Cohere Tiny Aya: 3B-Parameter Multilingual Model Outperforms Larger Competitors in 46 of 61 Languages

Kabui, Charles

2026-02-22

Read at ToKnow.ai

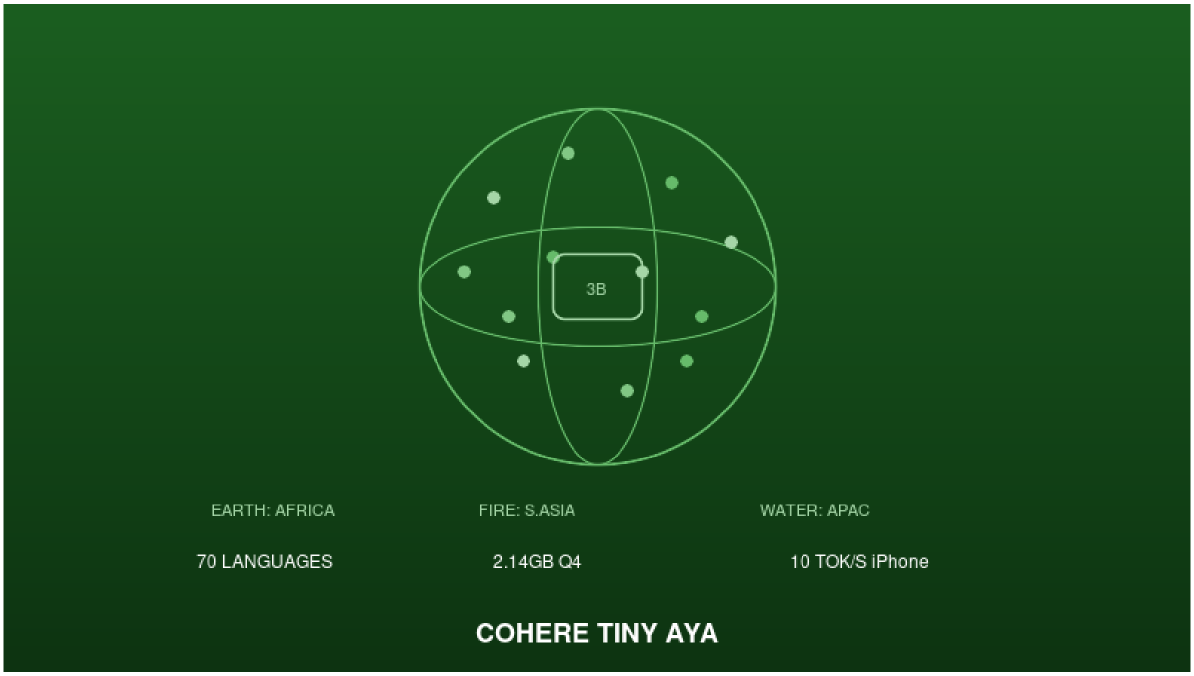

Figure 1: Cohere Tiny Aya

Cohere Labs released Tiny Aya, a 3.35B-parameter model family covering 70 languages. It ships as five models: a pretrained base, a globally balanced instruction-tuned variant, and three regional specializations for Africa/West Asia, South Asia, and Asia-Pacific/Europe. Training used Fusion-of-N, where multiple teacher models (Command A, Gemma3-27B-IT, DeepSeek-V3) generate responses and a judge picks the strongest parts to build synthetic training data. Regional checkpoints were merged with the global model using SimMerge so regional tuning doesn't erase safety behaviors. With 4-bit quantization, the whole model fits in 2.14 GB and runs at 10 tokens per second on an iPhone 13, losing only 1.4 points in quality. It beats Gemma3-4B in translation for 46 of 61 languages on WMT24++, and hits 39.2% on math reasoning for African languages where Gemma3-4B scores 17.6% and Qwen3-4B scores 6.25%.

Most multilingual models need cloud infrastructure or fall apart on low-resource languages. Tiny Aya runs on a phone, offline, no API required. The full post-training pipeline used a single 64-GPU H100 cluster. That is a small compute budget by current standards, and the results suggest that good data curation and distillation from larger teachers can do more than throwing extra hardware at the problem. Anyone building offline translation or education tools in areas with unreliable internet now has a viable local option.

Against Gemma3-4B, which is nearly identical in size, Tiny Aya is nearly identical in overall translation performance: it wins 46 of 61 language pairs, Gemma3-4B takes the other 15. The difference is that Tiny Aya does this in 2.14 GB on a phone with no cloud dependency. Where it clearly pulls ahead is low-resource languages. On African-language math reasoning it scores 39.2%, more than double Gemma3-4B's 17.6% and six times Qwen3-4B's 6.25%. That gap comes from a 262k-token vocabulary built for non-Latin scripts, synthetic data distilled from 27B-scale teachers, and region-specific model merging.

Sources:

- Cohere Labs Tiny Aya Blog
- Tiny Aya Technical Report
- MarkTechPost: Cohere Tiny Aya
- Tiny Aya on HuggingFace

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more: /terms-of-service***