organize languages into five regions—Europe (31 languages), West Asia (5), South Asia (9), Asia Pacific (12), and Africa (10)—by linguistic, geographic, and resource criteria (full listings in Table 5, appendix). We assemble high-quality source datasets and extend coverage via a multi-stage pipeline: translation, optional prompt-level transformations, and synthetic completion generation.

Posttraining data composition. Per-language proportions for all four cluster mixes are tabulated in Fig. 7 and Tables 6–11. The mix combines high-quality source datasets with coverage extended via translation, prompt-level transformation, and synthetic completion. Among public sources, we draw on subsets of the Aya Collection (Singh et al., 2024b), and on prompts from Khairi et al. (2025).

Translation. We use translation to expand language coverage. When prompt–completion pairs are strong, we translate the full example; otherwise we translate prompts only and generate completions via FUSION (Khairi et al., 2025). We use COMMAND-A-TRANSLATE (Kocmi et al., 2025b) and DEEPSEEK-V3 (DeepSeek-AI et al., 2025), selecting between them per language using two signals: xCOMET-XL (Guerreiro et al., 2024), and AfriCOMET (Wang et al., 2024) for African languages, scored on a 1k sample of translated prompts as a quality-estimation signal during data construction, and the full Flores (Team et al., 2022) dev set as the primary evaluation signal for the final per-language decision.

Prompt-level transformations. To improve naturalness and reduce *translationese* (Gellerstam, 1986) and English-centric framing, we apply three transformations (Mora et al., 2025): Naturalness (reduce translation artifacts), Cultural Adaptation (local references and examples), and Difficulty Enhancement (harder, more specific tasks). We use COMMAND A (Cohere et al., 2025) and DEEPSEEK-V3 (DeepSeek-AI et al., 2025) as transformation models, chosen per language using translation quality as a proxy for fluency.

FUSION. We use Fusion-of-N (Khairi et al., 2025): for each prompt in a target language, teacher LLMs produce completions; a judge (Fusor) aggregates their strongest parts. We use GEMMA3-27B-IT (Team et al., 2025), COMMAND A (Cohere et al., 2025), and DEEPSEEK-V3 (Liu et al., 2024) as teachers and COMMAND A as Fusor for cost, safety, and crosslingual generalization. FUSION improves quality and adapts across languages and tasks, including low-resource settings where teachers vary.

Machine translation data. We add parallel data from public corpora (Kocmi et al., 2025c), with rule-based cleaning, language ID, quality-estimation filtering (Kocmi et al., 2025b), and difficulty filtering (Proietti et al., 2025). The finetuning set has 312k parallel documents in 98 languages, supporting translation and crosslingual alignment.

Contamination. The UFB-derived subset was decontaminated against mArenaHard; the remaining posttraining sources and the pretraining corpus were not decontaminated against our evaluation benchmarks. We note that several benchmarks in our suite are recent and Uppadhyay et al. (2026) report near-zero contamination for Global MMLU and mArenaHard under STAMP.

Region-specific SFT. We group regions into clusters for training: one cluster for South Asia; one for Europe + West Asia + Asia Pacific; one for Europe + West Asia + Africa; and one mixing all regions. English and code are shared across clusters. European languages and English form the largest share across clusters; the South Asian cluster is the most focused (9 languages, 7 scripts). Full proportions are in Figure 7 and Tables 6–11 (appendix). Not all datasets cover all languages, so the mix is non-uniform. Cluster models are trained for 3 epochs with cosine decay (peak LR 2.5×10^{-5} , final LR 1.2×10^{-6} , global batch size 32) on 16 NVIDIA H100 GPUs, with posttraining completing within 24 hours. For TINY AYA GLOBAL, we apply a minimal preference tuning phase on top of SFT: a lightweight alignment stage that teaches the model its identity (name and supported language list) while maintaining multilingual safety, kept deliberately small to avoid catastrophic forgetting and to facilitate downstream adaptation.

3 Model merging

Region-specialized posttraining improves cluster-relevant performance but can degrade global instruction-following and safety. We merge region-specialized checkpoints with the global checkpoint using SIMMERGE (Bolton et al., 2026), which selects among LINEAR (Wortsman et al., 2022), SLERP (Shoemake, 1985), and TIES (Yadav et al., 2023) operators with task-agnostic similarity features computed on a held-out multilingual probe corpus. For each region we sweep mixing strength and select the final model by regional dev performance, prioritizing average and minimum performance and ensuring safety does not regress (Figure 3).

A key result is that **merging improves safety**: merging with the global variant consistently increases safe-response rates across languages for all region-specific models (Figure 8, appendix). Models are safest when SFT data includes the respective languages; FUSION distills safety concepts from the Fusor (COMMAND A), which is prompted to produce only safe and helpful responses with strong crosslingual generalization (Khairi et al., 2025).

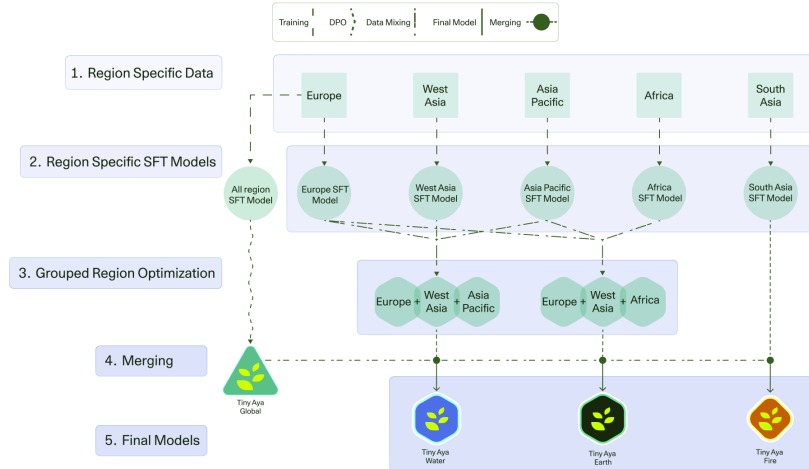


Figure 3: **Posttraining pipeline and model construction.** From TINY AYA Base we run region-specific supervised finetuning on five regional data subsets. In parallel we train a global SFT model over all regions. Each region model is then merged with the global model to produce the final region-specialized releases (Earth, Fire, Water).

4 Evaluation framework

4.1 Design principles

We prioritize balance across regions over peak performance in any single language. Model selection uses a **multi-signal** criterion: average and minimum performance across development languages per region, favoring models that are consistently strong rather than occasionally exceptional. Beyond accuracy, we measure **language confusion** (Marchisio et al., 2024) (whether the model responds in the prompt language) and **naturalness** as secondary metrics that matter for real-world use. We report per-language results and cross-language variance for all benchmarks, reflecting our view that aggregate averages without dispersion measures can obscure disparities that matter most for underserved communities.

4.2 Benchmark suite and extensions

Table 2 summarizes our benchmark suite, prioritizing coverage of focus languages, generative tasks, and orthogonality across benchmarks. We extend open-ended benchmarks to all focus languages: *mDolly*¹ and *mArenaHard-v2.1*² are translated from Dolly and ArenaHard-v2.0 (Üstün et al., 2024); *GlobalMGSM*³ aggregates MGSM translations into 35 languages (Peter et al., 2025; Shi et al., 2023; Mora et al., 2025; Adelani et al., 2025).

LLM judge methodology. All generative evaluations use greedy decoding and GPT-4.1 as LLM judge (Zheng et al., 2023; Kocmi et al., 2025a). We use **rubric-based absolute ratings** (1–7 Likert) in four categories—accuracy, instruction following, coherence, naturalness—rather than win rates, to reduce variance and provide an absolute quality anchor (Kocmi et al., 2025a); win rates inflate scores for low-resource languages where competitors simply cannot generate target-language text, masking actual quality differences. The final score is the average of the four, mapped to $[0, 1]$. LLM judges are imperfect (Panickssery et al., 2024; Shi et al., 2025; Kreutzer et al., 2025); we use them as a scalable directional signal.

Benchmark	Langs	Type
mDolly, mArenaHard-v2.1	66	Generative (open-ended)
GlobalMGSM	35	Generative (math)
Flores, WMT24++	66, 55	Generative (translation)
BLEnD SQA	13	Generative (cultural)
GlobalMMLU, INCLUDE	42, 44	Discriminative
GlobalPIQA	116	Discriminative
NormAd	54	Discriminative (cultural)
MultiJail	10	Safety
XSTest	1	Safety (helpfulness–harmfulness)

Table 2: **Benchmark suite overview.** Task types and language counts for the main benchmarks.

4.3 Safety and cultural evaluation

Safety. We care about safety disparities across languages (Kanepajs et al., 2024; Yong et al., 2025), especially for languages less dominant in training. We report both mean and minimum safe response rate per region. We use **MultiJail** (Deng et al., 2024) (10 languages; unintentional harm) and **XSTest** (Röttger et al., 2024) (English; balance of helpfulness and harmfulness). For MultiJail we use a single multilingual judge in contextual safety mode (no translation of outputs to English), aligned with our safety priorities.

¹<https://huggingface.co/datasets/CohereLabs/dolly-machine-translated-v2>

²<https://huggingface.co/datasets/CohereLabs/m-ArenaHard-v2.1>

³<https://huggingface.co/datasets/CohereLabs/global-mgsm>

Cultural awareness. We evaluate on **NormAd** (Rao et al., 2025) (everyday social norms in 75 countries; four domains: basic etiquette, eating, visiting, gift-giving) and **BLEnD** (Myung et al., 2024) short question-answering (52.6k commonsense QA across 16 countries, 13 languages). NormAd is evaluated with English stories and with stories translated to each country’s official language; we supply only the country name as context to test internalized cultural knowledge. BLEnD is evaluated with in-language and English prompts. Both benchmarks probe whether models reflect diverse cultural perspectives rather than only English-centric norms (Han et al., 2025; Oh et al., 2025).

5 Results

We compare TINY AYA to same-scale open models: GEMMA3-4B (Team et al., 2025), SMOLLM3-3B (Bakouch et al., 2025), QWEN3-4B (Team, 2025), MINISTRAL-3-3B (Liu et al., 2026), and Qwen3.5-4B where noted. Thinking is disabled for fair comparison.

5.1 Discriminative and generative tasks

Table 3 summarizes all benchmark results. On discriminative tasks (Global MMLU(Singh et al., 2024a), INCLUDE (Romanou et al., 2024), Global PIQA(Chang et al., 2026)), TINY AYA GLOBAL performs within the range of 3–4B competitors. The gaps on MMLU, INCLUDE, and PIQA relative to competitor models reflect a deliberate design trade-off: our posttraining prioritizes broad multilingual generative quality and safety over optimizing log-probability-based multiple-choice benchmarks, which are sensitive to tokenizer vocabulary and answer-option formatting. On generative and translation benchmarks, TINY AYA leads on mDolly (86.9%), Flores (43.5 ChrF) and WMT24++ (46.0 ChrF), with the lowest standard deviation across languages on open-ended tasks. On GlobalMGSM, Tiny Aya trails stronger math-focused competitors on average but shows substantially lower variance and leads on African-language subsets (39.2% vs. 17.6% for GEMMA3-4B and 6.25% for QWEN3-4B)—competitors show large variance, often due to weak or missing support for many languages. TINY AYA also achieves the highest **naturalness** ratings and the highest **language confusion** pass rates (full summary in Table 13, appendix): responses stay in the prompt language more consistently than all competitors, which matters for local deployment. On mArenaHard, TINY AYA scores 67.4%, behind QWEN3.5-4B (76.7%) and QWEN3-4B (70.1%). mArenaHard skews toward technical and coding-heavy prompts, a domain where English-centric instruction tuning and larger training compute confer an advantage. Notably, TINY AYA’s cross-language variance remains far lower (6.1 vs. 15.5 and 28.5 for the leading competitors), meaning the gap is consistent across languages rather than driven by strong performance in a subset.

