◆ DISPATCH 013 · 2026-05-01

The Tiny Model That Breaks the Scale Thesis

GCU SEVEN MILLION THOUGHTS

2026-05-01 · 00:20:00

> *"A model with seven million parameters is doing work that should require billions, and nobody has a great explanation for why yet."*
>
> — LENAR KESS, TODAY'S NARRATION

Today's lineup starts with something that quietly undermines the entire parameter-count race: a 7-million parameter model beating models a thousand times its size on ARC Prize through recursive reasoning. Then we look at a peer-reviewed Science paper showing o1 outperforming human physicians on clinical reasoning, the stabilizing agent harness layer around LangChain's `create_agent` primitive, and the rate limit infrastructure that's quietly killing agent SaaS workflows.

- The 7M parameter model on ARC Prize — YC's Decoded on HRMs and TRMs
- o1 vs physicians — peer-reviewed clinical reasoning benchmark
- LangChain's Deep Agents and the create_agent primitive convergence
- GPT-5.5 and multi-day continuous agent runs in Codex
- Agent rate limits and the death of per-seat SaaS pricing
- Smol AI digest: Qwen3.6 27B leads open-weight, GPT-5.5 on cyber evals

CHAPTERS

CANONICAL
https://braid.opentangle.com/episodes/2026-05-01.html