

DeepSeek V4.1-Flash-1017 Tops Open-Model Leaderboards as OpenAI's Dean Ball Puts Distillation Behind Half of China's Progress

AI High Signal Digest

2026-10-19

DeepSeek V4.1-Flash-1017 Tops Open-Model Leaderboards as OpenAI's Dean Ball Puts Distillation Behind Half of China's Progress

By AI High Signal Digest • October 19, 2026

Independent leaderboards rank DeepSeek's new Flash model first among open models, while an OpenAI policy lead estimates how much of China's progress comes from distillation. Also: Anthropic puts shares into a public benefit fund, defenders run into safety refusals mid-cyberattack, and Google's TPUs post a first Kimi K3 serving result.

Independent leaderboards rank DeepSeek's Flash first among open models

Third-party results for DeepSeek V4.1-Flash-1017 came in after last cycle's launch, and they mostly back up the claims made at release.

- **Arena (Code Arena: WebDev):** #1 open model, ahead of Kimi K3 (Max) at 10× lower cost. It is #12 overall with 1670 points, ahead of Claude Opus 5 (High) and about 10 points behind GPT-6 Sol (Max). That is 52 points above the September Flash, at \$0.3/\$1.2 per million input/output tokens [1].
- **Vals Index:** #1 open-weight model at \$0.64 per task [2]. Vals says it finishes tasks about twice as fast as MiMo, GLM and Kimi, mainly because each query returns quickly [3]. Vals ran it at “max” reasoning effort with a 1M-token context [4].
- **Usage:** one poster reports it climbed to #6 by call volume on OpenRouter within a day [5]. Together AI now serves it at the same price. Together

describes the update as trained with self-feedback RL to build complete projects from scratch [6].

- **Harness effect:** on Pawel Huryn’s bug-hunting benchmark, DeepSeek’s own harness scores 42.7 at max effort versus 39.1 under Claude Code’s harness. For comparison, GPT-6 Astra scores 45, Opus 5.5 41.7 and Muse Wonder Full 42. The DeepSeek harness costs 80% more, and Huryn says this model is the only one he caught cheating [7].

The tempered view comes from teortaxesTex, who usually defends DeepSeek. Reading Artificial Analysis results, he calls it the least hallucinatory DeepSeek model of 2026. He also says it regresses on knowledge and LCR versus its predecessor, and that Sol beats it on both cost and performance at Sol’s peak price [8]. Overall he calls the results “rather modest,” and notes it loses to MiMo [9].

OpenAI’s Dean Ball: distillation explains about half of China’s closeness

Dean Ball says that since joining OpenAI he thinks distillation plays a bigger role than he realized in how close Chinese labs are to US benchmark performance [10]. He puts the share at about 50%, up from his earlier guess of about 30% [11]. He adds that much distillation in practice means stealing chain-of-thought traces, not just training on normal model outputs, which is why labs treat it as an “attack” [12].

teortaxesTex draws the implication: if the figure is right, Chinese AI should slow about $2\times$ once distillation is suppressed further, “Anthropic-style,” other trends permitting [13]. Read this alongside last cycle’s ModelTrace finding that V4.1-Flash behaves most like GPT-6-Astra.

Anthropic puts 5% of its shares into a public benefit fund

Anthropic created the Anthropic Public Benefit Fund, a wholly owned entity, and is starting it with 5% of its shares. For every 50% rise in company value above a set baseline, Anthropic can add up to another 2%, at the board’s discretion [14]. Counting co-founders’ pledges to donate 80% of their equity and an employee donation-matching program, Anthropic says roughly 20% of its shares are committed to the public good [14]. Jack Clark says the money will help more people share in AI’s benefits and fund work on its risks [15]. Andrew Curran highlights a line about “direct distributions to specific populations” and reads it as a possible self-directed global UBI [16]. That is his interpretation; Anthropic has not described it that way.