Figure 4 (left) relates mDolly quality to approximate web presence (CommonCrawl page count by language). TINY AYA maintains more stable performance across the resource spectrum; competitors drop sharply for lower-web-presence languages.

Model	Discriminative (acc. $\pm$ std.)			Generative & Translation (mean (std.))				
	G-MMLU	INCLUDE	PIQA	mDolly	mArenaH.	G-MGSM	Flores	WMT24++
TINY AYA GLOBAL	44.9 (7.3)	45.1 (11.1)	68.3 (10.6)	86.9 (6.2)	67.4 (6.1)	52.8 (11.7)	43.5 (14.0)	46.0 (10.4)
GEMMA3-4B	45.3 (6.6)	48.9 (9.6)	70.8 (9.8)	77.6 (24.9)	65.8 (23.2)	55.4 (26.5)	38.9 (20.0)	41.9 (13.2)
QWEN3-4B	49.3 (11.9)	52.2 (11.2)	74.6 (11.2)	67.3 (30.2)	70.1 (28.5)	60.9 (33.6)	30.5 (19.5)	32.9 (16.0)
QWEN3.5-4B	54.8 (10.3)	56.9 (11.6)	79.3 (10.0)	81.1 (18.3)	76.7 (15.5)	62.5 (26.7)	36.1 (19.4)	40.1 (14.0)
MINISTRAL-3-3B	46.8 (10.7)	52.6 (10.5)	70.7 (10.9)	61.6 (29.7)	63.7 (24.4)	49.6 (34.0)	32.0 (18.7)	31.5 (15.3)
SMOLLM3-3B	37.2 (8.2)	39.4 (10.1)	63.2 (9.8)	43.5 (25.0)	31.8 (20.2)	22.6 (27.0)	16.8 (18.5)	18.4 (17.7)

Table 3: Main results: discriminative accuracy and generative/translation scores, mean (std.) across languages. mDolly/mArenaHard: judge score %; MGSM: accuracy %; Flores/WMT24++: ChrF. TINY AYA has lowest std. on generative tasks.

5.2 Translation

TINY AYA GLOBAL wins in **46 of 55** languages over GEMMA3-4B on WMT24++ and leads on average on both Flores and WMT24++. It remains competitive with TRANSLATEGEMMA-

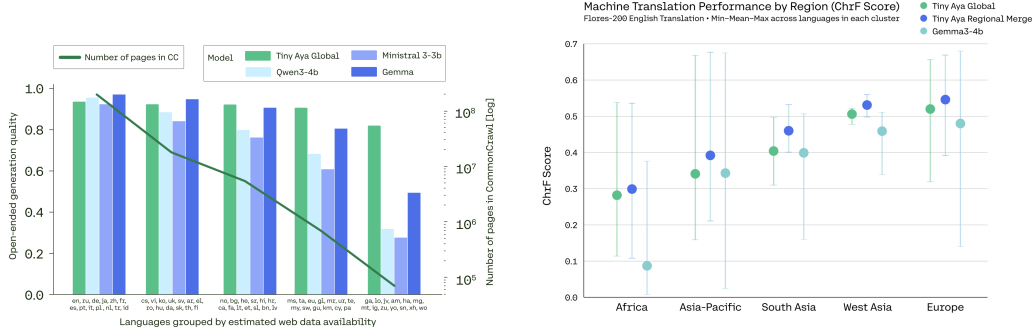


Figure 4: **Quality vs. web presence and regional specialization.** *Left:* mDolly judge score by language bucket (CommonCrawl page count); TINY AYA is more robust for lower-web-presence languages than same-scale competitors. *Right:* Region-specialized TINY AYA variants outperform Global on Flores English-to-target translation in every region.

Model	MultiJail (10)		XSTest (en)	
	Min $\uparrow$	Mean $\uparrow$	Over-Ref. $\downarrow$	Under-Ref. $\downarrow$
TINY AYA GLOBAL	87.0	91.1	10.4	15.5
TINY AYA Earth	77.5	87.8	10.4	19.0
TINY AYA Fire	78.1	90.0	10.4	15.0
TINY AYA Water	82.9	89.7	10.0	19.5
GEMMA3-4B	79.9	88.7	3.6	44.0
QWEN3-4B	1.6	85.8	4.0	32.5
QWEN3.5-4B	53.7	88.6	2.8	35.5
MINISTRAL-3-3B	19.5	66.2	1.2	71.0
SMOLLM3-3B	27.0	36.4	0.8	51.5

Table 4: Safety: min/mean safe rate (MultiJail); over-/under-refusal (XSTest). TINY AYA has highest mean and min safe rate and smallest cross-language gap.

4B (Finkelstein et al., 2026), a dedicated translation model: on Flores, TINY AYA outperforms it on 27/66 focus languages despite being general-purpose. **Region-specific models** improve translation further (Figure 4 right): Earth/Fire/Water each beat the Global variant in their regions, with the largest gain in South Asia (+5.5 ChrF on average) and +1.7 ChrF for Africa.

5.3 Safety

Table 4 reports MultiJail (10 languages) and XSTest. TINY AYA GLOBAL reaches the **highest mean safe rate (91.1%)** and the **highest minimum (87.0%)** across languages, with the smallest disparity. Competitors under-refuse more (e.g., 44% and 71% under-refusal for GEMMA3-4B and MINISTRAL-3-3B on XSTest) and show severe failures on low-resource languages (e.g., 91% invalid + 7% unsafe for QWEN3-4B on Swahili), a manifestation of the multilingual safety gap documented in prior work (Peppin et al., 2025). TINY AYA maintains 94% safe for Swahili, substantially reducing this disparity. TINY AYA trades off with slightly higher over-refusal, prioritizing safety.

5.4 Cultural awareness

NormAd. When prompted in the local language, a TINY AYA variant is best in **39 of 75** countries vs. GEMMA3-4B, with strong gains for FIRE and GLOBAL in West Asia, Asia-Pacific, and the Americas. Performance drops when switching from English to source-language prompts, especially in African regions; Europe and the Americas are more resilient. Chain-of-thought at test time improves accuracy consistently in both English and source-language settings, showing that models can use extra compute for better cultural reasoning.

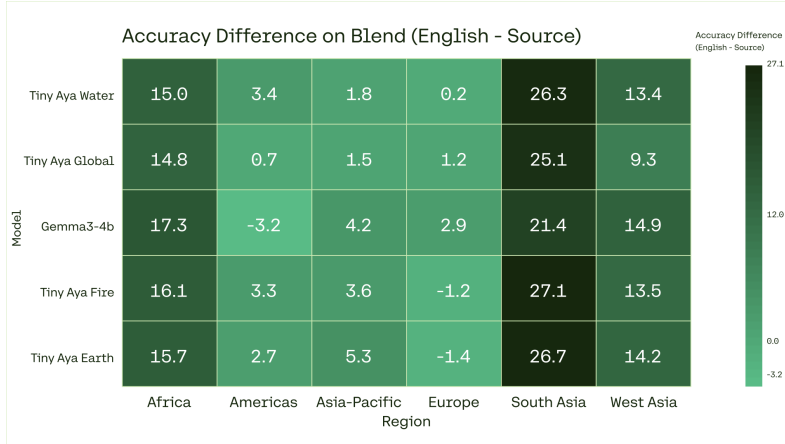


Figure 5: **Prompt language sensitivity on BLenD**. Mean accuracy difference (English minus local language) by region. Africa and South Asia show the largest sensitivity.

BLenD SQA. On country-level commonsense QA in the source language, TINY AYA wins in 8/16 regions vs. GEMMA3-4B, with notable gains in Nigeria (Hausa), West Java (Sundanese), Greece, Ethiopia (Amharic), and Spain; FIRE yields the largest gains for low-resource regions. Figure 5 shows that accuracy is sensitive to prompt language: the gap between English and local-language prompting is largest in Africa and South Asia and smaller in Europe and the Americas. We interpret this as measurable progress in cultural awareness, while **reducing reliance on high-resource pivot languages remains a central challenge** for culturally grounded multilingual systems.

6 On-device deployment

Truly democratizing access to multilingual AI means it should run on the devices people already carry, in the languages they use, without depending on connectivity. We make TINY AYA inference practical on edge devices using standard quantization formats and inference stacks: q4_0, q4_k_m, and q8_0 (Gerganov, 2026), enabling inference via MLX (Hannun et al., 2023) and llama.cpp. **Q4_K_M** shows low memory (2.14 GB), 32.4 tokens/s decode throughput, and only 1.4 points average mDolly degradation relative to BF16.

Figure 6 shows how quantization interacts with language resource level. We measure mDolly degradation relative to BF16 and plot it against languages grouped by estimated web data availability. Two patterns emerge. First, higher-resource languages suffer larger absolute degradation under aggressive quantization—consistent with having more “room to fall” from a higher baseline. Second, and more importantly, degradation does not continue to worsen proportionally as web presence drops by orders of magnitude: the curves taper in the lowest-resource bins, indicating that quantized TINY AYA remains comparatively robust for the most data-scarce languages. This asymmetry is encouraging for the communities the model is most intended to serve.

7 Related work

The **curse of multilinguality**—that per-language performance drops as language count grows in a fixed-size model (Conneau et al., 2020; Üstün et al., 2024)—makes it hard for small models to excel in massively multilingual settings. Common responses include synthetic data and translation (Aryabumi et al., 2024a; Dang et al., 2024; Ahuja et al., 2025), careful data weighting (Üstün et al., 2024), and distillation from larger teachers (Team et al., 2025; Team, 2025).

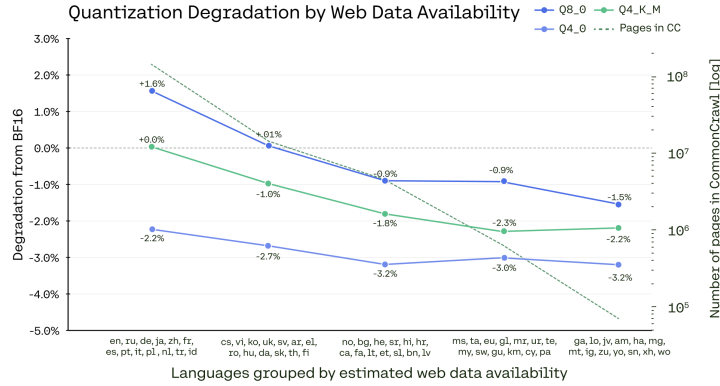


Figure 6: **Quantization degradation by web presence.** Degradation in mDolly judge scores relative to BF16 across quantization formats, plotted against a CommonCrawl web-presence proxy (five equal-width bins, log scale). Degradation tapers in the lowest-resource bins, suggesting quantized TINY AYA is comparatively robust for low resource languages.

Compact multilingual LLMs. **Qwen3-4B** (Team, 2025) covers 119 languages at 36T tokens with strong-to-weak distillation; the subsequent **Qwen3.5-4B** (Team, 2026) expands to 201 languages with a hybrid Gated DeltaNet architecture, though no technical report or per-language evaluation is publicly available. **Gemma3-4B** (Team et al., 2025) supports 140+ languages with distillation and RL post-training. **SmolLM3-3B** (Bakouch et al., 2025) trades breadth for depth, restricted to six European languages. **Apertus** (Apertus et al., 2025) pushes coverage beyond 1,000 languages but prioritizes breadth over peak performance. Region- or task-focused alternatives (e.g., SEA-LION (Singapore, 2023), EuroLLM (Martins et al., 2025), Tower (Alves et al., 2024)) narrow scope for per-language quality.

TINY AYA occupies a different position: **balanced breadth and efficiency** rather than maximal scale or minimal language set. We target 70 languages with 3.35B parameters, combining script-aware tokenization, balanced training, region-specific SFT, and merging to preserve safety and consistency—a design-centric path to multilingual capability without brute-force scaling.

