

Moonshot Says Kimi K3.1 Pro Nearly Matches Claude, With Hacking Skills Held Back on Purpose

2026-10-24 / 00:19:35

“offensive cyber capability is, by design, below what our pipeline can produce”

— from this episode's transcript

- Lenar Kess
- Damra Vol

Moonshot released Kimi K3.1 Pro with open weights promised for November 4th, and says it deliberately shipped a version that is worse at offensive cyber than the one it trained. The same day OpenAI described agents escaping its training sandbox, plus a Brotli remote code execution bug found along the way. Who sets the ceiling on what a model can do, and how would anyone outside check it?

- Kimi K3.1 Pro launch post: 81.0 percent on xdev-bench-hard against 90.5 for a checkpoint without risk-reduction steps, and Terminal-Bench 4.0 at 65.7 percent. The cyber gap can't be audited from outside, and the weights and a license are still pending.
- OpenAI, Securing our frontier model training environment: agents asked to break containment got out in novel ways, and a Brotli unauthenticated remote code execution bug surfaced. No CVE or patch status has been published in anything we saw.
- METR and Redwood three-week assessment: an outside team with internal access found ways agents could evade monitoring or bypass sandbox restrictions.

- Chris Lattner on Mojo and MAX: the language, the graph compiler and runtime internals are now open. The license is the thing to read.
- Anthropic's MCP Tasks and Elicitations: long-running plugin jobs and mid-task questions to the user, alongside Agent Skills going to a foundation.
- OpenAI's finance talk: Codex adoption from near zero to 90 percent, and a review-time threshold that explains when agents pay off.
- Carmen Li on GPU futures: reserved GPU contracts moved from backwardation to contango, so the forward curve says compute gets pricier.

SEGMENTS

- 00:00:04 Kimi K3.1 Pro
- 00:05:28 OpenAI's sandbox escapes and the Brotli bug
- 00:08:47 Builder releases: Mojo, Claude agent tooling, Chroma
- 00:11:35 What agents actually deliver
- 00:14:58 OpenAI's round and how GPU compute trades
- 00:17:44 Quick hits: small models and an agent-security rule

Transcript

1. Lenar Kess 00:00:04

Here's a sentence I haven't seen in a model launch post before: "offensive cyber capability is, by design, below what our pipeline can produce." That's Moonshot, announcing Kimi K3.1 Pro on Friday. In plain terms, they trained a version that was better at attacking software and then shipped a worse one on purpose.

- kimi.ai
- x.com
- x.com

- [x.com](#)
- [reddit.com](#)
- [openai.com](#)
- [x.com](#)
- [x.com](#)
- [metr.org](#)
- [youtube.com](#)
- [x.com](#)
- [x.com](#)
- [x.com](#)
- [reddit.com](#)
- [anthropic.com](#)
- [x.com](#)
- [x.com](#)
- [x.com](#)
- [x.com](#)
- [x.com](#)
- [x.com](#)
- [thinkrail.ai](#)
- [youtube.com](#)
- [youtube.com](#)
- [youtube.com](#)
- [youtube.com](#)
- [cnbc.com](#)
- [x.com](#)
- [youtube.com](#)
- [youtube.com](#)
- [x.com](#)
- [x.com](#)
- [bottlecapai.com](#)
- [huggingface.co](#)
- [reddit.com](#)
- [block.xyz](#)
- [youtube.com](#)
- [youtube.com](#)

2. Damra Vol 00:00:22

So the open-weights lab is the one holding back. That flips the usual story, where the open model is the one with no brakes. How much worse did they make it?

3. Lenar Kess 00:00:32

They gave a number. On a hard exploit-development benchmark called xdev-bench-hard, K3.1 Pro scores 81.0 percent at pass-at-three. A checkpoint without their risk-reduction steps scored 90.5. The previous Kimi, K3, scored 33.3. And they say they removed cyber-related data from both mid-training and the reinforcement learning environment mix. We'll come back to that. First, here's today. Kimi leads, because it was the biggest discussion on Hacker News yesterday. After that comes OpenAI describing how agents broke out of its own training sandbox, plus the Brotli bug that turned up along the way. Then builder news: Mojo going fully open, Anthropic's agent tooling, and Chroma joining Replit. Last are what coding agents and finance agents actually deliver, the money behind OpenAI's next round, and a lineup of small model drops.

4. Damra Vol 00:01:26

Start with the model itself. What did Moonshot say it shipped?

5. Lenar Kess 00:01:30

A mixture of experts model. The announcement tweets put it at 2.8 trillion total parameters with 104 billion active. The blog post itself just says "3T-class." It has a one million token context window and takes images as input. A Highspeed variant runs two to three times faster. The headline number is Terminal-Bench 4.0: Kimi at 65.7 percent, Claude Opus 5.5 at 66.4, and GPT-6 Astra at 58.2. The weights are promised by November 4th, and the post doesn't state a license.

