SemiAnalysis's GDPx Puts Every Model at 12% or Lower as Mythos 5.5 and GPT-6.2 Astra Surface

AI High Signal Digest

2026-11-02

SemiAnalysis's GDPx Puts Every Model at 12% or Lower as Mythos 5.5 and GPT-6.2 Astra Surface

By AI High Signal Digest • November 2, 2026

SemiAnalysis released GDPx, an agentic benchmark aimed at high-value domains, and argues that safety classifiers are skewing model rankings. Meanwhile, Anthropic and OpenAI both appear to be preparing new models, and evidence of AI-driven pressure on cybersecurity and compute keeps building.

GDPx: a frontier benchmark where safety classifiers change the rankings

SemiAnalysis released GDPx, a benchmark of agentic capability across what it calls "the most valuable domains." It describes GDPx as frontier-level and says even the best model scores just 12% [1]. The domains were chosen on purpose. SemiAnalysis expects the first trillion dollars of AI revenue to concentrate in sectors where frontier intelligence pays off most, so it included chip design, cyber and bio rather than government, real estate or insurance. It argues that benchmarks weighted by share of GDP or by occupational taxonomies are a mistake [2].

The main caveat: SemiAnalysis says Opus 5.5 and GPT-6 Astra have "sandbagged overall scores due to their safety classifiers." With cyber, biology and alignment excluded, the three domains where those classifiers can trigger, the ranking matched its own subjective impression of the models [3]. It also names Muse Wonder Full as the best available cyber model "by a meaningful margin," and says classifiers matter more for model choice than people assume [4].

The practical upshot is that one GDPx headline number can mislead buyers in regulated or dual-use domains.

Early reactions: - **Dylan Patel** issued an "official apology to Google," saying Gemini 4 Carbon is spikier across domains but "really good in a few areas" [5]. - **Elie Bakouch** found Sol 6.1 and Astra scoring very similarly, Carbon beating Argon across the board, and Sonnet strong on SWE-style tasks [6]. Once he removed the safeguard-affected benchmarks, Carbon looked very cheap for its performance and Sol 6.1 even cheaper [7]. - **Openness.** One critic, posting in a self-described joke, objects that GDPx is not open source, so its data and verifier code can't be audited [8].

New models from both labs appear close

Anthropic. Mythos 5.5 appears in the HTML source of Anthropic's official site, according to a tip passed to kimmonismus [9]. A follow-up says the entry was added within the last few days [10]. The tagline "our most capable model class for cybersecurity and biology research" refers to all Mythos models, not 5.5 specifically [11, 12]. There is no confirmed launch date.

OpenAI. Some Codex accounts now list `gpt-6.2-astra` as a separate line in usage analytics, while the model picker still shows GPT-6 Astra. The poster says multiple users see the same thing and calls it a gray rollout [13]. Kimmonismus expects a release "very soon," which is speculation [14]. Teortaxes argues that Astra 6.0 was already stronger than Opus 5.5 but lacked product-oriented post-training, and that 6.2 seems to fix this [15]. One user's complaint shows the cost problem: on corporate plans they exhaust Astra's weekly limit in a day, so they now use it only to review Opus 5.5's work [16].

On capability tracking, Teortaxes reports that DeepSeek V4.1F-1017 scores 159 on Epoch's Capabilities Index (ECI). That ties GPT 5.5 and is 10 points above the original V4-Pro, with Astra another 10 points higher [17]. He argues the index needs a refresh because Astra has saturated many of its evals [18]. Scaling01 counters that Chinese models still haven't reached Mythos Preview, at about 161, after nearly seven months [19].

DeepSeek's Super-agent: limits stated, critics speak up

Last period's Super-agent release reached Bloomberg, which framed it as a challenge to Claude Code [20]. Teortaxes noted the coverage came even though it was a "half-baked prototype" with no English announcement [21]. Tianyi Cui, one of the article's final reviewers, set out its limits: - "Super" refers to a layer above the working agent, not to greater intelligence [22]. - DeepSeek Harness had no long-term memory across sessions as of his post [22]. - The prototype implements only part of the design and has been validated only in a small, trusted internal environment [22].

Opinion split. Teortaxes calls the design "something of a hot mess" that leans on

crutches to work around hallucinations [23]. Mario Zechner (@badlogicgames) says Earendil is moving toward the same pattern: a single long-lived agent, recursive delegation, and event-driven triggers [24].