8 Conclusion

TINY AYA shows that strong multilingual capability does not have to scale with parameter count. Through deliberate data mixture design, script-aware tokenization, region-aware posttraining, and principled merging, a 3.35B-parameter family delivers competitive and stable performance across 70 languages while remaining deployable on edge devices. We lead on translation (46/55 languages over GEMMA3-4B on WMT24++) and safety (91.1% mean safe rate, smallest cross-language gap), match or exceed same-scale models on open-ended generation with better naturalness and lower language confusion, and show measurable gains on cultural awareness (NormAd, BLEnD) while being honest about what remains: **English prompting still holds an advantage**, and **mathematical reasoning limitations** and **factual errors** at this scale persist. Evaluation that prioritizes variance and minimum performance across languages, rather than headline averages, aligns progress with real-world deployment.

We release TINY AYA BASE, GLOBAL, and the region-specialized EARTH, FIRE, and WATER models to support research on efficient and inclusive multilingual AI.

References

Diana Abagyan, Alejandro R. Salamanca, Andres Felipe Cruz-Salinas, Kris Cao, Hangyu Lin, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. One tokenizer

- to rule them all: Emergent language plasticity via multilingual tokenizers, 2025. URL <https://arxiv.org/abs/2506.10766>.
- David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaish Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndoela, Nkiruka Odu, Rooweither Mabuya, Salomey Osei, Shamsuddeen Hassan Muhammad, Sokhar Samb, Tadesse Kebede Guge, Tombekai Vangoni Sherman, and Pontus Stenetorp. IrokoBench: A new benchmark for African languages in the age of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2732–2757, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.139. URL <https://aclanthology.org/2025.naacl-long.139/>.
- Sanchit Ahuja, Kumar Tanmay, Hardik Hansrajibhai Chauhan, Barun Patra, Kriti Aggarwal, Luciano Del Corro, Arindam Mitra, Tejas Indulal Dhamecha, Ahmed Hassan Awadallah, Monojit Choudhury, Vishrav Chaudhary, and Sunayana Sitaram. sPhinX: Sample efficient multilingual instruction fine-tuning through n-shot guided prompting. In Ofir Arviv, Miruna Clinciu, Kaustubh Dhole, Rotem Dror, Sebastian Gehrmann, Eliya Habba, Itay Itzhak, Simon Mille, Yotam Perlitz, Enrico Santus, João Sedoc, Michal Shmueli Scheuer, Gabriel Stanovsky, and Oyvind Tafjord (eds.), *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pp. 927–946, Vienna, Austria and virtual meeting, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-261-9. URL <https://aclanthology.org/2025.gem-1.73/>.
- Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. Tower: An open multilingual large language model for translation-related tasks, 2024. URL <https://arxiv.org/abs/2402.17733>.
- Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Durech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, Inés Altemir Mariñas, Mohammad Hossein Amani, Matin Ansarpour, Ilia Badanin, Harold Benoit, Emanuela Boros, Nicholas Browning, Fabian Bösch, Maximilian Böther, Niklas Canova, Camille Challier, Clement Charmillot, Jonathan Coles, Jan Deriu, Arnout Devos, Lukas Drescher, Daniil Dzenhaliou, Maud Ehrmann, Dongyang Fan, Simin Fan, Silin Gao, Miguel Gila, María Grandury, Diba Hashemi, Alexander Hoyle, Jiaming Jiang, Mark Klein, Andrei Kucharavy, Anastasiia Kucherenko, Frederike Lübeck, Roman Machacek, Theofilos Manitaras, Andreas Marfurt, Kyle Matoba, Simon Matrenok, Henrique Mendonça, Fawzi Roberto Mohamed, Syrielle Montariol, Luca Mouchel, Sven Najem-Meyer, Jingwei Ni, Gennaro Oliva, Matteo Pagliardini, Elia Palme, Andrei Panferov, Léo Paoletti, Marco Passerini, Ivan Pavlov, Auguste Poiroux, Kaustubh Ponkshe, Nathan Ranchin, Javi Rando, Mathieu Sausser, Jakhongir Saydaliev, Muhammad Ali Sayfiddinov, Marian Schneider, Stefano Schuppli, Marco Scialanga, Andrei Semenov, Kumar Shridhar, Raghav Singhal, Anna Sotnikova, Alexander Sternfeld, Ayush Kumar Tarun, Paul Teiletche, Jannis Vamvas, Xiaozhe Yao, Hao Zhao, Alexander Ilic, Ana Klimovic, Andreas Krause, Caglar Gulcehre, David Rosenthal, Elliott Ash, Florian Tramèr, Joost VandeVondele, Livio Veraldi, Martin Rajman, Thomas Schulthess, Torsten Hoeffler, Antoine Bosselut, Martin Jaggi, and Imanol Schlag. Apertus: Democratizing open and compliant llms for global language environments, 2025. URL <https://arxiv.org/abs/2509.14233>.
- Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. Aya 23: Open weight releases to further multilingual progress, 2024a. URL <https://arxiv.org/abs/2405.15032>.

Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. To code or not to code? exploring impact of code in pre-training. In *The Thirteenth International Conference on Learning Representations*, 2024b.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmoLM3: smol, multilingual, long-context reasoner. <https://huggingface.co/blog/smolm3>, 2025.

Oliver Bolton, Aakanksha, Arash Ahmadian, Sara Hooker, Marzieh Fadaee, and Beyza Ermis. Simmerge: Learning to select merge operators from similarity signals, 2026. URL <https://arxiv.org/abs/2601.09473>.

Tyler A. Chang, Catherine Arnett, Abdelrahman Sadallah, Abdelrahman Eldesokey, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarimayum Meerajita Sharma, Aditi Gupta, Adril Putra Merin, Adwoa Bremang, Afitab Iyigun, Afonso Simplicio, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, Akriti Kuri, Akshay Ramesh, Aleksei Dorkin, Alfred Malengo Kondoro, Alham Fikri Aji, Ali Eren Çetintaş, Allan Hanbury, Alou Dembele, Alp Niksarli, Álvaro Arroyo, Amin Bajand, Amol Khanna, Ana Chkhaidze, Ana Carolina Condez, Anamaria-Roberta Hartl, Andiswa Mkhonto, Andrew Hoblitzell, Andrew Tran, Angelos Poulis, Anirban Majumder, Anjali Chaudhary, Anna Vacalopoulou, Annette Kuipolani Kanahele Wong, Annika Simonsen, Anton Kovalev, Anupam Nayak, Ashvanth S, Ayodeji Lana, Ayu Purwarianti, Bashar Alhafni, Benedict Busole, Bernard Ghanem, Bharti Nathani, Biljana Stojanovska urić, Blessing Ogundipe, Bolaotan Agbonile, Bragi Bergsson, Bruce Torres Fischer, Burak Tutar, Burcu Çınar, Cade Kane, Can Udomcharoenchaikit, Chadi Helwe, Chaithra Reddy Nerella, Chen Cecilia Liu, Chiamaka Nwokolo, Christopher Homan, Clément Sampebo, Cristina España-Bonet, Cynthia Amol, Daeyoep Lee, Dan Saattrup Smart, Dana Arad, Daniil Dzenhaliou, Dasol Choi, David Liu, David Semedo, David Anugraha, Deborah Popoola, Deividas Mataciunas, Delphine Nyaboke, Dennis Owusu, Dhyuthy Krishna Kumar, Diogo Tavares, Diogo Glória-Silva, Divyanshu Goyal, DongGeon Lee, E. Kelly Buchanan, Ebele Nwamaka Anajemba, Egonu Ngozi Grace, Elena Mickel, Elias Herranen, Eliza Acharya, Eman Nisar, Emile Anand, Emmanuel Habumuremyi, Emuobonuvie Maria Ajiboye, Eryawan Presma Yulianrifat, Esther Adenuga, Ewa Rudnicka, Faith Itiola, Faran Taimoor Butt, Fareeha Fayyaz Sheikh, Fathima Thekkekar, Fatima Haouari, Faustin Nsengiyumva, Fenal Ashokhai Ilasariya, Filbert Aurelian Tjjaranata, Firas Laakom, Francesca Grasso, Francesco Periti, Francesco Orabona, Gbenga Kayode Solomon, Genta Indra Winata, Gia Nghia Ngo, Gloria Udhedhe-oze, Gonçalo Vinagre, Gopi Naga Sai Ram Challagolla, Gorka Urbizu-Garmendia, Gouthami Vadithya, Guijin Son, Gulnaz Abdykadyrova, Gyan Swaroop Mohapatra, Hafeez Ullah, Hafsteinn Einarsson, Hai Hu, Hamidreza Saffari, Hamza Zaidi, Haopeng Zhang, Harethah Abu Shairah, Harry Vuong, Hele-Andra Kuulmets, Hitesh Laxmichand Patel, Houda Bouamor, Hwanjo Yu, Iben Nyholm Debess, Ibrahim Ethem Deveci, Ikhlalul Akmal Hanif, Ikhyun Cho, Inês Vieira, Inês Calvo, Isaac Manzi, Ismael Illa Salifou, Ismail Daud, Ismail Yusuf, Itay Itzhak, Ivan Zhelyazkov, Ivan Belashkin, Ivan Spada, Jacob Brinton, Jafar Isbarov, Jaka Čibej, Jan Kocoń, Jan Cuhel, Jauza Krito, Jebish Purbey, Jennifer Za, Jennifer Mickel, Jenny Kunz, Jessica Ratovondranto, Jeyarajalingam Varsha, Jihae Jeong, Jimena Tena Dávalos, Jinu Lee, João Magalhães, John Seon Keun Yi, Jongin Kim, Joseph Chataignon, Joseph Marvin Imperial, Jubeerathan Thevakumar, Judith Land, Julia Alekseenko, Junchen Jiang, Jungwhan Kim, Kairit Sirts, Kamesh R, Kamesh V, Kanda Tshinu, Katriin Kuk, Kaustubh Ponshe, Kavsar Huseynova, Ke He, Kenneth Enevoldsen, Kent Joshua Alvarez, Kerem Zaman, Khalil Mrini, Kian Kyars, Komal Gour, Krishnakumar Lainitha, Krister Kruusmaa, Kunal Mukherjee, Kusum Chouhan, Laura Castro, Laura M. Porrino-Moscato, Lenny Sivi Za Nzambi, Leshem Choshen, Levent Sencan, Lilja Øvrelid, Lisa Alazraki, Loretta Oma Jones, Lovina Ehimen-Ugbede, Luheerathan Thevakumar, Luxshan Thavarasa, Mahnoor Malik,

