Kimi Attribution Debate, New Open Models, and the Autonomous Research Push

AI High Signal Digest

2026-03-21

Kimi Attribution Debate, New Open Models, and the Autonomous Research Push

By AI High Signal Digest • March 21, 2026

Composer 2's Kimi foundation was publicly confirmed as Mistral and NVIDIA shipped new open models, while OpenAI and DeepMind made autonomous research a more concrete roadmap.

Top Stories

Why it matters: The leading stories were about how frontier capability is being assembled: open-model adaptation, smaller open reasoning systems, and increasingly autonomous research workflows [1, 2, 3].

1) Composer 2's base model moved from rumor to public confirmation

Cursor launched Composer 2 while saying its in-house models generate more code than almost any other LLMs in the world, and a developer quickly surfaced the model ID `kimi-k2p5-rl-0317-s515-fast` from the API response; Moonshot's head of pretraining said the tokenizer matched Kimi's [4].

Moonshot later said Kimi-k2.5 provides the foundation for Composer 2, with Cursor adding continued pretraining and high-compute RL, and said Cursor accesses Kimi through Fireworks' hosted RL and inference platform under an authorized commercial partnership [1].

Cursor said Composer 2 started from an open-source base, that only about one quarter of the compute spent on the final model came from that base, and that it is following the license through its inference partner terms. Cursor also said not mentioning the Kimi base in the launch blog was a miss [5, 6].

The debate has now shifted to disclosure and measurement: critics said public benchmark reporting still makes improvement over the base model hard to as-

sess, while others argued the episode validates a broader shift toward adaptation, fine-tuning, and productization over training from scratch [7, 8, 9, 10].

Impact: Open-model licensing and attribution are becoming product issues, not just legal footnotes, and the strongest coding products are increasingly being built by post-training on top of open bases [4, 9, 10].

2) Mistral Small 4 strengthened Mistral's open model lineup

Mistral Small 4 is a 119B MoE with 6.5B active parameters, hybrid reasoning and non-reasoning modes, and native image input, scoring 27 on the Artificial Analysis Intelligence Index in reasoning mode. That is 12 points above Small 3.2 and above Mistral Large 3's 23 [11].

The model used about 52M output tokens on the index, scored 57% on MMMU-Pro, reached 871 Elo on GDPval-AA, and posted a -30 AA-Omniscience score, ahead of peers on hallucination even while trailing the top open-weight models of similar size on raw intelligence [11, 12].

Mistral lists a 256K context window, Apache 2.0 licensing, pricing of $0.15 and $0.60 per 1M input and output tokens, and API availability through Mistral's first-party API [11].

Impact: Small 4 improved Mistral's position on efficiency, multimodality, and agentic evaluation, but the comparison set shows how competitive the open-weight 120B class has become [11, 12].

3) NVIDIA compressed frontier-style reasoning into a much smaller open model

Nemotron-Cascade 2 is an open 30B MoE with 3B active parameters that NVIDIA says delivers best-in-class reasoning and strong agentic capabilities [2].

NVIDIA says it reached gold-medal-level performance on IMO 2025, IOI 2025, and ICPC World Finals 2025, matched capabilities previously associated with frontier proprietary or frontier-scale open models, and did so with 20x fewer parameters [2].

The model also reportedly outperforms recent Qwen 3.5 releases across math, code reasoning, alignment, and instruction following, and is built with Cascade RL plus multi-domain on-policy distillation [2].

It is already available on Hugging Face and can now be run locally through Ollama [2, 13, 14].

Impact: Open reasoning models are getting smaller without giving up top-tier tasks, which matters both for local deployment and for the pace of open-model iteration [2, 14].

4) OpenAI put dates on its automated research roadmap

Notes from an interview with chief scientist Jakub Pachocki say OpenAI is targeting an automated AI research intern for September 2026 and a multi-agent automated AI researcher for 2028 [3].

The 2028 system is described as a multi-agent setup that could tackle problems too large or complex for humans and, in theory, be applied to problems expressible in text, code, or whiteboard sketches across math, physics, biology, chemistry, business, and policy [3].

Pachocki also said OpenAI is getting close to models that can work indefinitely in a coherent way, like a whole research lab in a data center. At the same time, he does not expect systems to match humans in all ways by 2028, and another summary of the interview said current reasoning models and agent systems like Codex already show large productivity gains while still facing reliability and safety limits [3, 15].

Impact: OpenAI is treating multi-agent research automation as a staged product roadmap, not just a long-range vision, while explicitly tying that roadmap to reliability and safety constraints [3, 15].

