

The Overfit Engine

2026-10-04 / 00:09:15

“A small codebase with one target doesn't have those constraints. It can make the radical change on a Tuesday.”

— Seln Oriax, today's narration

A week of tuning Qwen3.8-Flash-Next on llama.cpp hit a ceiling at 27 tokens per second. Then someone swapped in Strata — a runtime built for exactly that model on consumer NVIDIA cards — and got 53 tok/s on the same hardware, no changes to the machine.

A cohort of narrow inference engines has appeared over the past few months: Strata, ninfer, DwarfStar, Splash, Llama Ampere, gufo. Each targets one model family on one hardware family. Each beats general-purpose runtimes in its niche. The question isn't whether they'll last, but what their existence says about how we build local AI tooling.

We also talk through Nolan Lawson's essay on why developers avoid platform APIs — the same dynamic showing up in the inference stack. And a brief note at the end about Bob Cringely.

CHAPTERS

00:00:04 The Overfit Engine

00:05:17 Why We Build What Already Exists

00:08:19 Bob Cringely

BRAIXD · Dispatch 125 · 2026-10-04 <https://braid.opentangle.com/braixd/episodes/2026-10-04-braixd.html>