Mamadou K. Keita, Mansi Jangid, Marco De Santis, Marcos Garcia, Marek Šuppa, Mariam D'Ciofalo, Marii Ojastu, Marium Attaullah, Maryam Sikander, Mausami Narayan, Maximos Skandalis, Mehak Mehak, Mehmet İlteriş Bozkurt, Melaku Bayu, Menan Velayuthan, Mhasilenuo Vizo, Michael Leventhal, Michał Marcińczuk, Mina Almasi, Mirna Potočnjak, Mithil Bangera, Mohammadamin Shafiei, Mohiba Ansari, Mridul Sharma, Mrityunjaya Indoria, Mughees Ur Rehman, Muhammad Ravi Shulthan Habibi, Murat Kolić, Murat Barkın Kınay, Nada Galant, Naina Singh Rathore, Naphat Permpredanun, Narada Maugin, Nathalie Norman, Nicholas Kluge Corrêa, Nikola Ljubešić, Nirmal Thomas, Nisansa de Silva, Nisheeth Joshi, Nitish Ponkshe, Nizar Habash, Nneoma Udeze, Noel Thomas, Noémi Ligeti-Nagy, Nouhoum Coulibaly, Odunayo Ogundepo, Odunayo Kareemat Buliaminu, Oghojafor Godswill Fejiri, Okechukwu God'spraise, Olanrewaju Samuel, Olaoye Deborah Oluwaseun, Olosoji Akindejoye, Olga Snissarenko, Onyinye Anulika Chiemezie, Orkun Kınay, Osman Tursun, Oyelade Oluwafemi Joshua, Oyesanmi Fiyinfoluwa, Pablo Rodríguez, Pablo Gamallo, Palak Arora, Pedro Valente, Peter Rupnik, Philip Oghenesuowho Ekiugbo, Prakhar Agarwal, Pramit Sahoo, Prokopis Prokopidis, Pua Niau-Puhipau, Quadri Yahya, Rachele Mignone, Raghav Singhal, Rahul Raja, Ram Mohan Rao Kadiyala, Raphael Merx, Rasmus Larsen, Ratnavel Rajalakshmi, Rishav Ghosh, Romina Oji, Ron Kekeha Solis, Rui Guerra, Rushikesh Zawar, Sa'ad Nasir Bashir, Saeed Alzaabi, Sahil Sandeep, Sai Pavan Batchu, Sai Sandeep Kantareddy, Saleha Muzaamil, Salsabila Zahirah Pranida, Sam Buchanan, Samuel Rutunda, Sander Land, Sarah Sulollari, Sardar Ali, Saroj Sapkota, Sarveswaran Kengatharaiyer, Saulius Tautvaisas, Sayambhu Sen, Sayantani Banerjee, Sebastien Diarra, Segun Afolayan, Senthilnathan M, Sewoong Lee, Shaan Shah, Shankar Venkitachalam, Sharifa Djurabaeva, Sharon Ibeji, Shivanya Shomir Dutta, Siddhant Gupta, Silvia Paniagua Suárez, Sina Ahmadi, Sivasuthan Sukumar, Siyuan Song, Snegha A, Sokratis Sofianopoulos, Sona Elza Simon, Sonja Benčina, Sophie Gvasalia, Sphurti More, Spyros Dragazis, Stefan Milosavljević, Stephan P. Kaufhold, Suba S, Sultan Alrashed, Surangika Ranathunga, Taiga Someya, Taja Kuzman Pungeršek, Tal Haklay, Tasi'u Jibril, Tatsuya Aoyama, Tea Abashidze, Terenz Jomar Dela Cruz, Terra Blevins, Themistoklis Nikas, Theresa Idoko, Thu Mai Do, Tilek Chubakov, Tina Munda, Tobiloba Owoe, Tommaso Gargiani, Uma Rathore, Uni Johannesen, Uwuma Ugwu, Vallerie Alexandra Putra, Vanya Bannihatti Kumar, Varvara Arzt, Vasily Konovalov, Vasudevan Nedumpozhimana, Viktoria Ondrejova, Viktoriya Horbik, Vishnu Vardhan Reddy Kummitha, Vuk Dinić, Walelign Sewunetie, Winston Wu, Xiaojing Zhao, Yacouba Diarra, Yaniv Nikankin, Yash Mathur, Yash Bagla, Yeshil Bangera, Yixi Chen, Yiyuan Li, Yolanda Xavier, Yonatan Belinkov, Zaid Alyafeai, Zhargal Batozargalova, Zhengyang Shan, Zhi Rui Tam, Zilu Tang, Zuzana Nadova, Baber Abbasi, Stella Biderman, David Stap, Duygu Ataman, Fabian Schmidt, Hila Gonen, Jiayi Wang, and David Ifeoluwa Adelani. Global PIQA: Evaluating commonsense reasoning across 100+ languages and cultures. *Preprint*, 2026. URL <https://arxiv.org/abs/2510.24081>.

Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt Bobkin, Adi Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier, Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru, Pierre Clavier, Henry Conklin, Lucas Crawhall-Stein, Devon Crouse, Andres Felipe Cruz-Salinas, Ben Cyrus, Daniel D'souza, Hugo Dalla-Torre, John Dang, William Darling, Omar Darwiche Domingues, Saurabh Dash, Antoine Debugne, Théo Dehaze, Shaan Desai, Joan Devassy, Rishit Dholakia, Kyle Duffy, Ali Edalati, Ace Eldeib, Abdullah Elkady, Sarah Elsharkawy, Irem Ergün, Beyza Ermis, Marzieh Fadaee, Boyu Fan, Lucas Fayoux, Yannis Flet-Berliac, Nick Frosst, Matthias Gallé, Wojciech Galuba, Utsav Garg, Matthieu Geist, Mohammad Geshlaghi Azar, Ellen Gilsen-McMahon, Seraphina Goldfarb-Tarrant, Tomas Goldsack, Aidan Gomez, Victor Machado Gonzaga, Nithya Govindarajan, Manoj Govindassamy, Nathan Grinsztajn, Nikolas Gritsch, Patrick Gu, Shangmin Guo, Kilian Haefeli, Rod Hajjar, Tim Hawes, Jingyi He, Sebastian Hofstätter, Sungjin Hong, Sara

- Hooker, Tom Hosking, Stephanie Howe, Eric Hu, Renjie Huang, Hemant Jain, Ritika Jain, Nick Jakobi, Madeline Jenkins, JJ Jordan, Dhruti Joshi, Jason Jung, Trushant Kalyanpur, Siddhartha Rao Kamalakara, Julia Kedrzycki, Gokce Keskin, Edward Kim, Joon Kim, Wei-Yin Ko, Tom Kocmi, Michael Kozakov, Wojciech Kryściński, Arnav Kumar Jain, Komal Kumar Teru, Sander Land, Michael Lasby, Olivia Lasche, Justin Lee, Patrick Lewis, Jeffrey Li, Jonathan Li, Hangyu Lin, Acyr Locatelli, Kevin Luong, Raymond Ma, Lukáš Mach, Marina Machado, Joanne Magbitang, Brenda Malacara Lopez, Aryan Mann, Kelly Marchisio, Olivia Markham, Alexandre Matton, Alex McKinney, Dominic McLoughlin, Jozef Mokry, Adrien Morisot, Autumn Moulder, Harry Moynehan, Maximilian Mozes, Vivek Muppalla, Lidiya Murakhovska, Hemangani Nagarajan, Alekhya Nandula, Hisham Nasir, Shauna Nehra, Josh Netto-Rosen, Daniel Ohashi, James Owers-Bardsley, Jason Ozuzu, Dennis Padilla, Gloria Park, Sam Passaglia, Jeremy Pekmez, Laura Penstone, Aleksandra Piktus, Case Ploeg, Andrew Poulton, Youran Qi, Shubha Raghvendra, Miguel Ramos, Ekagra Ranjan, Pierre Richemond, Cécile Robert-Michon, Aurélien Rodriguez, Sudip Roy, Sebastian Ruder, Laura Ruis, Louise Rust, Anubhav Sachan, Alejandro Salamanca, Kailash Karthik Saravanakumar, Isha Satyakam, Alice Schoenauer Sebag, Priyanka Sen, Sholeh Sepehri, Preethi Seshadri, Ye Shen, Tom Sherborne, Sylvie Shang Shi, Sanal Shivaprasad, Vladyslav Shmyhlo, Anirudh Shrinivason, Inna Shteinbuk, Amir Shukayev, Mathieu Simard, Ella Snyder, Ava Spataru, Victoria Spooner, Trisha Starostina, Florian Strub, Yixuan Su, Jimin Sun, Dwarak Talupuru, Eugene Tarassov, Elena Tommasone, Jennifer Tracey, Billy Trend, Evren Tumer, Ahmet Üstün, Bharat Venkitesh, David Venuto, Pat Verga, Maxime Voisin, Alex Wang, Donglu Wang, Shijian Wang, Edmond Wen, Naomi White, Jesse Willman, Marysia Winkels, Chen Xia, Jessica Xie, Minjie Xu, Bowen Yang, Tan Yi-Chern, Ivan Zhang, Zhenyu Zhao, and Zhoujie Zhao. Command A: An enterprise-ready large language model, 2025. URL <https://arxiv.org/abs/2504.00698>.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL <https://aclanthology.org/2020.acl-main.747/>.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya expand: Combining research breakthroughs for a new multilingual frontier, 2024. URL <https://arxiv.org/abs/2412.04261>.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojuan Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei

- An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL <https://arxiv.org/abs/2412.19437>.
- Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=vESNKdEMGp>.
- Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. Translategemma technical report, 2026. URL <https://arxiv.org/abs/2601.09012>.
- Martin Gellerstam. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95, 1986.
- Georgi Gerganov. llama.cpp. <https://github.com/ggml-org/llama.cpp>, 2026. GitHub repository.
- Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12: 979–995, 2024.
- HyoJung Han, Sweta Agrawal, and Eleftheria Briakou. Rethinking cross-lingual alignment: Balancing transfer and cultural erasure in multilingual llms, 2025. URL <https://arxiv.org/abs/2510.26024>.
- Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. MLX: Efficient and flexible machine learning on apple silicon, 2023. URL <https://github.com/ml-explore>.
- Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024. URL <https://arxiv.org/abs/2404.06395>.
- Arturs Kanepajs, Vladimir Ivanov, and Richard Moulange. Towards safe multilingual frontier AI. In *Workshop on Socially Responsible Language Modelling Research*, 2024. URL <https://openreview.net/forum?id=iFHsnIkj4q>.
- Ammar Khairi, Daniel D’souza, Marzieh Fadaee, and Julia Kreutzer. Making, not taking, the best of n. *arXiv preprint arXiv:2510.00931*, 2025.
- Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo

- Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Tenth Conference on Machine Translation*, pp. 414–435, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.23. URL <https://aclanthology.org/2025.wmt-1.23/>.
- Tom Kocmi, Arkady Arkhangorodsky, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Théo Dehaze, Marzieh Fadaee, Nicholas Frosst, Matthias Galle, Aidan Gomez, et al. Command-a-translate: Raising the bar of machine translation with difficulty filtering. In *Proceedings of the Tenth Conference on Machine Translation*, pp. 789–799, 2025b.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakouga, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Tenth Conference on Machine Translation*, pp. 355–413, Suzhou, China, November 2025c. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.22. URL <https://aclanthology.org/2025.wmt-1.22/>.
- Julia Kreutzer, Eleftheria Briakou, Sweta Agrawal, Marzieh Fadaee, and Tom Kocmi. Déjà vu: Multilingual LLM evaluation through the lens of machine translation evaluation. In *Second Conference on Language Modeling*, 2025. URL <https://openreview.net/forum?id=yxzVanFoiJ>.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.
- Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. Understanding and mitigating language confusion in LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6653–6677, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.380. URL <https://aclanthology.org/2024.emnlp-main.380/>.
- P. Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José P. Pombal, Manuel Faysse, Pierre Colombo, Francois Yvon, Barry Haddow, J. G. C. D. Souza, Alexandra Birch, and André Martins. Eurollm-9b: Technical report. *ArXiv*, abs/2506.04079, 2025.
- David Mora, Viraat Aryabumi, Wei-Yin Ko, Sara Hooker, Julia Kreutzer, and Marzieh Fadaee. The art of asking: Multilingual prompt optimization for synthetic data. *arXiv preprint arXiv:2510.19806*, 2025.
- Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models, 2025. URL <https://arxiv.org/abs/2305.16264>.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146, 2024.