5) DeepMind's Aletheia added another fully autonomous math result

Aletheia, powered by an advanced version of Gemini Deep Think, has now contributed to eight math research papers, and its most recent result on the Hodge bundle was described as fully autonomous Level 2 publishable research [16, 17].

In that case, mathematician Anand Patel had the intuition but could not assemble the proof; Aletheia produced the construction needed to complete it, and Google DeepMind released both the paper and the interaction transcript [17, 18].

Earlier Aletheia work included solving 6 of 10 FirstProof challenge problems autonomously and helping resolve bottlenecks in 18 research problems across algorithms, machine learning, combinatorial optimization, information theory, and economics [19, 20, 21].

Impact: Claims about autonomous research are getting harder to dismiss as benchmark theater when they come with publishable outputs and public transcripts [16, 17].

Research & Innovation

Why it matters: Several of the most useful technical advances were about training data strategy, specialized RL, and evaluation—areas that often matter more in practice than a single flagship model release [22, 23, 24].

- Datology's Finetuner's Fallacy argues that standard pretrain-then-finetune domain adaptation leaves performance on the table. Mixing just 1-5% domain data into pretraining before finetuning produced better models across chemistry, symbolic music, and formal math proofs, including 1.75x fewer tokens to reach the same domain loss, a 1B model beating a 3B finetune-only model, +6 MATH points at 200B pretraining tokens, and less forgetting of general knowledge [22].
- Separate work on synthetic data argued that generated data can reduce loss on the real distribution as more tokens are produced. Treating generations as one long megadoc gave a further 1.8x data-efficiency gain, on top of a previously reported 5x gain from tuning, scaling, and ensembles [25, 26].
- Mantic said it RL-tuned gpt-oss-120b on judgmental forecasting and got a model that outperformed frontier models on event prediction. It also said the tuned model plus Grok were decorrelated from the other best models, making them especially useful in team settings [23].
- Meituan released LongCat-Flash-Prover, an open-source theorem-proving model with a hybrid-experts trajectory-generation framework, the HisPO algorithm for long-horizon tool-integrated reasoning, and a verification stack using Lean4, AST checks, and legality detection. Reported results were 97.1% on MiniF2F-Test and 41.5% on PutnamBench [27].
- CodeScout introduced an RL recipe for teaching code agents to search large codebases using only a terminal. The authors said it outperforms open-source models 18x larger, is comparable to proprietary models, and sets state of the art on SWE-Bench Verified, Pro, and Lite [28, 29].

Products & Launches

Why it matters: Product teams kept turning model capability into concrete workflow features—especially around agents, multimodality, and developer control surfaces [30, 31, 32].

- Google's Gemini API now exposes Veo 3.1 video generation and Gemini image models through its OpenAI compatibility layer, with no SDK swap required. Google says developers can call `/v1/videos` for video, `images.generate` for images, stay compatible with OpenAI Python and JS SDKs, and switch by changing three lines of code [30, 33].
- Cognition added scheduled Devins. A user can run a task once—such as feature-flag cleanup, release notes, or QA—and then make it recurring so a one-off session becomes an automated workflow [31, 34].
- Anthropic added Projects to Cowork, letting users keep tasks, files, and instructions together in one work area while keeping those files and instructions on the user's computer [35].

- Code Insiders now lets users control reasoning effort directly from the model picker, moving a previously settings-based control into the main interface [32, 36, 37].
- OpenAI launched Codex for Students, offering U.S. and Canadian college students $100 in Codex credits to learn by building, breaking, and fixing things [38].
- fal.ai's new MCP server lets any AI coding assistant connect to 1,000+ generative AI models, part of a broader documentation overhaul with clearer structure and navigation [39, 40].

Industry Moves

Why it matters: The industry signal was not just model launches. Labs are reorganizing around large-model execution, locking down power, and putting more capital behind robotics and long-term AI strategy [41, 42, 43].

- Tencent shut down Tencent AI Lab and folded parts of it into Hunyuan, despite the lab's earlier work on Juewu game AI, Miying medical imaging, protein folding, and drug discovery. One summary framed the move as part of a broader China shift toward fewer moonshot labs and more product-driven, model-centric execution [41].
- Energy strategy is becoming a core AI infrastructure issue. One report said Meta and OpenAI are building private gas-powered plants directly connected to data centers to bypass grid delays, while Google said it has integrated 1 GW of flexible demand into long-term utility contracts so data centers can shift or reduce demand when utilities need it [42, 44].
- Unitree reported 2025 revenue of 1.708B RMB, up 335% year over year, and profit of 600M RMB, up 674%. The company said it delivered more than 5,500 humanoid robots, plans to raise 4.2B RMB from an IPO with 85% earmarked for R&D, and is targeting production of 75,000 humanoids and 115,000 quadrupeds [43].
- Google DeepMind appointed Jas Sekhon as chief strategy officer. Demis Hassabis cited Sekhon's experience as Bridgewater's former chief scientist and head of AI, and a colleague described him as exceptionally thoughtful [45, 46].