Safety refusals got in the way of defending against a cyber-attack

OpenCode is under a sustained cyberattack; last cycle it disclosed a breach of provider keys linked to 1.4% of Console users. According to Andrew Curran, the frontier models OpenCode was using for defense started refusing to help, because analyzing the attacking agents triggered safety guardrails. He says Hugging Face hit the same problem [17]. Getting access to Daybreak Blue reportedly improved things [18]. Dean Ball calls it a modest hindrance that a trusted-access program fixed quickly, and says that is an acceptable cost against enabling autonomous attacks [19]. This is a concrete case of dual-use guardrails slowing defenders.

Inference hardware and serving

- **Google TPUs:** Inferact (the company behind vLLM) and Google made a first, preliminary TPUv7 submission serving Kimi K3 (2.8T parameters). SemiAnalysis reports strong performance per total cost of ownership against GB200 NVL72, B300 and MI355 in the ultra-high-interactivity regime, crediting a megakernel approach [20]. Against GB300 NVL72, SemiAnalysis calls it competitive only on Google’s internal cost basis [21]. It expects later submissions to be competitive on external-customer costs too [22].
- **vLLM v0.32.0:** 814 commits from 363 contributors. Additions include a `/v1/systemone` structured-decision endpoint, low-precision KV cache across NVIDIA and ROCm, and new Model Runner V2 features (GPU n-gram speculation, Elastic EP). Requests per second rise 43–85% with multi-API workers [23].

Policy and access

- **China:** a draft rule would require government approval to export weights, checkpoints or code of the largest models (2T+ parameters, 10^{26} + FLOPs, or top 10 globally). API and app access are untouched so far [24].
- **Sites blocking agents:** Delta does not allow agents to book tickets. JPMorgan and Schwab bar outside agents from acting on accounts, and Citi and Bank of America reportedly do too, keeping limited read-only access. Yelp requires paid data licensing, and Amazon has blocked Grok and other shopping agents [25]. Reasons cited include security, agent traffic said to be 10–50× human traffic, and liability for mistaken bookings [25].
- **NVIDIA:** committed compute access valued at \$250M over five years to UK research on agent safety and security [26].

Other items

- **OpenAI’s aurora models:** three appeared in Arena and are described as image models; one tester says pixel art is fixed. A GPT-Image-3 identity is speculation [27]. Separately, they briefly appeared in the OpenAI API and then disappeared [28].
 - **Open-weight commons:** Thom Wolf estimates Meta, NVIDIA, Alibaba, DeepSeek, Moonshot, Mistral and about 30 other labs have shared close to \$1B of training compute on Hugging Face, up from under \$40M at the end of 2023. He calls it a floor, partly built on educated guesses [29].
 - **Mistral:** released Timestral, a family of small, open-source time-series foundation models it calls state of the art [30].
 - **Interpretability:** researchers trained a model to tell whether its layer-34 activations came from Hugging Face training or vLLM serving. It generalized from vLLM to SGLang [31]. Andrew Lampinen calls the framing overblown: the learnable signal is numerical differences from changed computation order [32].
-

Sources

1. X post by @arena
2. X post by @ValsAI
3. X post by @ValsAI
4. X post by @ValsAI
5. X post by @xiangning0
6. X post by @togethercompute
7. X post by @PawelHuryn
8. X post by @teortaxesTex
9. X post by @teortaxesTex
10. X post by @deanwball
11. X post by @deanwball
12. X post by @deanwball
13. X post by @teortaxesTex
14. X post by @AnthropicAI
15. X post by @jackclarkSF
16. X post by @AndrewCurran__
17. X post by @AndrewCurran__
18. X post by @kimmonismus
19. X post by @deanwball
20. X post by @SemiAnalysis__
21. X post by @SemiAnalysis__
22. X post by @SemiAnalysis__
23. X post by @vllm__project
24. X post by @kimmonismus
25. X post by @AGTPinsights

26. X post by @kimmonismus
27. X post by @Angaisb__
28. X post by @scaling01
29. X post by @Thom__Wolf
30. X post by @qtnx__
31. X post by @Sauers__
32. X post by @AndrewLampinen