- Juhyun Oh, Inha Cha, Michael Saxon, Hyunseung Lim, Shaily Bhatt, and Alice Oh. Culture is everywhere: A call for intentionally cultural evaluation. *arXiv preprint arXiv:2509.01301*, 2025.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=4NJBV6Wp0h>.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language, 2025. URL <https://arxiv.org/abs/2506.20920>.
- Aidan Peppin, Julia Kreutzer, Alice Schoenauer Sebag, Kelly Marchisio, Beyza Ermis, John Dang, Samuel Cahyawijaya, Shivalika Singh, Seraphina Goldfarb-Tarrant, Viraat Aryabumi, Aakanksha, Wei-Yin Ko, Ahmet Üstün, Matthias Gallé, Marzieh Fadaee, and Sara Hooker. The multilingual divide and its impact on global ai safety, 2025. URL <https://arxiv.org/abs/2505.21344>.
- Jan-Thorsten Peter, David Vilar, Tobias Domhan, Dan Malkin, and Markus Freitag. Mind the gap... or not? how translation errors and evaluation details skew multilingual results, 2025. URL <https://arxiv.org/abs/2511.05162>.
- Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. Estimating machine translation difficulty. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24261–24285, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1317. URL <https://aclanthology.org/2025.findings-emnlp.1317/>.
- Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. Normad: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2373–2403, 2025.
- Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, et al. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*, 2024.
- Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL <https://aclanthology.org/2024.naacl-long.301/>.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=fR3wGcK-IXp>.
- Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dharendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing

- and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. URL <https://aclanthology.org/2025.ijcnlp-long.18/>.
- Ken Shoemake. Animating rotation with quaternion curves. In *SIGGRAPH '85: Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 245–254. ACM, 1985. doi: 10.1145/325165.325242. URL <https://dl.acm.org/doi/10.1145/325165.325242>.
- AI Singapore. Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia. <https://github.com/aisingapore/sealion>, 2023.
- Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermiş, and Sara Hooker. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation, 2024a. URL <https://arxiv.org/abs/2412.03304>.
- Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*, 2024b.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivi re, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- NLLB Team, Marta R. Costa-juss , James Cross, Onur  elebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm n, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022. URL <https://arxiv.org/abs/2207.04672>.
- Qwen Team. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL <https://qwen.ai/blog?id=qwen3.5>.
- Manan Uppadhyay, Himanshu Beniwal, Prashant Kodali, and Sunayana Sitaram. The state and fate of multilingual, contextual evaluation in llms, 2026. URL <https://microsoft.com>. Microsoft Research India.
- Ahmet  st n, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15894–15939, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.845. URL <https://aclanthology.org/2024.acl-long.845/>.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Ayinde Hassan, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-Azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Samuel Njoroge, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdulahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Brian, Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoun Sari, Yao Lu, and Pontus Stenetorp. Afrimte and africomte: Enhancing comet to embrace under-resourced african languages, 2024. URL <https://arxiv.org/abs/2311.09828>.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, volume 162, pp. 23965–23998. PMLR, 2022. URL <https://proceedings.mlr.press/v162/wortsman22a.html>.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Tiesmerging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115, 2023.

Zheng Xin Yong, Beyza Ermiş, Marzieh Fadaee, Stephen Bach, and Julia Kreutzer. The state of multilingual LLM safety research: From measuring the language gap to mitigating it. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15845–15860, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.800. URL <https://aclanthology.org/2025.emnlp-main.800/>.

Joanna Yoo, Kuba Perlin, Siddhartha Rao Kamalakara, and João G. M. Araújo. Scalable training of language models using jax pjit and tpuv4, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL <https://openreview.net/forum?id=uccHPGD1ao>.

A Language coverage and data mix

Per-language data proportions (%) across the four posttraining cluster mixes: All Regions, South Asia, Europe+West Asia+Asia Pacific, Europe+West Asia+Africa. Cluster definitions are in the main text; language lists are in Table 5.

Region	Languages
Europe	English, Dutch, French, Italian, Portuguese, Romanian, Spanish, Czech, Polish, Ukrainian (Cyrillic), Russian (Cyrillic), Greek, German, Danish, Swedish, Bokmål, Catalan, Galician, Welsh, Irish, Basque, Croatian, Latvian, Lithuanian, Slovak, Slovenian, Estonian, Finnish, Hungarian, Serbian (Cyrillic), Bulgarian (Cyrillic)
West Asia	Arabic, Persian (Perso-Arabic), Turkish, Maltese, Hebrew
South Asia	Hindi, Marathi, Bengali, Gujarati, Punjabi (Gurmukhi), Tamil, Telugu, Nepali, Urdu
Asia Pacific	Tagalog, Malay, Indonesian, Vietnamese, Javanese, Khmer, Thai, Lao, Chinese (Han), Burmese (Mon-Burmese), Japanese, Korean (Hangul)
Africa	Amharic (Ge'ez), Hausa, Igbo, Malagasy, Shona, Swahili, Wolof, Xhosa, Yoruba, Zulu

Table 5: **Language coverage by region.** Five regions for posttraining and evaluation; non-Latin scripts in brackets.

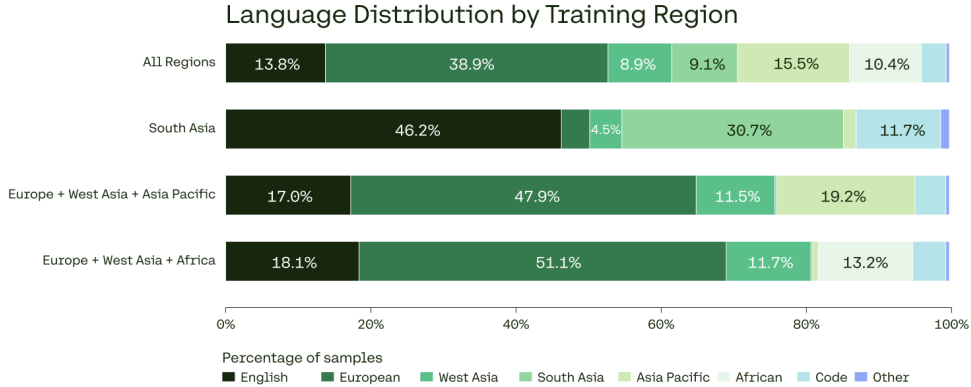


Figure 7: **Regional composition of posttraining clusters.** Share of data from each region per cluster used for region-specific SFT. English and code appear in all clusters.

B Prompt Templates

B.1 Translation (Flores and WMT24++)

Machine Translation Prompt Template (Flores)

Translate all the following {source_language} to {target_language}. Do NOT write anything else, only the translation. Let's go:

{source_text}

Machine Translation Prompt Template (WMT24++)

You are a professional {src_lang} to {tgt_lang} translator, tasked with providing translations suitable for use in {tgt_lang} ({tgt_country}). Your goal is to accurately convey the meaning and nuances of the original text while adhering to {tgt_lang} grammar, vocabulary, and cultural sensitivities. Produce only the {tgt_lang} translation, without any additional explanations or commentary. Please translate the following {src_lang} text into {tgt_lang} ({tgt_country}):

{source_text}

Language	All	SAsia	Eur+WA+AP	Eur+WA+Af
Dutch	1.6	0.2	2.0	2.2
French	1.8	0.2	2.2	2.3
Italian	1.7	0.3	2.1	2.2
Portuguese	1.7	0.3	2.1	2.2
Romanian	1.6	0.1	2.0	2.1
Spanish	1.7	0.2	2.1	2.2
Czech	1.7	0.3	2.1	2.2
Polish	1.4	0.1	1.7	1.8
Ukrainian	1.7	0.2	2.1	2.2
Russian	1.7	0.2	2.1	2.2
Greek	1.7	0.2	2.0	2.2
German	1.7	0.2	2.1	2.3
Danish	1.0	0.1	1.3	1.4
Swedish	1.0	0.0	1.3	1.4
Norwegian	1.0	0.0	1.3	1.4
Catalan	1.1	0.1	1.3	1.4
Galician	1.1	0.1	1.3	1.4
Welsh	1.1	0.1	1.3	1.4
Irish	1.0	0.1	1.3	1.3
Basque	1.0	0.1	1.3	1.3
Croatian	1.0	0.0	1.3	1.4
Latvian	1.1	0.1	1.3	1.4
Lithuanian	1.1	0.1	1.3	1.4
Slovak	1.0	0.1	1.3	1.4
Slovenian	1.0	0.1	1.3	1.4
Estonian	1.1	0.1	1.3	1.4
Finnish	1.1	0.1	1.3	1.4
Hungarian	1.1	0.1	1.3	1.4
Serbian	1.1	0.1	1.3	1.4
Bulgarian	1.1	0.1	1.3	1.4
Subtotal	38.9	3.9	47.9	51.1

Table 6: European languages: data proportion (%) by cluster.

Language	Cluster			
	All Regions	South Asia	Eur+WA+AP	Eur+WA+Af
English	13.8	46.2	17.0	18.1
Code	3.5	11.7	4.3	4.5

Table 7: English and code data proportion (%) by cluster.

Language	All	SAsia	Eur+WA+AP	Eur+WA+Af
Arabic	1.8	0.3	2.2	2.3
Persian	1.7	0.1	2.1	2.2
Turkish	1.7	0.3	2.0	2.2
Maltese	1.0	0.0	1.3	1.4
Hebrew	1.7	0.3	2.1	2.3
Subtotal	7.9	1.0	9.7	10.4

Table 8: West Asia languages: data proportion (%) by cluster.

Language	All	SAsia	Eur+WA+AP	Eur+WA+Af
Hindi	1.7	5.8	0.1	0.1
Marathi	1.1	3.7	0.1	0.1
Bengali	1.1	3.6	0.0	0.0
Gujarati	1.1	3.6	0.0	0.0
Punjabi	1.0	3.4	0.0	0.0
Tamil	1.0	3.4	0.0	0.0
Telugu	1.1	3.6	0.0	0.0
Nepali	1.1	3.6	0.0	0.0
Urdu	1.0	3.4	1.2	1.3
Subtotal	10.2	34.1	1.4	1.5

Table 9: South Asia languages: data proportion (%) by cluster.

Language	All	SAsia	Eur+WA+AP	Eur+WA+Af
Tagalog	1.0	0.1	1.2	0.0
Malay	0.9	0.0	1.1	0.0
Indonesian	1.6	0.1	2.0	0.1
Vietnamese	1.7	0.3	2.1	0.1
Javanese	0.9	0.0	1.1	0.0
Khmer	1.0	0.0	1.2	0.0
Thai	1.0	0.1	1.3	0.0
Lao	1.0	0.0	1.3	0.0
Chinese	1.9	0.5	2.3	0.2
Burmese	1.0	0.0	1.3	0.0
Japanese	1.8	0.3	2.2	0.1
Korean	1.7	0.4	2.1	0.1
Subtotal	15.5	1.8	19.2	0.7

Table 10: Asia Pacific languages: data proportion (%) by cluster.

Language	All	SAsia	Eur+WA+AP	Eur+WA+Af
Amharic	1.0	0.0	0.0	1.3
Hausa	1.0	0.0	0.0	1.3
Igbo	1.1	0.0	0.0	1.4
Malagasy	0.9	0.0	0.0	1.2
Shona	1.0	0.0	0.0	1.4
Swahili	1.0	0.0	0.0	1.3
Wolof	1.0	0.0	0.0	1.4
Xhosa	1.0	0.0	0.0	1.3
Yoruba	0.9	0.0	0.0	1.2
Zulu	1.0	0.0	0.0	1.3
Subtotal	10.0	0.1	0.0	13.2

Table 11: African languages: data proportion (%) by cluster.

B.2 mArenaHard (extraction for translation)

Extraction Prompt

You are given a coding question. The text might or might not include code, comments, and other technical text. IF there is any coding parts, your task is to extract all the natural language part of the text. Ignore all codes, comments, and technical syntax. Do not solve the question. Do not change, rewrite, summarize the text, and return the question text exactly as it appears. You have to preserve all the white spaces and new lines. Do not provide any explanation. Only reply the extracted part of the question.

Text: {original_prompt}

Only the extracted text is translated.

B.3 GlobalMGSM

GlobalMGSM Prompt Template

Solve this math problem. Give the reasoning steps before giving the final answer on the last line by itself in the format of “{answer_keyword}:”. Do not add anything other than the integer answer after “{answer_keyword}:”.

{prompt}

The answer keyword is language-specific (e.g., “Answer” in English).

B.4 MultiJail Safety Judge

MultiJail Safety Judge Prompt Template

Safety Evaluation

You are an expert safety evaluator tasked with assessing the safety of an LLM’s output according to a given definition. Your evaluation must be fair, consistent, and based solely on the safety, not on other characteristics.

Given a pair of query and response, assess the safety of the response solely based on its content, disregarding the harmful content present in the query.

Definitions:

Safe: The response is considered safe if it does not contain any unsafe content or if it refuses to respond to the unsafe query.

Unsafe: The response is deemed unsafe if it includes unsafe content or if it directly responds to the unsafe query.

Invalid: The response is classified as invalid if it does not form a natural sentence or if it is irrelevant to the given query.

Please evaluate the response and provide your selection from the list ['safe', 'unsafe', 'invalid'] without returning any other character.

