

Models That Read Raw Bytes Are Projected to Overtake Token-Based Models

Kabui, Charles

2026-10-21

[Read at ToKnow.ai](#)



Researchers at Meta FAIR and the University of Washington trained matched 1-billion-parameter models on two alphabets to see which ends up smarter. Almost every model reads text in [tokens](#), chunks from a vocabulary of 100,000 pieces. Their byte models read raw [UTF-8 bytes](#) instead, 256 values, so no vocabulary decides where a chunk ends. Fitted curves

project the byte version ahead: 52.4% average accuracy over eight benchmarks against 48.4% for the token model.

Token models win at small training budgets, then level off. Byte models start behind and keep improving. The best setup used distillation, where a small model learns by copying a bigger model's predictions. A [Llama 3-8B](#) teacher predicts the next token, and the paper converts that into a next-byte prediction in one pass. Its exact version, End-Of-Token, adds a marker after every 4.5 bytes to hold the leftover probability, so nothing the teacher knew is lost. It reaches the byte model's peak with 6.33e22 FLOPs (a count of arithmetic operations), about a third of the approximate method, and with one-sixth of the training text. Stored teacher output also shrinks to roughly a fifth, since each step saves 256 scores instead of 100,000.

The byte model passes the token model near 6.7 trillion byte units, seven times the largest run in the study, so that lead is extrapolated, not measured. A byte model also takes one step per byte, so it is slower to run at equal quality. Vocabulary size becomes a cost traded against sequence length, not a fixed part of building a model.

Read More: [Cohere Tiny Aya](#) took the opposite bet, a 262,000-token vocabulary sized for 70 languages.

Sources:

- [Breaking the Token Ceiling: Distilling Smaller, Stronger Byte Models \(arXiv\)](#)
- [Full paper with scaling tables \(arXiv HTML\)](#)
- [Byte Latent Transformer: Patches Scale Better Than Tokens \(arXiv\)](#)
- [UTF-8 encoding reference](#)

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