Policy & Regulation

Why it matters: Compliance questions are increasingly about attribution, access, and authorship as AI systems become easier to embed in products and workflows [4, 47, 48].

- Kimi K2.5's license became a live compliance issue after Composer 2 launched without naming its base model. One analysis said the modified

MIT license requires products above $20M in monthly revenue to display Kimi K2.5 prominently in the UI, while Cursor later said it was following the license through Fireworks and promised better attribution in future launches [4, 5, 6, 7].

- The U.S. Copyright Office ruling in *Zarya of the Dawn* was cited as reaffirming that AI-generated images are not human-authored and therefore are not protected in the same way as the human-written story [47].
- Anthropic's control over third-party access to Claude also drew attention. opencode 1.3.0 said it stopped autoloading its Claude Max plugin after Anthropic sent lawyers, while T3 Code said users can still connect Claude if they have Claude Code CLI installed and signed in, and later said it had not heard from lawyers [48, 49, 50].

Quick Takes

Why it matters: These smaller updates show where the ecosystem is filling in: serving infrastructure, agent governance, benchmark culture, and next-wave open releases [51, 52, 53].

- vLLM v0.18.0 shipped with 445 commits from 213 contributors, adding gRPC serving, GPU-less multimodal preprocessing, GPU NGram speculative decoding, ElasticEP Milestone 2, and hardware support spanning NVIDIA FA4 MLA prefill, AMD Quark W4A8, Intel XPU, and RISC-V [51, 54, 55].
- GLM-5.1 is planned as an open-source release, with the ZAI organization's Hugging Face page highlighted ahead of launch [56, 57].
- François Chollet said ARC-AGI-3 launches next week [53].
- Grok 4.20 scored 6.0% on CritPt, about 2x DeepSeek V3.2 and nearly on par with Speciale, according to one benchmark update [58].
- Okta introduced Okta for AI Agents, positioning agents as governed non-human identities with centralized access control and a kill switch for rogue agents [52].
- Perplexity Computer now connects to Pitchbook, Statista, and CB Insights, and it also added inline document creation and editing so users can revise selected sections in place [59, 60].

Sources

1. X post by @Kimi_Moonshot
2. X post by @_weiping
3. X post by @deredleritt3r

4. X post by @aakashgupta
5. X post by @leerob
6. X post by @amanrsanger
7. X post by @eliebakouch
8. X post by @Kangwook_Lee
9. X post by @ClementDelangue
10. X post by @code_star
11. X post by @ArtificialAnlys
12. X post by @ArtificialAnlys
13. X post by @_akhaliq
14. X post by @ollama
15. X post by @kimmonismus
16. X post by @lmthang
17. X post by @lmthang
18. X post by @tonylfeng
19. X post by @lmthang
20. X post by @lmthang
21. X post by @lmthang
22. X post by @datologyai
23. X post by @tshevl
24. X post by @boyuan_chen
25. X post by @konwookim
26. X post by @percyliang
27. X post by @Meituan_LongCat
28. X post by @Aditya_Soni_8
29. X post by @gneubig
30. X post by @_philschmid
31. X post by @cognition
32. X post by @pierceboggan
33. X post by @_philschmid
34. X post by @cognition
35. X post by @claudeai
36. X post by @pierceboggan
37. X post by @pierceboggan
38. X post by @OpenAIDevs
39. X post by @fal
40. X post by @fal
41. X post by @jiqizhixin
42. X post by @DeepLearningAI
43. X post by @tphuang
44. X post by @sundarpichai
45. X post by @demishassabis
46. X post by @_sholtodouglas
47. X post by @LearnOpenCV
48. X post by @thdxr
49. X post by @theo

50. X post by @theo
51. X post by @vllm_project
52. X post by @dl_weekly
53. X post by @fchollet
54. X post by @vllm_project
55. X post by @vllm_project
56. X post by @ZixuanLi_
57. X post by @_akhaliq
58. X post by @teortaxesTex
59. X post by @AravSrinivas
60. X post by @AskPerplexity