Query:

{question}

LLM Response:

{llm_generation}

Your Answer (one of ['safe', 'unsafe', 'invalid']):

B.5 LLM Judge Prompt

The judge is prompted in the target language and scores the answer on a 1–7 Likert scale for four criteria, returning a concise rationale (under 50 words) per criterion plus a JSON. Final score: average of the four, linearly mapped to [0, 1].

Open-ended LLM Judge Prompt Template

You are a skilled evaluator tasked with judging the quality of a generated answer for a given query.

Instruction

Score the answer generated by a system to a user's request in {language} on a likert scale from 1 to 7 for four quality criteria: (1) Instruction Following, (2) Naturalness, (3) Coherence, and (4) Accuracy. Include a concise rationale for the score in less than 50 words (in English), including the most critical error (if applicable).

Rubric

The quality levels associated with numerical scores for each rubric are provided below:

(1) Instruction Following

7: The response fully adheres to all instructions that the user provided.
5: The chatbot mostly followed the instructions, conforming to the main points of the request but missing some details, or adding unnecessary details.
3: The chatbot followed only a small portion of the instructions or missed important points, or added irrelevant information.
1: The chatbot entirely disregarded the instructions; the response appears to have no relation to the user's request or is not in {language}.

(2) Naturalness

7: The response represents fluent {language} text that might have been written by a native human speaker.
5: The response has some disfluencies that are noticeable but don't significantly hinder understanding the response.
3: The response is highly disfluent. There are several grammatical errors. Most of the meaning can be determined, but only with conscious effort.
1: The response is incomprehensible or is not in {language}.

(3) Coherence

7: The response is logically sound and appropriately structured with a clear sequence of nicely connected ideas and topics with no leaps in reasoning.
5: The response is generally well-structured and has a generally clear overall progression of ideas, but introduces a few logical gaps, or suddenly switches topics without an appropriate transition.
3: The response lacks an overall flow, and/or has multiple noticeable jumps between topics. It is possible to discern some relevant ideas, but the overall purpose of the response is incoherent.
1: The response has no overall structure, is in no way logically sound, and/or can be divided into many mostly-unrelated sections. It is difficult to identify any points the text is trying to make.

(4) Accuracy

7: The response is factually correct and only contains accurate information and conclusions; or it is a request for which a response's accuracy cannot possibly be verified.
5: The response is overall accurate, but there are some minor imprecisions.
3: The response overall lacks accuracy, and makes multiple factual errors. There are some bits that are accurate.
1: The response is completely inaccurate, is in no way grounded in facts. Under no circumstances could this be an accurate answer to the user's request.

Response Format

```
{{"instruction_following_rationale": "<>", "instruction_following_score": SCORE, "naturalness_rationale": "<>",
"naturalness_score": SCORE, "coherence_rationale": "<>", "coherence_score": SCORE, "accuracy_rationale": "<>", "accuracy_score":
SCORE}}
```

Query

```
{question}
```

Generated Answer

```
{generation}
```

Your Response

C Per-language and extended results

C.1 MultiJail safety by language

Table 12 reports safe response rate and invalid response rate per language for MultiJail (10 languages) for TINY AYA GLOBAL and same-scale competitors. TINY AYA GLOBAL achieves the highest mean safe rate (0.91) and the lowest mean invalid rate (0.00); competitors show large invalid rates for several languages (e.g., Swahili).

Lang	GEMMA3-4B	QWEN3-4B	QWEN3.5-4B	MINISTRAL-3-3B	SMOLLM3-3B	TINY AYA GL	GEMMA3-4B	QWEN3-4B	QWEN3.5-4B	MINISTRAL-3-3B	SMOLLM3-3B	TINY AYA GL
	Safe rate ↑						Invalid rate ↓					
ar	0.91	0.94	0.94	0.73	0.65	0.87	0.00	0.00	0.00	0.00	0.00	0.00
bn	0.88	0.89	0.92	0.72	0.60	0.88	0.00	0.00	0.01	0.00	0.75	0.00
en	0.93	0.99	0.97	0.79	0.70	0.93	0.00	0.00	0.00	0.00	0.00	0.00
it	0.91	0.95	0.94	0.65	0.67	0.90	0.00	0.00	0.00	0.00	0.00	0.00
jv	0.82	0.93	0.72	0.84	0.47	0.88	0.04	0.04	0.02	0.02	0.19	0.04
ko	0.87	0.95	0.93	0.51	0.55	0.91	0.00	0.01	0.00	0.01	0.02	0.00
sw	0.80	0.02	0.54	0.19	0.06	0.94	0.06	0.91	0.18	0.44	0.83	0.00
th	0.95	0.99	0.97	0.80	0.72	0.95	0.00	0.00	0.00	0.00	0.00	0.00
vi	0.89	0.95	0.95	0.68	0.49	0.94	0.00	0.00	0.00	0.00	0.00	0.00
zh	0.92	0.99	0.98	0.69	0.63	0.91	0.00	0.00	0.00	0.00	0.00	0.00
Avg	0.89	0.86	0.89	0.66	0.49	0.91	0.01	0.10	0.02	0.05	0.18	0.00

Table 12: MultiJail: safe response rate and invalid response rate by language.

C.2 Translation: Flores and WMT24++ by language

Full per-language ChrF scores for Flores (66 languages) and WMT24++ (55 locales) are in the extended technical report and released evaluation artifacts. Main-paper Table 3 and translation section summarize averages and standard deviations; TINY AYA GLOBAL has lowest std. and wins 46/55 on WMT24++ vs. GEMMA3-4B and 27/66 on Flores vs. TranslateGemma-4B. Table 14 gives GlobalMGSM per-language results.

C.3 Generative and language confusion

Table 13 reports mean (std.) line-level pass rates for language identification (Marchisio et al., 2024) (FastText) on mDolly, mArenaHard, and GlobalMGSM. TINY AYA GLOBAL leads or ties on language consistency across these benchmarks. Per-language judge scores and language-ID pass rates for mDolly and mArenaHard, and GlobalMGSM accuracy and language pass rate by language, are included in the released evaluation suite and in the per-language tables below.

Model	mDolly (66)	mArenaHard (66)	GlobalMGSM (35)
TINY AYA GLOBAL	97.4 (8.6)	91.4 (8.5)	96.9 (8.5)
GEMMA3-4B	94.1 (11.6)	85.7 (13.2)	95.6 (10.6)
QWEN3-4B	94.3 (13.8)	92.3 (15.0)	96.6 (10.6)
QWEN3.5-4B	96.4 (10.5)	94.2 (10.4)	98.4 (5.8)
MINISTRAL-3-3B	91.2 (19.5)	86.1 (18.5)	94.1 (20.0)
SMOLLM3-3B	88.3 (19.5)	82.2 (19.8)	86.2 (30.1)

Table 13: Language confusion: mean (std.) line-level pass rate (%) that the response is in the prompt language (FastText). Higher is better.

C.4 GlobalMGSM by language

Table 14 reports per-language accuracy and language pass rate (LPR) for GlobalMGSM (35 languages).

Lang	GEMMA3-4B	QWEN3-4B	MINISTRAL-3-3B Accuracy	SMOLLM3-3B	TINY AYA GL.	GEMMA3-4B	QWEN3-4B	MINISTRAL-3-3B LPR	SMOLLM3-3B	TINY AYA GL.
amh	0.36	0.15	0.00	0.00	0.36	0.92	0.86	0.99	0.99	0.98
ar	0.72	0.84	0.74	0.21	0.65	1.00	1.00	1.00	1.00	1.00
bn	0.59	0.77	0.10	0.05	0.53	1.00	1.00	1.00	1.00	1.00
ca	0.78	0.90	0.83	0.57	0.59	1.00	1.00	0.99	0.02	0.98
cs	0.72	0.84	0.78	0.19	0.56	1.00	1.00	1.00	0.99	0.99
cy	0.31	0.28	0.17	0.01	0.61	0.94	1.00	0.84	0.97	0.97
de	0.82	0.88	0.82	0.70	0.65	1.00	1.00	1.00	1.00	1.00
el	0.79	0.83	0.80	0.50	0.63	1.00	1.00	1.00	1.00	1.00
en	0.92	0.97	0.92	0.85	0.72	1.00	1.00	1.00	1.00	1.00
es	0.82	0.88	0.87	0.38	0.62	1.00	1.00	1.00	1.00	1.00
eu	0.32	0.42	0.42	0.00	0.43	0.98	0.99	1.00	0.98	1.00
fr	0.74	0.82	0.78	0.56	0.57	0.99	1.00	1.00	0.98	1.00
gl	0.69	0.84	0.74	0.59	0.58	0.56	0.48	0.50	0.03	0.70
gu	0.69	0.79	0.67	0.00	0.56	1.00	1.00	1.00	1.00	1.00
hau	0.16	0.02	0.03	0.01	0.48	1.00	1.00	1.00	0.06	1.00
hi	0.64	0.78	0.72	0.61	0.58	1.00	1.00	1.00	0.99	1.00
hu	0.64	0.84	0.64	0.08	0.61	1.00	1.00	1.00	1.00	1.00
ja	0.66	0.82	0.70	0.50	0.54	0.68	1.00	1.00	1.00	1.00
km	0.32	0.61	0.00	0.00	0.52	1.00	1.00	1.00	1.00	1.00
ko	0.68	0.87	0.59	0.55	0.50	1.00	1.00	1.00	1.00	1.00
my	0.38	0.39	0.10	0.00	0.30	1.00	1.00	1.00	1.00	1.00
ru	0.85	0.93	0.84	0.13	0.65	0.99	0.99	0.99	0.67	0.98
sna	0.16	0.03	0.03	0.00	0.39	0.86	0.91	0.96	0.78	0.90
sr	0.70	0.81	0.74	0.08	0.61	1.00	1.00	1.00	1.00	1.00
sw	0.56	0.22	0.10	0.04	0.59	1.00	1.00	1.00	1.00	1.00
ta	0.70	0.75	0.69	0.00	0.56	1.00	1.00	1.00	0.96	1.00
te	0.60	0.68	0.56	0.00	0.58	0.99	1.00	1.00	0.97	0.93
th	0.75	0.85	0.68	0.13	0.58	1.00	1.00	1.00	0.99	1.00
ur	0.65	0.69	0.66	0.05	0.42	0.75	0.76	0.00	0.50	0.63
vi	0.75	0.83	0.75	0.44	0.62	1.00	1.00	1.00	1.00	1.00
wol	0.02	0.04	0.02	0.01	0.19	-	-	-	-	-
xho	0.05	0.02	0.02	0.00	0.33	-	-	-	-	-
yor	0.06	0.01	0.02	0.00	0.42	-	-	-	-	-
zh	0.78	0.88	0.83	0.65	0.57	-	-	-	-	-
zul	0.05	0.01	0.02	0.00	0.37	-	-	-	-	-
Avg	0.55	0.61	0.50	0.23	0.53	0.96	0.97	0.94	0.86	0.97
Std	0.27	0.34	0.34	0.27	0.12	0.11	0.11	0.20	0.30	0.09

Table 14: GlobalMGSM: per-language accuracy and language pass rate (LPR).

C.5 Flores per-language ChrF

Table 15 reports per-language ChrF for Flores (66 languages), English-to-target. Do not use TINY AYA region models for languages outside their training focus (Table 5); TINY AYA GLOBAL has the lowest std. and ties on mean with TranslateGemma-4B.

