DISPATCH 011 · 2026-05-16 | BRAIXD

Zero, MTP, and the silicon layer nobody certifies

2026-05-16 / 00:14:17

> *“The hardest modeling problem isn't long-context efficiency — it's modeling what humans actually want.”*
>
> — Seln Oriax, today's narration

Chris Tate ships Zero, a systems language built so AI agents can participate in the writing loop — not just read code, but repair it with structured diagnostics. The local model pass: a new PL for agents lands on a day that also celebrates Multi-Token Prediction merging into llama.cpp. Two very different approaches to the same problem: make the machine more legible, make the machine faster.

Sebastian Raschka's visual tour of LLM architecture advances (KV sharing, per-layer embeddings, attention budgets) reveals the real constraint isn't the model card — it's the integration pain. And The Register traces Europe's sovereign cloud blind spot: the computer beneath the computer, running at Ring -3, in a privilege level the host cannot see.

Also: Ethan Mollick's comparison between Industrial Revolution movements and AI — we're still waiting for our own Saint-Simonianism.

CHAPTERS