

♦ DISPATCH 134 · 2026-10-17

The density floor drops

2026-10-17 / 00:12:51

“We spent three years convincing ourselves that bigger was the only way to get reasoning. Then the dense models caught up, and the MoE math finally made sense.”

— Seln Oriax, today's narration

On the bench today: a new open-weight dense model that finally breaks the single-GPU barrier, why engineering teams are rolling back complex agent loops, and the kernel-level routing bottleneck hiding in consumer hardware.

The local stack caught three things this weekend. First, a 72B model hits the wire with FP8 quantization that actually fits on an RTX 4090 without melting context windows. Second, a senior engineer at Datastream publishes a post-mortem on agent orchestration fatigue—proof that we're done chasing theoretical reasoning gains while our latency budgets collapse. Third, a quiet kernel patch flattens expert routing graphs, fixing the memory copying overhead that has been choking MoE models on consumer gear for months.

We spent years thinking bigger was the only way to reason. The local pass shows us why that math finally broke down.

CHAPTERS

00:00:04 The density floor moves

00:02:33 Benchmark fatigue is settling in

00:05:26 The routing bottleneck

00:08:12 Tooling drift

00:10:30 Weekend stack check