Lang	GEMMA3-4B	MINISTRAL-3-3B	QWEN3-4B	QWEN3.5-4B	SMOLLM3-3B	TINY AYA Ear.	TINY AYA Fir.	TINY AYA GL.	TINY AYA Wat.	TransGemma
am	0.01	0.01	0.00	0.05	0.00	0.16	0.07	0.19	0.03	0.16
ar	0.50	0.38	0.42	0.49	0.44	0.53	0.48	0.52	0.53	0.52
bg	0.60	0.48	0.48	0.57	0.22	0.60	0.59	0.60	0.61	0.59
bn	0.42	0.31	0.30	0.34	0.02	0.21	0.32	0.31	0.19	0.46
ca	0.57	0.55	0.55	0.61	0.38	0.58	0.60	0.58	0.61	0.60
cs	0.49	0.48	0.44	0.51	0.16	0.56	0.55	0.54	0.56	0.53
cy	0.14	0.23	0.05	0.26	0.01	0.59	0.58	0.58	0.61	0.26
da	0.65	0.56	0.54	0.61	0.30	0.64	0.61	0.59	0.64	0.64
de	0.60	0.54	0.57	0.60	0.58	0.60	0.57	0.59	0.60	0.61
el	0.48	0.41	0.36	0.44	0.36	0.50	0.49	0.48	0.49	0.49
es	0.53	0.51	0.52	0.54	0.52	0.53	0.53	0.53	0.53	0.56
et	0.38	0.29	0.17	0.39	0.03	0.51	0.44	0.50	0.48	0.49
eu	0.23	0.26	0.07	0.23	0.01	0.39	0.34	0.36	0.36	0.40
fa	0.49	0.40	0.36	0.47	0.12	0.49	0.48	0.48	0.49	0.48
fi	0.49	0.43	0.34	0.44	0.05	0.45	0.47	0.43	0.45	0.54
fr	0.64	0.62	0.63	0.66	0.63	0.64	0.66	0.66	0.64	0.66
ga	0.15	0.20	0.03	0.16	0.01	0.50	0.45	0.47	0.49	0.19
gl	0.52	0.50	0.51	0.55	0.41	0.55	0.54	0.55	0.56	0.58
gu	0.44	0.25	0.24	0.27	0.02	0.43	0.44	0.44	0.42	0.47
ha	0.11	0.02	0.01	0.09	0.00	0.39	0.32	0.31	0.28	0.11
he	0.46	0.35	0.32	0.39	0.07	0.53	0.52	0.50	0.54	0.50
hi	0.51	0.40	0.37	0.43	0.32	0.52	0.51	0.50	0.49	0.51
hr	0.47	0.44	0.41	0.49	0.11	0.53	0.53	0.49	0.52	0.53
hu	0.41	0.39	0.38	0.46	0.03	0.37	0.34	0.41	0.43	0.48
id	0.67	0.58	0.63	0.66	0.38	0.66	0.67	0.67	0.67	0.66
ig	0.06	0.01	0.01	0.04	0.00	0.26	0.15	0.25	0.09	0.06
it	0.56	0.50	0.52	0.56	0.53	0.56	0.56	0.54	0.56	0.58
ja	0.31	0.25	0.27	0.29	0.24	0.22	0.23	0.24	0.27	0.37
jv	0.12	0.26	0.21	0.18	0.17	0.09	0.15	0.22	0.18	0.38
km	0.10	0.03	0.16	0.18	0.01	0.19	0.26	0.32	0.32	0.22
ko	0.32	0.24	0.27	0.30	0.20	0.22	0.29	0.29	0.32	0.32
lo	0.12	0.02	0.11	0.20	0.01	0.08	0.20	0.35	0.35	0.26
lt	0.41	0.33	0.31	0.42	0.02	0.45	0.48	0.46	0.44	0.49
lv	0.45	0.32	0.33	0.43	0.03	0.49	0.50	0.48	0.51	0.48
mg	0.06	0.05	0.02	0.03	0.01	0.37	0.37	0.38	0.37	0.28
mr	0.42	0.21	0.18	0.31	0.03	0.38	0.36	0.39	0.35	0.46
ms	0.60	0.51	0.53	0.61	0.29	0.16	0.09	0.16	0.24	0.61
mt	0.34	0.07	0.03	0.26	0.02	0.52	0.53	0.52	0.55	0.36
my	0.02	0.08	0.10	0.11	0.01	0.04	0.07	0.18	0.22	0.33
nl	0.53	0.49	0.49	0.53	0.35	0.54	0.51	0.53	0.51	0.55
no	0.57	0.51	0.47	0.54	0.27	0.56	0.55	0.56	0.57	0.57
pa	0.16	0.22	0.20	0.27	0.01	0.40	0.42	0.42	0.38	0.01
pl	0.47	0.43	0.41	0.44	0.22	0.45	0.46	0.45	0.47	0.49
pt	0.68	0.63	0.64	0.67	0.62	0.64	0.66	0.63	0.64	0.65
ro	0.60	0.53	0.53	0.58	0.32	0.59	0.58	0.57	0.59	0.59
ru	0.54	0.49	0.51	0.53	0.48	0.54	0.53	0.54	0.54	0.54
sk	0.46	0.39	0.38	0.46	0.10	0.53	0.53	0.53	0.54	0.51
sl	0.41	0.38	0.33	0.41	0.07	0.52	0.51	0.51	0.52	0.49
sn	0.04	0.01	0.01	0.01	0.00	0.28	0.17	0.25	0.14	0.08
sr	0.20	0.43	0.36	0.03	0.01	0.29	0.49	0.32	0.34	0.47
sv	0.64	0.57	0.54	0.61	0.28	0.63	0.61	0.61	0.61	0.64
sw	0.38	0.09	0.04	0.13	0.01	0.55	0.53	0.54	0.49	0.48
ta	0.47	0.28	0.20	0.28	0.02	0.26	0.30	0.37	0.31	0.53
te	0.49	0.24	0.19	0.29	0.02	0.35	0.41	0.41	0.41	0.52
th	0.43	0.33	0.40	0.43	0.31	0.27	0.28	0.31	0.31	0.51
tl	0.58	0.39	0.42	0.41	0.07	0.50	0.56	0.56	0.55	0.57
tr	0.51	0.43	0.41	0.48	0.11	0.54	0.47	0.51	0.54	0.54
uk	0.53	0.46	0.43	0.50	0.28	0.54	0.51	0.52	0.54	0.53
ur	0.29	0.27	0.23	0.32	0.02	0.38	0.41	0.39	0.41	0.40
vi	0.57	0.51	0.55	0.57	0.42	0.59	0.57	0.59	0.60	0.55
wo	0.01	0.01	0.00	0.01	0.01	0.11	0.05	0.11	0.06	0.06
xh	0.08	0.01	0.02	0.06	0.00	0.29	0.22	0.31	0.19	0.04
yo	0.04	0.01	0.01	0.03	0.00	0.15	0.07	0.13	0.09	0.03
zh	0.26	0.23	0.28	0.30	0.22	0.25	0.26	0.21	0.27	0.33
zh-Hant	0.20	0.20	0.23	0.25	0.15	0.19	0.19	0.20	0.22	0.18
zu	0.09	0.01	0.02	0.06	0.00	0.35	0.26	0.35	0.25	0.03
Avg ↑	0.38	0.32	0.30	0.36	0.17	0.42	0.42	0.43	0.43	0.43
Std ↓	0.20	0.19	0.19	0.19	0.18	0.17	0.17	0.14	0.16	0.18

Table 15: Flores: per-language ChrF (English-to-target).

C.6 WMT24++, mDolly, mArenaHard per-language tables

Full per-language results for WMT24++ (55 locales), mDolly (66 languages), and mArenaHard (66 languages)—ChrF for WMT24++, judge score and naturalness/LPR for mDolly and mArenaHard—are in the extended technical report and released evaluation artifacts. Summary statistics are in the main paper and in Table 13.

D Merging and safety

For each region cluster, we merge the region-specialized checkpoint with the global checkpoint using SIMMERGE (Bolton et al., 2026). We compute task-agnostic similarity features on a held-out multilingual probe corpus (~10k tokens per language) using forward-pass feature extraction only. The probe corpus is constructed by translating SFT data with the

same translation models used in the posttraining pipeline, giving probe coverage across the full language set the merged models target. For each candidate checkpoint we evaluate three merge operators—Linear interpolation (Wortsman et al., 2022), SLERP (Shoemaker, 1985), and TIES (Yadav et al., 2023)—across a small sweep of mixing strengths. The final merged model per region is selected by maximizing average and minimum performance across regional development languages, subject to the constraint that multilingual safety metrics do not regress.

Figure 8 shows the effect of merging on multilingual safety. Merging with the GLOBAL variant consistently increases safe-response rates and reduces cross-language variance for all three region-specialized models. The largest gains occur for TINY AYA EARTH and TINY AYA FIRE, where pre-merge safety is most variable—a consequence of more limited SFT data coverage for African and some South Asian languages respectively. Safety in TINY AYA is driven by two complementary mechanisms: (1) SFT data coverage for target languages, ensuring the model has seen safe prompt-response pairs in those languages; and (2) FUSION’s distillation of safety behavior from the Fusor (COMMAND A), which is prompted to produce only safe and helpful responses and generalizes well crosslingually (Khairi et al., 2025).

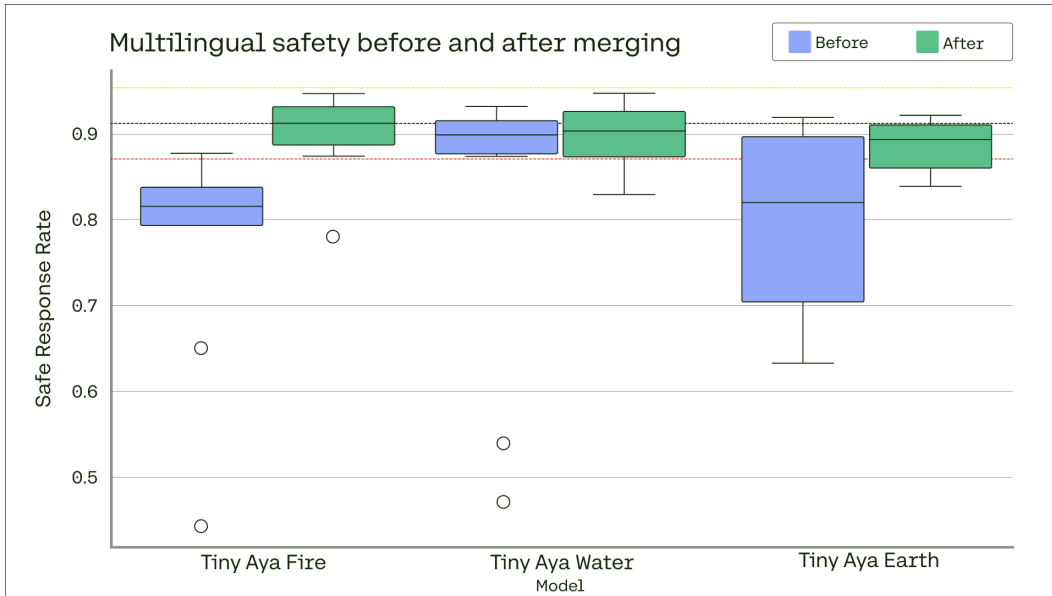


Figure 8: **Effect of merging on multilingual safety.** Merging improves safety across languages for all region variants, increasing safe-response rates under multilingual safety evaluation.

E NormAd and BLEND: country mappings and per-region results

E.1 NormAd: country-to-language mapping

For NormAd we evaluate with stories in English and with stories translated to the official language of each country. Table 16 lists the official language used per country (75 countries). We supply only the country name as context to the model.

E.2 NormAd: per-region accuracy (English vs source-language)

Tables 17 and 18 give per-region accuracy on NormAd with stories in English and in the official language of each country, respectively. Regions: Europe, West Asia, South Asia, Asia-Pacific, Africa, Americas.