6. Damra Vol 00:02:04

So "open weights" is a promise for eleven days from now, and the chart is Moonshot's own. Still, seven and a half points ahead of Astra on a terminal task, and under a point behind Opus, is a strange place for an open model to be. Who's vouching for it?

7. Lenar Kess 00:02:21

Two people whose reads I'd weigh. Simon Willison says, in effect, that this is the first time he's seen an open-weights model compete this credibly with the frontier. Nathan Lambert reads the benchmarks as evidence that the lab fixed its reinforcement learning, which had been the weak spot in the earlier Kimi. Moonshot also posted its own head-to-head of K3.1, GPT-6 and Claude on a resource-constrained terminal task, which is the sort of test a builder would run.

8. Damra Vol 00:02:47

The RL point fits their own numbers, for what internal numbers are worth. Their long-horizon Kimi Code Bench goes from about 56 for K3 to 61.2 for K3.1. And the hallucination bench is the odd one. On the hard set with tools, claim-level hallucination drops from 23.4 for K3 to 8.3, which is better

than Opus at 15.0 and worse than Astra at 5.5. Lower is better there. I'd still rather see someone else run it.

9. Lenar Kess 00:03:18

There's also a spot where the gap shows up. On BenchCAD without tools, K3.1 is at 72.8 and Opus is at 73.0, basically level. With tools, it's 92.7 against 96.2 for Opus and 95.9 for Astra.

10. Damra Vol 00:03:35

That's the pattern I'd expect from a model whose raw ability got fixed before its tool-use habits did. Close on the model, behind once it's driving something. Which loops back to your opening, because the cyber ceiling is exactly the kind of thing that depends on what happens when the model drives.

11. Lenar Kess 00:03:52

Right. The post says its cyber capability trails proprietary frontier models that have been public for at least three months. And then the policy sentence: "we aim not to release open weights that would extend the absolute frontier of cyber capability." Followed by: "we intend to progressively raise the cyber capability available in open weights." The Hacker News discussion, 939 points and 559 comments, leaned heavily into that paragraph, about corporate control of what an open model is allowed to know.

12. Damra Vol 00:04:22

I read "progressively raise" as a ratchet, and Moonshot holds the handle. Nobody outside can check the gap between 81 and 90.5, because the checkpoint that proves it stays inside the lab. And on November 4th, anyone who downloads the weights can fine-tune cyber data back in. Taking it out of mid-training raises the cost of that. It doesn't lock anything. Still, it's more restraint than I expected from anyone shipping a model near the top of a terminal benchmark.

13. Lenar Kess 00:04:50

One more item, and it needs a flag. A screenshot gallery on the singularity subreddit alleges that Kimi told influencers not to disclose paid promotion and pushed specific talking points about the new model. That's an allegation from screenshots, and I can't confirm it. It doesn't touch the benchmark numbers, but it does mean some of the early enthusiasm has uncertain provenance.

14. Damra Vol 00:05:11

Which is another reason the interesting date is November 4th. That's when independent people can run Terminal-Bench against weights they hold, and when anyone has to figure out how to serve a

model this size. Until then it's a vendor chart, two respected outside readers, and a cyber ceiling no one can audit.

15. Lenar Kess 00:05:28

OpenAI published a post Friday titled "Securing our frontier model training environment." Andrew Curran quoted from it: teams of frontier-model agents, asked to break containment, got out of the virtual machine defenses in novel ways. Greg Brockman followed with his own take, that OpenAI's frontier cyber models now secure its training environment far beyond what the company could do unassisted. The methods live in the post, so I'll stay with what Curran and Brockman said.

16. Damra Vol 00:05:55

What stands out is that the attackers and the defenders are the same class of model. Agents try to get out, other models help build the walls. On Wednesday we did the story where the monitor lived in a repository the agent could write to. This is OpenAI saying it hardened the room and then sent in agents to pick the lock.

17. Lenar Kess 00:06:13

That Wednesday story was about monitoring being weakened. This one is about containment, and it's a separate artifact. There's also an outside check. METR and Redwood Research did a three-week security assessment with access to OpenAI's internal infrastructure. Their write-up, as TLDR summarizes it, identifies weaknesses that could let agents evade monitoring, bypass sandbox restrictions, or conceal what they did.

18. Damra Vol 00:06:37

Three weeks of access is what makes me trust that one more. A lab saying its sandbox is good and an outside team getting inside for three weeks are different kinds of statement, and OpenAI let the second one happen and got it published.

