

DeepSeek’s V4.1-Flash-1017 Reaches GPT-5.6 Sol’s Level on Benchmarks at a Fraction of the Price

AI High Signal Digest

2026-10-18

DeepSeek’s V4.1-Flash-1017 Reaches GPT-5.6 Sol’s Level on Benchmarks at a Fraction of the Price

By AI High Signal Digest • October 18, 2026

DeepSeek shipped an open-weight Flash update that early testers place above GPT-5.6 Sol on several coding and security tests. The same testers raise concerns about benchmark gaming and distillation. Also: the OpenCode key breach, Terence Tao on OpenAI’s math result, and new training and inference research.

DeepSeek V4.1-Flash-1017: open weights at near-frontier scores

DeepSeek released V4.1-Flash-1017 on its website, app and API. It claims the update beats the previous version on benchmarks while finishing tasks faster with shorter outputs, which it calls “a Pareto improvement in performance and cost” [1]. The weights are on Hugging Face and ModelScope [2]. One summary lists the license as MIT and the context window as 1M tokens [3], and a tester puts the size at 763B parameters [4]. DeepSeek also opened its in-house Search and Fetch APIs to everyone at \$0.75 per 1,000 calls per endpoint. Calls made through server-side tools in DeepSeek Harness, Claude Code and Codex stay free for DeepSeek API users until billing starts on Oct 31 [5].

Reported numbers. These are DeepSeek’s own selected benchmarks, as summarized by one observer [3]:

Benchmark	Flash-1017	Previous Flash	GPT-5.6 Sol	Opus 5
DeepSWE	77.0	74.2	73.0	74.0
Terminal-Bench 4.0	42.3	31.2	39.9	51.8
Automation-Bench	60.5	54.8	45.8	50.3

Opus 5 still leads on Terminal-Bench 4.0. API prices are \$0.15/\$0.60 per million input/output tokens off-peak and \$0.30/\$1.20 at peak, against \$4/\$20 for Sol and \$5/\$25 for Opus 5. GPT-6 Luna is cheaper per token, at \$0.10/\$0.50 [3]. Cline says the model beats Sol on Terminal-Bench at about 17 $\times$ lower cost. It adds that the 35% gain over the previous Flash came from RL training on multiple coding-agent harnesses, not from “algorithmic innovation” [6].

Early outside tests mostly agree on the size of the jump: - **BugBunny VulnPR-100:** 54 of 100 known vulnerabilities found for \$7.53, the best open-weight result on that leaderboard. That matches Grok 4.6 at about 16 $\times$ lower cost and is three findings behind Gemini 4 Argon at about 22 $\times$ lower cost. The previous Flash found 44 for \$12.79 [7]. - **Real bug-fixing (Pawel Huryn’s test, 105 hidden bugs in two repos):** the model averaged 38.7 over three runs, with individual runs of 34, 37 and 45. That compares with 21.7 for the old Flash, 41.7 for Opus 5.5 and 42 for Muse Wonder Full [8]. - **Token use:** one tester saw an order of magnitude fewer chain-of-thought tokens in some tests [9]. - **Category breakdown:** on a leaderboard the poster himself calls not very high-signal, results were uneven. Agentic and coding scores rose, while browser use and research regressed [10], with Constraint Adherence down 33 points [11].

Hosting followed the same day. Ollama Cloud offers the model with US-based, zero-retention hosting and launch commands for Claude Code and OpenCode [12], and Databricks says it will host it [13].

Skeptics’ main points: - **Self-reported scores:** Lisan (@scaling01) urged waiting for private third-party benchmarks before trusting DeepSeek’s DeepSWE results [14]. - **LisanBench:** on his own benchmark he says he is “100% certain” DeepSeek trains on it with RL, citing a telltale chain of 64 two-letter words [15]. He also reports the model reasons about the grader in about half of runs, more often at higher reasoning budgets [16]. - **Distillation:** ModelTrace fingerprinting finds its numerical behavior most resembles GPT-6-Astra [17]. teortaxesTex, a regular DeepSeek defender, concedes it is unclear how much of Chinese model progress comes from distillation and how much from internal RL [18]. - **Agent behavior:** users say it finishes the core of a project quickly and then falls into “useless OCD fiddling” [19].

