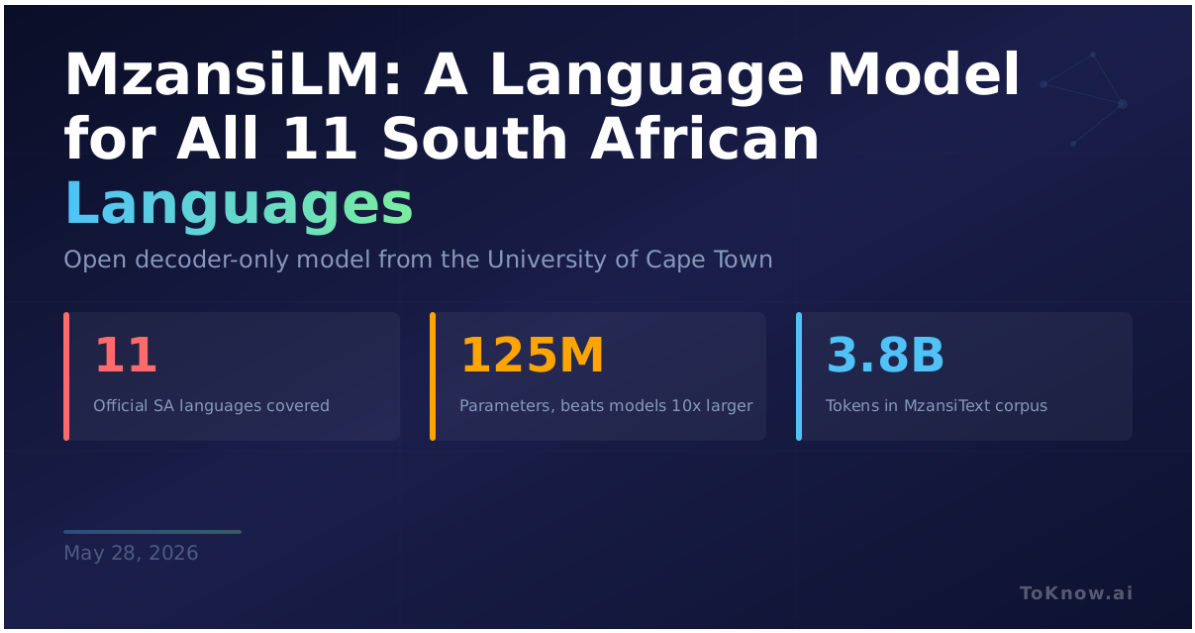
MzansiLM: A Language Model for All 11 South African Official Languages

Kabui, Charles

2026-05-28

Read at ToKnow.ai



Researchers at the University of Cape Town released MzansiLM, a 125M-parameter decoder-only language model trained from scratch on all 11 official South African languages. The model was pretrained on MzansiText, a curated 3.8-billion-token corpus aggregated from public sources including mC4, CulturaX, and WURA. It uses a LLaMA-based architecture with a custom 65,536-token BPE tokenizer designed for South African languages. Nine of the 11 target

languages are Bantu languages with limited digital text. Despite its modest size, MzansiLM reaches 20.65 BLEU on isiXhosa data-to-text generation, competing with encoder-decoder models over 10 times larger. Multilingual finetuning pushes isiXhosa news classification to 78.5% macro-F1. The model, corpus, training code, and a reproducible data pipeline are all released under Apache 2.0. The paper was accepted at LREC 2026.

Only 8.7% of South Africans speak English at home, yet most AI tools serve only English. MzansiLM covers isiZulu (24.4% of home speakers), isiXhosa (16.3%), Afrikaans (10.6%), and eight more official languages. The full open release (weights, data, code, pipeline) means developers can fine-tune it for local tasks: chatbots in Sesotho, document classification in Tshivenda, or text generation in siSwati. At 125M parameters, it runs on consumer hardware without a GPU.

The corpus reveals a harder truth about "coverage." Afrikaans and English together account for 84% of training tokens, while isiNdebele has just 818,000. For low-resource African languages, the bottleneck is not model architecture or compute. It is finding enough clean text in the languages that need AI tools the most.

Read more: Google's WAXAL dataset covers speech for 27 African languages, while Cohere's Tiny Aya tackles multilingual text across 70 languages on-device.

Sources:

- MzansiText and MzansiLM: An Open Corpus and Decoder-Only Language Model for South African Languages
- MzansiLM-125M on HuggingFace
- MzansiText Dataset on HuggingFace
- Training Code and Reproducible Pipeline on GitHub