Cybersecurity: exploit data offers, an exploit anecdote, a fake leak

- **Exploit-data offer.** Jun Song says an anonymous black-hat group offered "an astronomical amount" for post-training open-weight models on exploit data, and claimed an attack corpus larger than anything he had seen. He says a white-hat GLM-5.3 project using under 1% of that volume "outperformed frontier models by a huge margin." He declined, and warns someone else will accept [25].
- **CVE-2026-98432.** One researcher calls this Linux TCP use-after-free severe: Go 1.24+ servers are exposed because MPTCP is on by default. He adds that it is likely exploitable mainly by an attacker on the same network [26]. His Mythos run spent about $2k over three days without producing a reliable proof of concept, and GPT-5.6 Sol couldn't get a memory leak. Anthropic, by contrast, reportedly needed 40 minutes. He suspects Anthropic used an unreleased model or an expert exploit developer. This is one person's anecdote [26].
- **Fake "Numbat" leak.** A Hugging Face repo claiming to hold a 14T-parameter model in the Mythos class drew attention [27]. Parts of it exactly match randomly generated numbers, which suggests the weights were never trained and the leak is likely fake [28].

Compute economics

- **NVIDIA inference results.** SemiAnalysis reports an NVIDIA AgentX submission reaching over 1,000 tok/s/user on 700B+ models, using software written by agent swarms. On Rubin NVL72 it claims 20.9× better performance per TCO than GB300 NVL72 at 325 tok/s/user [29]. The gain relies on Mooncake Store, which pools KV cache in Vera LPDDR5X memory across 18 trays over NVLink instead of RDMA, for over 10× the bandwidth [30].
- **China pricing.** Citing industry contacts, one poster says Blackwell prices in China tripled in four months, with an eight-GPU server at RMB 16M [31]. Teortaxes converts that to about $300K per B300 and estimates roughly a year to break even, even at DeepSeek's revenues [32].
- **US buildout.** SemiAnalysis says US datacenter capacity for 2027 is largely locked in, but a prolonged Texas permit freeze is now the biggest single risk for 2028 [33].

Briefly

- **Karpathy's HackerTube.** Claude filtered 43M Hacker News comments down to 6,454 repeatedly recommended videos, in about 2 hours with Opus 5.5 [34]. His only manual steps were an API key, a VPS and buying the domain; Claude secured, deployed and optimized the site [34].
- **Sakana Sawara.** Sakana AI is training its own language-model series from pretraining. Its first model, Sawara Small, is targeted for spring 2027 [35]. David Ha calls it an agent-native LLM built at Sakana's RSI Lab [36].
- **Suleyman on consciousness.** Mustafa Suleyman says he has seen no evidence that AIs are conscious, and that he would call for a halt if consciousness emerged during training [37]. He argues AIs should be designed to act as non-conscious, and calls for safety information sharing among labs [37].
- **MiniMax M3.1 Flash.** MiniMax highlighted community quantizations of the 428B model, including 2-bit GGUF weights at 143.9 GB and recipes for a single DGX Spark and for 4× RTX 3090 [38].
- **AI and productivity.** John Burn-Murdoch argues the US productivity trend predates AI and appears only in the US, so it doesn't look like an AI effect [39]. Gavin Baker reads the same FT charts as early "ROI on AI," citing productivity growth rising from 2.2% to 5% in the three industries that adopted AI most [40].

Sources

1. X post by @SemiAnalysis_
2. X post by @SemiAnalysis_
3. X post by @SemiAnalysis_
4. X post by @SemiAnalysis_
5. X post by @dylan522p
6. X post by @eliebakouch
7. X post by @eliebakouch
8. X post by @dejavucoder
9. X post by @kimmonismus
10. X post by @kimmonismus
11. X post by @scaling01
12. X post by @scaling01
13. X post by @TokenGremlin
14. X post by @kimmonismus
15. X post by @teortaxesTex
16. X post by @paul_cal
17. X post by @teortaxesTex
18. X post by @teortaxesTex
19. X post by @scaling01

20. X post by @business
21. X post by @teortaxesTex
22. X post by @ZhihuFrontier
23. X post by @teortaxesTex
24. X post by @badlogicgames
25. X post by @jun_song
26. X post by @S1r1u5_
27. X post by @OxShamss
28. X post by @kimmonismus
29. X post by @SemiAnalysis_
30. X post by @SemiAnalysis_
31. X post by @jukan05
32. X post by @teortaxesTex
33. X post by @SemiAnalysis_
34. X post by @karpathy
35. X post by @SakanaAILabs
36. X post by @hardmaru
37. X article by @mustafasuleyman
38. X post by @MiniMax_AI
39. X post by @jburnmurdoch
40. X post by @GavinSBaker