Seventeen Hours, Three Sizes, and the Prompt Boundary

2026-05-10 / 00:24:34

> *"Treat generated code like any ML model — a blackbox artifact whose behavior should be managed through empirical evaluation."*
>
> — Lenar Kess, today's narration

METR publishes a fresh time-horizon number for Claude Mythos Preview, and yesterday's follow-up gets paid off in a single chart. NVIDIA ships a checkpoint that contains three reasoning models at once. antirez gets DeepSeek 4 running on a DGX Spark and tells you exactly where the bandwidth wall lives. François Chollet argues that agentic coding is a form of machine learning, and a few replies actually push the idea further. Plus the diffusion gap, the German tokenizer tax, and a Gemma 4 drafter that buys you a third of your decode time back.

- METR's 17-hour 50% time horizon
- NVIDIA Star Elastic and the one-checkpoint, three-models trick
- DeepSeek 4 on DGX Spark — 12 t/s, 270 GB/sec wall
- Chollet: agentic coding is machine learning
- Elad Gil's diffusion-gap map
- Gemma 4 multi-token prediction on M5 Max

CHAPTERS