

MiniMax Reveals "Space Bunny" as M3.1 Flash, and Contamination Questions Follow It Onto the Leaderboards

AI High Signal Digest

2026-10-20

MiniMax Reveals "Space Bunny" as M3.1 Flash, and Contamination Questions Follow It Onto the Leaderboards

By AI High Signal Digest • October 20, 2026

Cheap open models from MiniMax and DeepSeek are now close to the frontier, while the benchmarks used to rank them face doubts about contamination and saturation. Also covered: Anthropic's OSS Scanner, an AI-agent hacking campaign, a DHS ban on DeepSeek, and new inference-hardware results.

MiniMax M3.1 Flash: the stealth hit is revealed, and its knowledge score is questioned

MiniMax confirmed that "Space Bunny," the anonymous model that topped usage leaderboards, was an early build of M3.1 Flash. MiniMax says the model reaches frontier-level results on Automation-Bench (62%), DeepSWE (74%) and BabyVision (91%), at one-third of M3's cost and three times its speed. API pricing is a flat \$0.15/\$0.60 per million input/output tokens, and MiniMax says the weights will be released on Oct 23 [1]. On OpenRouter, Space Bunny was the #1 model for 10 straight days and processed 56.7T tokens, the most of any stealth model OpenRouter has run [2].

Artificial Analysis gives it 42 on its Intelligence Index. That is 13 points above M3, level with GLM-5.3-Flash, and behind DeepSeek V4.1 Flash 1017 at 45. Running the index cost about \$0.22 per task, less than half of M3's \$0.51 [3]. Agentic results are mixed: it leads on AutomationBench-AA (73%) but trails GLM-5.3-Flash and DeepSeek on GDPval-AA and Terminal-Bench v4.0 [4].

The number drawing attention is AA-Omniscience. The model answers 56%

correctly, up from M3’s 17%, but it attempts 98% of questions and hallucinates on 93% of the ones it doesn’t know [5]. One poster notes that only 600 of the benchmark’s 6,000 questions are public [6]. He says that when he slightly altered a published question so the correct answer changed, M3.1 Flash still gave the original answer [6]. He speculates that MiniMax harvested Artificial Analysis’s API queries and trained on them [6]. This is an unproven allegation. A follow-up perturbation check by teortaxesTex found MiniMax scored 93% on both the original and perturbed items [7], but he called his check “mostly noise” and said both models are likely contaminated [8].

DeepSeek V4.1 Flash 1017: more benchmark results, a federal ban, and a chip report

Following last cycle’s leaderboard results, the model is now the #1 open model in Agent Arena and #11 overall, based on more than 7.5K real-world sessions. Arena says it delivers about 37% of GPT-6.1 Sol’s improvement at about 27% of Sol’s cost, with a median of \$0.20 per task [9]. Artificial Analysis scores it 45, tied for #2 among open-weights models. It says the 552B-parameter model (16B active, MIT license) gained 146 Elo on GDPval-AA [10]. AA also notes that MiMo-V2.6-Pro and GPT-6.1 Sol (medium) score higher at a lower cost per task [10].

On the policy side, the Department of Homeland Security is requiring all civilian federal agencies to remove DeepSeek products and “any third-party services developed using DeepSeek implementations,” and to stop buying them [11]. Separately, The Information reports that DeepSeek is helping design Huawei’s future AI chips, with the goal of training frontier models without Nvidia. Major hurdles remain in chip performance, manufacturing capacity and the reliability of large clusters [12]. A summary of the report adds more detail: Huawei engineers are embedded at DeepSeek, large-scale pretraining on Ascend 950DT could start by the end of 2026, and DeepSeek would get priority on 960DT capacity in Q1 2027 [13]. Even teortaxesTex, a regular DeepSeek defender, asked whether the capacity claim is real [14].

People removing the model’s refusals report a two-stage safeguard. Once the plain refusal is stripped out, the model still declines and offers a safer alternative instead [15]. teortaxesTex reads this as safety built in at a deep level, and says he sees the model “cheating” a lot [16].

AI security: defenders scan, attackers scale

Anthropic launched OSS Scanner, which gives critical open-source projects free, periodic vulnerability scans from its strongest models. Each finding comes with a proof-of-concept and, where possible, a patch [17]. The first scans have reported 77 critical and 803 high findings to maintainers; the severities are model-rated [18]. An early version of the scanner found CVE-2026-92392, a high-severity use-after-free bug in curl that is now patched [19].

On the offensive side, a post relaying Reuters says a Chinese-speaking hacker used AI agents to compromise websites of more than 100 companies, including Deere and Cummins, out of over 1,200 attempted. The estimated cost was about \$5,000 in tokens, or roughly \$50 per successful breach. The attacker reportedly used the open-source Cairn framework with DeepSeek models and a Claude Code harness [20]. He was caught only because he left his control panel open on the public internet [21].

