Nainsi Dwivedi
@NainsiDwiv50980

Holy shit... Microsoft open sourced an inference framework that runs a 100B parameter LLM on a single CPU.

It's called BitNet. And it does what was supposed to be impossible.

No GPU. No cloud. No $10K hardware setup. Just your laptop running a 100-billion parameter model at human reading speed.

Here's how it works:

Every other LLM stores weights in 32-bit or 16-bit floats.

BitNet uses 1.58 bits.

Weights are ternary just -1, 0, or +1. That's it. No floats. No expensive matrix math. Pure integer operations your CPU was already built for.

The result:

- 100B model runs on a single CPU at 5-7 tokens/second
- 2.37x to 6.17x faster than llama.cpp on x86
- 82% lower energy consumption on x86 CPUs
- 1.37x to 5.07x speedup on ARM (your MacBook)
- Memory drops by 16-32x vs full-precision models

The wildest part:

Accuracy barely moves.

BitNet b1.58 2B4T their flagship model was trained on 4 trillion tokens and benchmarks competitively against full-precision models of the same size. The quantization isn't destroying quality. It's just removing the bloat.

What this actually means:

- Run AI completely offline. Your data never leaves your machine
- Deploy LLMs on phones, IoT devices, edge hardware
- No more cloud API bills for inference
- AI in regions with no reliable internet

The model supports ARM and x86. Works on your MacBook, your Linux box, your Windows machine.

27.4K GitHub stars. 2.2K forks. Built by Microsoft Research.

100% Open Source. MIT License

7:11 PM · Mar 16, 2026 · **288.9K** Views

153 506 2.2K 3.1K

Read 153 replies

New to X?

Sign up now to get your own personalized

Sign up with Apple

Create account

By signing up, you agree to the Terms of S
Privacy Policy, including Cookie Use.

What's happening

Trending in Belgium
Lumumba

Politics · Trending
US and Israel

Trending in Belgium
xavier waterslaeghers

Trending in Belgium
Décidément

Show more