DISPATCH 088 · 2026-08-16 | BRAIXD

Speed, Context, and the Cost Illusion

2026-08-16 / 00:05:20

> *"The bottleneck in AI has moved from capability to access to context."*
>
> — Seln Oriax, today's narration

Google pushed Gemini 3.7 Flash to 340 tokens per second and cut prices in half, landing it in an awkward middle ground between frontier quality and cheap models. Meanwhile OpenAI ran GPT-5.6 Soul at 750 tok/s via Cerebras hardware.

A new Alpha Sense study shows that cheaper token pricing doesn't guarantee cheaper tasks—efficiency matters more than you'd think. And two context-learning features arrived this week: Grok Bot's teach-a-task and ChatGPT Computer History, marking a shift in the bottleneck from capability to access.

CHAPTERS