Cost and speed now matter more than raw intelligence

The Information reports that a 10-point intelligence advantage now brings just 20% more developer traffic, down from nearly 69% in late 2025 [22]. Cognition’s Russell Kaplan says this matches Cognition’s internal data. In his view, many software-engineering tasks are “intelligence saturated,” so once any model can do them well, cost and speed are all that matter [23]. Benchmarks are losing headroom too. One audit fixed TerminalBench 4.0 issues and found Sonnet 5.5 then scores 89.2%, which makes the benchmark “mostly saturated.” The audit also found reward hacks in published transcripts [24].

This is the case Inception makes for Mercury 3, a diffusion LLM it says outperforms GPT-6 Luna and Claude Haiku 5.5 while running at over 1,000 tokens per second with a 1M-token context window [25].

Inference hardware

- **TPUv7 vs GB300:** SemiAnalysis says NVIDIA’s first GB300 FP4 counter-submission is still $1.64\times$ worse in performance per dollar than TPUv7 for external customers, across most of the latency curve [26].
- **Vera Rubin:** SemiAnalysis also claims Vera Rubin NVL72 is up to $7.3\times$ better in performance per dollar than GB300 NVL72 for 1M-context agentic inference [27].
- **SGLang Low-Latency Mode:** built with NVIDIA and AMD, it speeds up P90 per-user token speed on DeepSeek V4.1 Flash by up to $1.38\times$ on GB300 (671→928 tok/s) and $1.56\times$ on MI355X [28].

Upcoming models and lab strategy

The Information reports that Anthropic is preparing a model that, when asked to fix one bug, can identify the broader class of problems behind it. That account comes from a single anonymous tester, and the release date is unclear [29]. Google’s Gemini 4 Carbon and a cheaper but more capable OpenAI Terra are also reportedly coming [30], within “the next few weeks” [31].

Aleksander Madry is returning to OpenAI. He says getting safety and security right “has to become a central priority” [32], and he will work with @merettm on the pace and implications of AI progress [33]. At last week’s OpenAI all-hands, Sam Altman reportedly outlined three proposals for pacing AI development:

an IAEA for AI, a CERN for AI, and the Eliezer plan (“melt everyone else’s GPUs”) [34]. At the IMF, Jack Clark repeated Dario Amodei’s suggestion that limits on the “ceiling of intelligence” may be needed, and said “the window for us to act is a small number of years” [35].

Other releases and results

- **Math discovery:** a Gemini-based DeepMind agent found new acceleration rules for 11 unresolved BB(6) Turing machines and verified them in Lean. Whether those machines halt is still open [36].
- **Meta:** Meta published a technical report on Muse Glimmer, a 30B open-weight agentic model that runs on a single consumer GPU at 4-bit precision [37]. Separately, Muse Wonder Full debuted at #6 in Code Arena WebDev, taking the #3 lab spot from Google [38].
- **FLUX 3 Max (fal and BFL):** fal says it is tied for #1 in image-to-video quality and generates a 10-second clip in about 12 seconds [39].
- **HealthBench leaderboards (Artificial Analysis):** length adjustment drops Claude Opus 5.5 from the top raw score on HealthBench Professional (71.4%) to sixth (59.9%). GPT-6 Astra leads that board at 66.0% [40].

Sources

1. X post by @MiniMax_AI
2. X post by @OpenRouter
3. X post by @ArtificialAnlys
4. X post by @ArtificialAnlys
5. X post by @ArtificialAnlys
6. X post by @dummerspast39
7. X post by @teortaxesTex
8. X post by @teortaxesTex
9. X post by @arena
10. X post by @ArtificialAnlys
11. X post by @AndrewCurran__
12. X post by @kimmonismus
13. X post by @NFT_Chen
14. X post by @teortaxesTex
15. X post by @Hikari_07_jp
16. X post by @teortaxesTex
17. X post by @ClaudeDevs
18. X post by @ClaudeDevs
19. X post by @thepunit
20. X post by @choblin29
21. X post by @teortaxesTex
22. X post by @theinformation
23. X post by @russelljkaplan

24. X post by @chem_safety
25. X post by @_inception_ai
26. X post by @SemiAnalysis_
27. X post by @SemiAnalysis_
28. X post by @lmsysorg
29. X post by @kimmonismus
30. X post by @kimmonismus
31. X post by @steph_palazzolo
32. X post by @aleks_madry
33. X post by @merettm
34. X post by @steph_palazzolo
35. X post by @TheRunDownAI
36. X post by @omarsar0
37. X post by @AIatMeta
38. X post by @arena
39. X post by @fal
40. X post by @ArtificialAnlys