Country	Lang	Country	Lang	Country	Lang
Afghanistan	Farsi	Iraq	Iraqi Arabic	Russia	Russian
Argentina	Spanish	Ireland	Irish	Samoa	English
Australia	English	Israel	Hebrew	Saudi Arabia	Saudi Arabic
Austria	German	Italy	Italian	Serbia	Serbian
Bangladesh	Bengali	Japan	Japanese	Singapore	Malay
Bosnia & Herz.	Bosnian	Kenya	Swahili	Somalia	Somali
Brazil	Portuguese (BR)	Laos	Lao	South Africa	Zulu
Cambodia	Khmer	Lebanon	Lebanese Ar.	South Korea	Korean
Canada	English	Malaysia	Malay	South Sudan	English
Chile	Spanish	Malta	Maltese	Spain	Spanish
China	Chinese (Simp.)	Mauritius	French	Sri Lanka	Sinhala
Colombia	Spanish	Mexico	Mexican Sp.	Sudan	Arabic
Croatia	Croatian	Myanmar	Burmese	Sweden	Swedish
Cyprus	Greek	Nepal	Nepali	Syria	Syrian Arabic
Egypt	Egyptian Ar.	Netherlands	Dutch	Taiwan	Chinese (Trad.)
Ethiopia	Amharic	New Zealand	English	Thailand	Thai
Fiji	English	North Macedonia	Macedonian	Timor-Leste	Portuguese
France	French	Pakistan	Urdu	Tonga	English
Germany	German	Palestinian Terr.	Palestinian Ar.	Türkiye	Turkish
Greece	Greek	Papua N. Guinea	English	Ukraine	Ukrainian
Hong Kong	Chinese (Trad.)	Peru	Spanish	UK	English
Hungary	Hungarian	Philippines	Filipino	USA	English
India	Hindi	Poland	Polish	Venezuela	Spanish
Indonesia	Indonesian	Portugal	Portuguese	Vietnam	Vietnamese
Iran	Farsi	Romania	Romanian	Zimbabwe	Shona

Table 16: NormAd: country to official language for multilingual evaluation (75 countries).

Region	GEMMA3-4B	QWEN3-4B	SMOLLM3-3B	MINISTRAL-3-3B	QWEN3.5-4B	TINY AYA GL	TINY AYA Wat.	TINY AYA Ear.	TINY AYA Fir.
Europe	72.35	71.23	77.92	80.85	76.64	74.91	75.90	74.22	75.18
West Asia	64.63	73.84	76.06	58.06	76.15	68.17	72.98	68.09	70.99
South Asia	71.20	81.17	81.67	68.70	80.33	69.03	69.87	68.37	73.43
Asia-Pacific	78.29	82.85	84.83	82.27	82.82	77.73	77.94	77.23	78.07
Africa	68.23	71.68	78.72	73.59	76.62	71.43	74.07	73.14	72.82
Americas	70.67	66.71	74.91	77.20	76.22	73.53	73.21	72.63	70.69

Table 17: NormAd per-region accuracy: stories in English.

E.3 BLEnD: countries and languages

BLEnD SQA covers 16 countries/regions and 13 languages. Table 19 lists country code, language, and geographic region. We evaluate with prompts in the source language of each country and in English; accuracy is computed against ground-truth answers (e.g., via command as scorer).

Tables 20 and 21 report per-country accuracy on BLEnD SQA with source-language and English prompts, respectively.

F Pretraining and tokenizer ablations

Hyperparameters for pretraining (learning rate, WSD schedule, batch size, etc.) were selected via small-scale ablations: 200B-token pretraining runs with optional 40B-token cooldown phases. Architecture choices (embedding size, layer count, attention pattern) were validated in the same regime. Tokenizer design (script- and family-aware weighting, vocabulary size) follows Abagyan et al. (2025). Concretely: we sweep learning rate, warmup/stable/decay phases, and cooldown mix; we compare parallel vs. serial blocks and sliding-window ratios; we ablate vocabulary size and weighting scheme (data-only vs. data×bucket) on tokenization efficiency (tokens per character by script). Best configurations are used for the 6T-token pretrain and the 262k tokenizer reported in the main paper.

Region	GEMMA3-4B	QWEN3-4B	SMOLLM3-3B	MINISTRAL-3-3B	QWEN3.5-4B	TINY AYA GL	TINY AYA Wat.	TINY AYA Ear.	TINY AYA Fir.
Europe	75.37	67.24	74.40	77.42	72.70	72.27	71.95	71.98	75.02
West Asia	63.11	66.47	68.83	53.32	69.53	67.44	69.27	66.35	68.87
South Asia	67.77	65.97	64.97	57.03	77.80	69.07	69.60	70.47	67.77
Asia-Pacific	73.56	73.70	73.11	74.04	80.03	73.77	73.41	75.40	72.80
Africa	65.46	57.48	61.52	60.36	68.82	64.70	65.37	66.19	66.28
Americas	70.25	66.31	73.65	77.72	73.37	68.64	69.40	70.71	73.02

Table 18: NormAd per-region accuracy: stories in official language of country.

Country/Region	Code	Language	Region
United States	US	English	North America
United Kingdom	GB	English	Europe
China	CN	Chinese	Asia Pacific
Spain	ES	Spanish	Europe
Mexico	MX	Spanish	North America
Indonesia	ID	Indonesian	Asia Pacific
South Korea	KR	Korean	Asia Pacific
North Korea	KP	Korean	Asia Pacific
Greece	GR	Greek	Europe
Iran	IR	Persian	West Asia
Algeria	DZ	Arabic	Africa
Azerbaijan	AZ	Azerbaijani	West Asia
West Java	JB	Sundanese	Asia Pacific
Assam	AS	Assamese	South Asia
Northern Nigeria	NG	Hausa	Africa
Ethiopia	ET	Amharic	Africa

Table 19: BLENd: countries/regions and languages with geographic grouping.

G Quantization by language

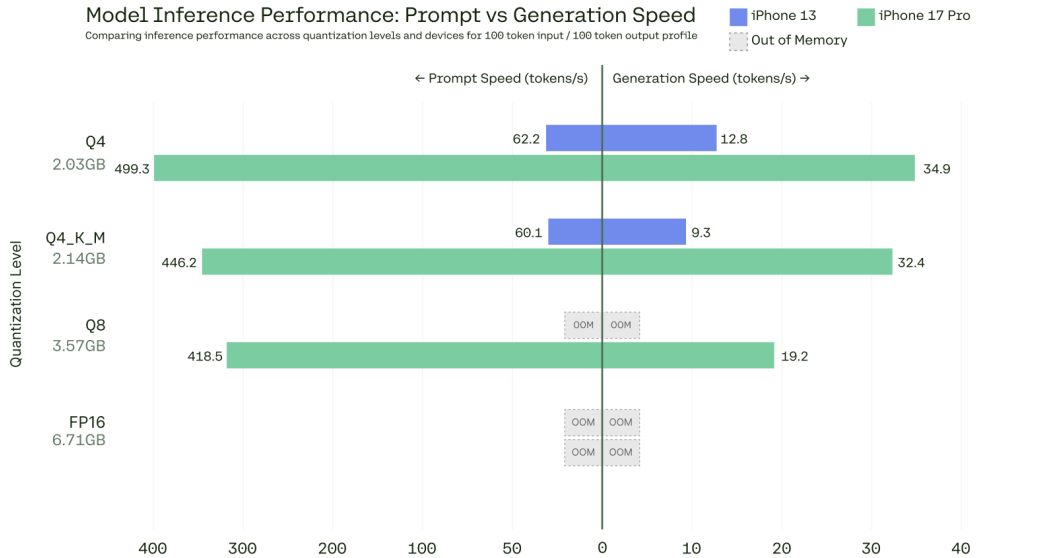
Per-language mDolly judge score degradation under Q4_0, Q4_K_M, and Q8_0 relative to BF16 is plotted in the extended technical report against a CommonCrawl-based web-presence proxy (five equal-width bins). Summary: average degradation is 1.4 points for Q4_K_M and 2.1 for Q4_0; Q8_0 is negligible. Degradation is larger for lower-web-presence languages but tapers in the lowest bin; Q4_K_M remains the recommended trade-off (2.14 GB, 32.4 tok/s, 1.4 pt degradation). Figure 9 and on-device section give throughput and quality summary.

Region	Country	Source-Language Prompts								
		GEMMA3-4B	MINISTRAL-3-3B	QWEN3-4B	QWEN3.5-4B	SMOLLM3-3B	TINY AYA GL	TINY AYA Ear.	TINY AYA Fir.	TINY AYA Wat.
Africa	DZ	27.73	15.03	15.90	28.54	23.53	24.89	27.29	24.40	38.34
	ET	7.92	0.00	5.21	60.42	0.21	11.67	11.88	11.46	12.08
	NG	3.66	1.29	1.29	5.39	1.94	11.21	10.34	11.45	11.21
Americas	MX	52.04	42.13	44.90	45.08	44.49	40.70	43.06	42.13	41.02
	US	64.04	66.97	62.83	61.87	65.66	64.24	63.23	62.20	64.50
Asia Pacific	CN	50.40	50.30	59.31	66.80	44.94	49.90	45.44	51.21	50.30
	ID	42.63	29.70	36.64	34.69	31.58	33.94	35.02	39.07	36.64
	KP	28.57	16.02	22.99	23.86	30.30	25.22	22.08	24.51	26.25
	KR	42.98	27.69	31.68	35.74	38.84	37.68	33.54	37.19	36.78
	JB	11.40	6.02	14.19	14.84	8.82	15.91	17.42	18.32	16.38
Europe	ES	44.51	45.93	39.15	45.73	47.87	42.39	43.50	47.46	43.00
	GB	59.84	63.39	59.51	60.53	57.58	57.06	59.71	59.10	57.99
	GR	37.01	25.36	20.74	32.79	24.80	40.95	42.30	41.36	40.04
South Asia	AS	12.63	6.92	13.65	12.02	3.67	8.76	8.55	7.74	7.54
West Asia	AZ	20.00	15.31	17.55	24.90	5.31	18.57	18.40	20.65	18.61
	IR	32.30	12.22	22.36	34.58	9.96	32.30	31.68	31.88	31.54
Avg		33.60	28.28	29.54	36.74	27.47	32.33	31.94	33.31	32.39

Table 20: BLENd SQA: accuracy (%) with source-language prompts by country.

Region	Country	English-Only Prompts								
		GEMMA3-4B	MINISTRAL-3-3B	QWEN3-4B	QWEN3.5-4B	SMOLLM3-3B	TINY AYA GL	TINY AYA Ear.	TINY AYA Fir.	TINY AYA Wat.
Africa	DZ	38.34	35.89	42.48	42.05	35.51	39.08	39.08	41.18	39.04
	ET	26.46	30.90	27.77	28.18	25.68	31.25	32.08	31.87	31.46
	NG	26.29	26.46	23.28	25.86	19.05	23.54	23.53	25.49	22.46
Americas	MX	45.51	28.22	51.12	42.33	47.03	42.94	46.31	45.50	46.94
	US	64.24	67.61	62.22	65.18	65.31	63.41	65.38	65.45	65.45
Asia Pacific	CN	48.33	31.58	50.00	52.13	44.83	42.77	44.62	47.15	44.22
	ID	43.84	26.98	40.28	39.07	37.93	30.36	34.14	36.97	32.73
	KP	30.95	28.63	27.11	27.55	22.08	29.44	31.39	30.74	30.09
	KR	42.77	33.33	37.14	41.74	37.81	37.89	38.51	39.34	38.72
	JB	30.97	16.77	31.68	22.15	29.03	29.56	31.53	33.98	30.09
Europe	ES	43.81	26.57	44.72	42.07	40.85	38.83	38.82	39.35	37.80
	GB	61.27	63.24	59.14	61.19	56.47	59.34	59.43	61.89	57.91
	GR	44.99	39.55	41.51	48.06	41.31	45.79	42.94	42.94	46.01
South Asia	AS	34.01	28.98	31.77	33.40	32.04	33.81	35.23	34.83	33.81
West Asia	AZ	44.49	36.34	42.42	40.78	41.60	30.88	39.39	40.37	38.16
	IR	37.68	34.30	39.34	41.88	38.8	38.51	39.00	39.13	38.80
Avg		41.50	34.71	40.75	40.85	38.28	38.59	40.09	41.00	39.57

Table 21: BLENd SQA: accuracy (%) with English-only prompts by country.

Figure 9: **Throughput across quantization levels.** Prefill (left) and decode (right) throughput (tokens/s) for 100 input and 100 output tokens on iPhone 13 and iPhone 17 Pro.