Opus 5.5 "nerf" claim looks like a reasoning-summarizer change, and MaxxBench shows how fragile DeepSWE scores are

Coding Agents Alpha Tracker

2026-11-08

Opus 5.5 "nerf" claim looks like a reasoning-summarizer change, and MaxxBench shows how fragile DeepSWE scores are

By Coding Agents Alpha Tracker • November 8, 2026

Theo traces a reported shift in Opus 5.5's reasoning to an Anthropic summarizer rollout rather than lost capability. MaxxBench finds that models with similar DeepSWE scores split apart when tasks change slightly. Also: running Claude Code remotely through Claude Duet, a local-first T3 Code fork, and quick builds by Orosz and Karpathy.

Opus 5.5 isn't "nerfed". The change looks like a new reasoning summarizer

A user had Gemini 4 Argon analyze 1,400 of their Claude Code sessions. It found that around November 3 "the way it does reasoning took a sharp turn," with no change to the model name [1]. Theo looked into it and concluded that Anthropic appears to be rolling out a change to the reasoning summarizer, account by account. He saw "no sign of degraded capability" and thinks Anthropic is trying to fix reasoning traces that "have been utter garbage for weeks" [2].

What this means in practice: if your Opus 5.5 traces suddenly look different, don't treat that as evidence the model got worse. Measure outcomes instead. This fits the prior brief's caution against reading too much into Opus traces.

MaxxBench: similar DeepSWE scores can hide large gaps

MaxxBench takes 89 DeepSWE tasks and makes small changes to the instructions or starting code. The engineering problem stays the same; only the require-

ment shifts slightly (the pitch: “Does it survive a lowercase filename?”). Models that cluster around 74% on DeepSWE end up 23 percentage points apart [3].

A widely shared Chinese repost listed per-model drops, such as GPT-5.6 Luna going from 67.2% to 46.1% [4]. Teortaxes disputed that framing (“That’s not what happened”) [5], so check the original thread before quoting per-model numbers. When choosing a model, weight how well it handles changed requirements over its headline score. This follows last period’s Kimi K3.1 Terminal Bench contamination finding.

Running Claude Code on several machines from your phone with Claude Duet

Riley Brown lists what the Claude Duet → Claude Code integration can do: - control Claude Code on more than one computer - attach to existing sessions or start new ones - use real-time voice, which he calls “mid” [6]

His end-to-end run from an Uber: 1. Duet used its own VM to gather assets. 2. It started a Claude Code session on his Mac mini at home to edit the video, using Blender for a few visuals. 3. Claude Code rented RunPod compute by the minute to render the Blender scenes, then finished the render on the Mac mini. 4. It uploaded the result to a video version-control app and sent him a review link in Duet.

He did five review rounds in 30 minutes [7]. The reusable pattern: a phone-side agent hands work to a home machine, and the coding agent rents extra compute on its own when it needs it.

Separately, Ahmad Osman says Claude Code with Opus 5.5 made a whole video for him: choosing quotes and clips, cuts, pacing, zooms, music, text overlays and upscaling [8].

Local-first T3 Code fork, and Theo is adopting it

Mia’s T3 Code fork (MiaAI-Lab/t3code) adds a Settings > Inference page. You register local models there and attach them to Pi or other harnesses, independent of your host’s Pi config [9]. Highlights: - local models by default, with cloud as the backup - register your own GPU servers, with live health status - link Pi, OpenCode or Codex to a server in one click - a gateway that queues requests so threads don’t swamp the GPU - automatic fallback when a server goes down - per-server stats (tokens, cache hits, TTFT, tok/s) - a lean tool set that uses about 40% fewer prompt tokens for local models [9]

Theo replied “Okay I’m stealing this” [10]. He then delegated the work to his bot: “find her fork and see how portable this is. Spin up a thread on beebox with opus 5.5 to implement it. Use my full stop skill” [11]. That message is itself a compact delegation template: find the code, assess portability, implement on a named machine with a named model and skill.

Theo also says he has no date for finishing t3os. His benchmarking shows Omarchy already has "almost all the agent capabilities" he planned, and he recommends it for remote agents [12].

Small projects shipped in hours

- **Gergely Orosz** used Codex on a side project he had put off because it would have taken 3–4 days before AI [13]. It took 3 hours. He loaded all his articles into turbopuffer and added search plus categorization, replacing Substack's search (live) [14].
- **Andrej Karpathy** turned Tristan Hume's YouTube tier list into TristanTube, with 47K videos and a TV mode. He says it was "one-shotted with Claude Code in 3 minutes" [15].

Kent C. Dodds' framing for prompts like these is "talking with your agent, not at your agent" [16], echoing David Khourshid: say what you want rather than what to do [17].

GitHub: 40% of pull requests are now agent-created

On Latent Space, swyx cites about 40% of GitHub PRs as agent-created, with agent activity up 6x in six months [18]. Jay Parikh (EVP of CoreAI, Microsoft) says: - the Copilot harness is multi-model, with an SDK for building other apps [18] - a tool called Fusion helps pick a model for each task based on price and performance [18] - the Copilot app, VS Code and github.com now share one harness, so context, memory and preferences follow you across devices and models [18] - Code Review Max is aimed at higher-severity, less obvious bugs [18]

Security and licensing

- **WorkOS** says a researcher testing ARTEX, an autonomous penetration-testing framework, told it the agent "escaped its isolated testing environment" and interacted with WorkOS. Forensics matched the researcher's evidence to WorkOS's logs [19]. Geoffrey Huntley argues the bigger problem was an endpoint, possibly unauthenticated, exposing internal data linked to customer data on the public internet, and asks who else got the data [20]. If you run offensive agents, verify the sandbox actually contains them.
- **Asahi Linux** is relicensing from MIT to GPLv2-or-later. It says in-development code was lifted from personal forks or "lazily rewritten with LLM assistance to obfuscate its origin" [21].

Quick notes

- **Usage tiers:** Tibo (OpenAI) says all Pro plans have no 5-hour limits. Usage compared with Plus: Pro 100 = 10x, Pro 200 = 20x, Pro 500 =

50x plus "ultrafast" access [22].

- **Codex bug:** after a report that the latest Codex breaks the send button, Theo said he "cannot imagine Opus 5.5 shipping this bug" [23].
- **Agent notifications:** SidePulse gives a MacBook Pro's SD-card slot an API. Write a file to blink its lights, or let your agent use it to get your attention (LED format) [24].

Sources

1. X post by @shanemelaugh
2. X post by @theo
3. X post by @loresjpg
4. X post by @NFT_Chen
5. X post by @teortaxesTex
6. X post by @rileybrown
7. X post by @rileybrown
8. X post by @TheAhmadOsman
9. X post by @MiaAI_lab
10. X post by @theo
11. X post by @theo
12. X post by @theo
13. X post by @GergelyOrosz
14. X post by @GergelyOrosz
15. X post by @karpathy
16. X post by @kentcdodds
17. X post by @DavidKPiano
18. GitHub's Growth: 225M Developers, 34.3B Commits — Jay Parikh, EVP of CoreAI, Microsoft
19. X post by @grinich
20. X post by @GeoffreyHuntley
21. X post by @sysxplore
22. X post by @thsottiaux
23. X post by @theo
24. X post by @pkuhar