OpenCode breach exposes stored provider keys

OpenCode disclosed a breach tied to a “large scale distillation attack” that affected 1.4% of Console users, all of whom had linked external providers. It says the incident is ongoing and that it is working with Anthropic, OpenAI and

Google to fix it [20]. Users report unauthorized API charges, including \$400 on one Anthropic account. OpenCode’s customer email says third-party keys stored in workspaces may have been exposed through a legacy system that is being shut down [21]. If you have linked provider keys, check for unexpected charges.

Tao: OpenAI’s quasi-Riemann result is the biggest advance on the problem in 50 years

In a quoted interview, Terence Tao calls OpenAI’s quasi-Riemann Hypothesis work the biggest advance on the subject in 50 years: “a massive new zero-free region”. He says it combines three or four existing ideas “in a very unusual way” that no single living mathematician would likely have found [22].

Anthropic policy criticism

Sneha Revanur argues Anthropic plans to tell Congress it should be *harder* for the government to block deployment of a model judged unacceptably dangerous, including internal deployment. She notes this comes as Jack Clark says full recursive self-improvement may be possible within a year [23].

Research

- **Tencent, terminal-agent self-improvement:** starting from Qwen3.8-27B, five stages alternate between fine-tuning on verified trajectories and PPO. The four-suite Terminal-Bench average rose from 20.68% to 50.58%. A reward audit found failed runs collected 50.09% of the dense progress reward [24].
- **NVIDIA, open midtraining recipe:** NVIDIA added one stage of 256K-context training on 500B tokens of software-engineering data and left SFT unchanged. SWE-bench Multilingual on Nemotron 3.5 Lightning went from 55.8 to 62.8 [25]. NVIDIA found that base-model proxy metrics did not predict post-training results [25].
- **GLM 5.3:** in a quote shared by SemiAnalysis, a speaker says the model was built entirely by post-training the previous model, with a “night and day” improvement in cyber capabilities [26].
- **Next-Concept Modeling (Stanford/Together):** a new text-generation method that produces text by refining compressed codes from coarse to fine. It is 10–31× faster than the strongest diffusion baselines at comparable or better quality, but has only been tested on small models [27].
- **Apple RSI-Inference Bench:** tests whether agents can do real vLLM and SGLang engineering work. The best agent scores 26.7% pass@1 [28].

Inference and model sightings

- **Kimi K3 kernels:** vLLM reports 842 tokens/s at batch 1 decoding Kimi K3 on GB300 NVL72, using fused CuTe DSL kernels [29]. SemiAnalysis credits NVIDIA with a $1.68\times$ speedup on the same hardware, to 423 tok/s per user [30].
 - **Arena Alignment Index:** new entries are Muse Wonder (77.5%), Reflection’s Beam 1.0 (75.3%) and StepFun’s Step 5 Preview (75.1%) [31].
 - **OpenAI API:** three models named aurora-blue-alpha, aurora-green-alpha and aurora-red-alpha appeared. Their purpose is unconfirmed; one observer guesses specialized Codex models or fast GPT-6 variants [32].
-

Sources

1. X post by @deepseek_ai
2. X post by @deepseek_ai
3. X post by @kimmonismus
4. X post by @PawelHuryn
5. X post by @deepseek_ai
6. X post by @cline
7. X post by @Zaddyzaddy
8. X post by @PawelHuryn
9. X post by @janxin
10. X post by @teortaxesTex
11. X post by @teortaxesTex
12. X post by @ollama
13. X post by @Yuchenj_UW
14. X post by @scaling01
15. X post by @scaling01
16. X post by @scaling01
17. X post by @scaling01
18. X post by @teortaxesTex
19. X post by @teortaxesTex
20. X post by @opencode
21. X post by @kimmonismus
22. X post by @MehmetMars7
23. X post by @SnehaRevanur
24. X post by @omarsar0
25. X post by @dair_ai
26. X post by @SemiAnalysis__
27. X post by @TheTuringPost
28. X post by @onloglogn
29. X post by @vllm_project
30. X post by @SemiAnalysis__
31. X post by @arena

32. X post by @scaling01