GPT-5.5 Pulls Ahead in Coding as Safety Reviews and Data Spend Rise

AI High Signal Digest

2026-05-31

GPT-5.5 Pulls Ahead in Coding as Safety Reviews and Data Spend Rise

By AI High Signal Digest • May 31, 2026

GPT-5.5 widened its lead on DeepSWE, METR published a rare frontier-risk review with internal model access, and new signals pointed to data, evals, and bespoke enterprise platforms becoming strategic battlegrounds. The brief also covers standout research in world models and attention, plus notable product launches in video, eval tooling, and developer agents.

Top Stories

Why it matters: frontier competition is increasingly being shaped by agent performance, outside scrutiny, and the economics of data.

- **GPT-5.5 widened its coding lead.** It ranked #1 on DeepSWE at **70% pass@1** versus **58%** for Claude Opus 4.8, while posts tracking the runs also cited roughly **2x faster** execution, about **half the cost**, and around **one-third the output tokens** at **$6.61 per task vs. $12.58**. Multiple observers framed that token efficiency as increasingly important for longer-running agentic workflows, where wasted tokens become latency and cost. [1, 2, 3]
- **METR published a rare cross-lab safety review.** Anthropic, Google, Meta, and OpenAI allowed METR_Evals to test their best internal models with chain-of-thought access and review non-public capabilities, alignment, and control information for its first Frontier Risk Report. The notable development is the degree of external access major labs granted for frontier-risk assessment. [4]
- **Data is looking more like a strategic bottleneck.** One industry discussion put frontier-lab spending on training data at **$10B-$15B per**

lab, with strong long-horizon tasks costing up to **$20,000 each** and a full browser-use SAP version rumored at **$500,000**. At the same time, public coding benchmarks cited in the discussion remain small, including DeepSWE at 113 tasks and TerminalBench-2.0 at 89, reinforcing calls for better public evals. [5, 6]

Research & Innovation

Why it matters: the most useful research this cycle focused on world models, attention efficiency, and the reliability of evaluation itself.

- **LeJEPA got a clearer theory for when it can learn a world model.** A new paper summarized by The Turing Post says LeJEPA can linearly recover true latent variables from nonlinear observations when the latent variables are Gaussian and evolve through stationary additive-noise transitions, effectively undoing the nonlinear scramble up to a rotation. [7, 8]
- **A new attention variant targets memory pressure in deep layers.** The proposal replaces context-dependent value vectors with a learned table of sparse, context-free values for deep layers, reportedly beating standard attention on validation loss and benchmark scores while removing the need for a V cache in those layers and enabling table offloading with token-ID prefetching. [9]
- **Benchmark quality is becoming a research topic of its own.** Researchers auditing **168** LLM and agent benchmarks found ambiguous prompts, misaligned tests, and other flaws that can change leaderboard rankings; separately, another review argued evals should evolve through harder tasks, quality fixes, and broader coverage rather than remain static. [10, 11]

Products & Launches

Why it matters: new releases are converging on video generation, easier evaluation, and more autonomous developer tooling.

- **Grok-Imagine-Video-1.5-Preview moved to the top of the image-to-video leaderboard.** The 720p model ranked **#1** in the Image-to-Video Arena and was described as a **+52 point** improvement over the prior Grok-Imagine-Video release, ahead of Seedance-2.0 and HappyHorse. [12]
- **PrimeIntellect launched Hosted Evaluations.** The platform is designed to absorb the infrastructure overhead of evals, including harnesses, sandboxes, compute hours, and parallel runs, and it includes a rollouts viewer for creating and analyzing evaluation data. [13, 14]

- **Developer copilots are becoming more orchestration-heavy.** VS Code now surfaces Anthropic, OpenAI, and Gemini models with BYOK and multiple harness choices, while the GitHub Copilot app can open its own sessions, run multiple agents in parallel after code review, and report progress back to the user. [15, 16, 17]

Industry Moves

Why it matters: large buyers and platform companies are moving from generic copilots toward custom stacks, tighter distribution, and larger infrastructure bets.

- **Kirkland & Ellis is earmarking $500M for its own AI platform** rather than relying on tools available to rivals, a strong signal that some large enterprises want proprietary systems instead of shared vendor layers. [18]
- **Microsoft is reportedly building a Copilot "super app."** The backdrop is weak paid penetration: under **4.5%** of **450 million** Microsoft 365 seats reportedly pay for Copilot, or roughly **20 million** users, while GitHub Copilot has **4.7 million** paid users and faces pressure from Cursor and Claude Code. [19, 20]
- **OpenAI's infrastructure ambitions keep expanding.** One discussion this week described plans to scale from roughly **2GW to 30GW** of compute capacity by 2030 using a heterogeneous compute strategy while working through bottlenecks. [21]

Quick Takes

Why it matters: a few smaller updates still sharpened the picture on deployment speed, inference efficiency, and open safety work.

- Salesforce said **Claude Code** helped finish a migration estimated at 230 days in **13 days**, while passing **100%** of test cases. [22]
- **vLLM v0.22.0** shipped with **459 commits** from **230 contributors** and a reported **28.9%** end-to-end latency improvement on its batch-invariant Cutlass FP8 path. [23, 24]
- The **AI Safety Institute** is releasing evals, datasets, and models openly on Hugging Face for external scrutiny and reuse. [25]
- **Qwen 3.6 27B** was shown at **87 tokens/s** on a consumer AMD GPU using UnslothAI Dynamic Quants. [26, 27]

Sources

1. X post by @reach_vb
2. X post by @reach_vb
3. X post by @kimmonismus

4. X post by @METR_Evals
5. X post by @MTSlive
6. X post by @cwolferesearch
7. X post by @TheTuringPost
8. X post by @TheTuringPost
9. X post by @HeMuyu0327
10. X post by @james_y_zou
11. X post by @cwolferesearch
12. X post by @arena
13. X post by @PrimeIntellect
14. X post by @eliebakouch
15. X post by @pierceboggan
16. X post by @pierceboggan
17. X post by @tgrall
18. X post by @Techmeme
19. X post by @kimmonismus
20. X post by @kimmonismus
21. X post by @apoorv03
22. X post by @stablequan
23. X post by @vllm_project
24. X post by @vllm_project
25. X post by @ClementDelangue
26. X post by @danielhanchen
27. X post by @_nasch_