19. Lenar Kess 00:06:50

One piece of this reached outside OpenAI. Its account posted that during this red-teaming it found an unauthenticated remote code execution bug in Brotli, the compression library that sits in browsers and web servers. Brockman says disclosure is under way and that it shows how hard coordination is across downstream projects. No CVE number, affected versions, or patch status appears in anything I've read, so I won't say who's exposed.

20. Damra Vol 00:07:15

A compression library is a nasty place for one. It sits in front of almost everything and chews on bytes an attacker controls. And everyone who bundles it has to ship their own fix on their own schedule, which is what I take Brockman to mean by coordination.

21. Lenar Kess 00:07:29

The same week produced a pile of other bugs. Fireship's roundup covers a PHP remote code execution flaw nicknamed PHP2Shell that he says was found with a model. A patch leaked on Hacker News before release, and a public exploit video opened roughly a 24-hour unpatched window. Node.js patched a code-execution flaw triggered by malformed compressed files. GitLab fixed a regex engine bug that had affected self-hosted installs since 2014. A LiteLLM release shipped credential-stealing code before it was pulled. And an Indian certificate authority generated private keys from tiny integers instead of proper 384-bit values, until browsers revoked its certificates within three days. That's a YouTuber's summary, and nothing in it ties the Node.js flaw to Brotli.

22. Damra Vol 00:08:14

The PHP one is where I'd stay. A model finds the bug, the fix goes into a public commit, and anyone with a model can read the diff. That's my read, but it squeezes the disclosure window from both ends. Finding gets faster, and reverse-engineering the patch gets faster.

23. Lenar Kess 00:08:31

Which makes the Brotli disclosure a live test. OpenAI pointed its own cyber agents at its own stack, found something that lives outside its stack, and now has to hand it to maintainers who don't work for it. The advisory, once it has a CVE number, will say how much time they had.

24. Lenar Kess 00:08:47

Chris Lattner announced that Mojo and MAX are now fully open for development, extension, and commercial use. That covers the full Mojo language, the MAX graph compiler, and the internals of the MLRT runtime. It drew more attention than anything else I went through this morning, and so far it's one tweet. The license and the repository are what I'd read next.

25. Damra Vol 00:09:08

"Runtime internals" is the phrase I'd circle. Plenty of languages are open while the compiler and runtime that make them fast stay closed. The graph compiler is where decisions about fusing operations and targeting devices get made. If that's open, someone can add a backend for hardware Modular doesn't care about. "Commercial use" is in the tweet, but patent and trademark terms are where "fully open" tends to get complicated.

26. Lenar Kess 00:09:33

Anthropic had a busy Friday. The ClaudeDevs account announced support for MCP Tasks, which are long-running plugin jobs that report back when they finish, deep research being the example. It also announced Elicitations, where a plugin can stop mid-task and ask you for input, like confirming a refund. Overnight, a Reddit post summarized an Anthropic guide dated October 22 arguing that most agents should have fewer subagents: one main loop, with skills and tool search. The post quotes the guide as saying "we cut Claude Code's system prompt by over 80%." I've only seen that through the Reddit summary. TLDR also reports Anthropic donated the Agent Skills standard to the Agentic AI Foundation.

27. Damra Vol 00:10:16

Those all point the same way. Long jobs move into the protocol, the model asks a human when it needs one, and the swarm of subagents gets replaced by one loop that looks things up. An 80 percent cut, if accurate, probably means instructions moved out of the prompt and into skills loaded on demand. And donating Skills to a foundation means Anthropic gave up the pen on a format other companies already use. OpenAI's finance talk later today lists skills as part of its plugin stack.

28. Lenar Kess 00:10:44

Harrison Chase put a conceptual spin on it: skills should be thought of as memory, not static files, meaning something an agent writes and revises as it works.

29. Damra Vol 00:10:54

That's a different bet than a foundation standard, though. A standard wants skills to be portable files. Chase wants them to be alive. A format can serve both, but the two pull it in different directions.

30. Lenar Kess 00:11:05

And one acquisition-style item. Chroma, the vector database company, announced it is joining Replit to build what they call the context layer for the self-driving company. Replit's Amjad Masad said the Chroma team will help give agents the context they need.

31. Damra Vol 00:11:20

What I'd want to know is what happens to Chroma as an open-source project. The announcement carries no deal terms or open-source commitments. A retrieval library inside an agent platform is a different animal from a retrieval library that anyone can embed.

32. Lenar Kess 00:11:35

Tzvetan Mikov posted that an agent produced 66 commits for a feature that should have been two or three. Next to that is a Google Cloud AI Research paper, which I'm only seeing through DAIR.AI's summary. It reports that the best agent configuration averages 50.6 percent on professional engineering deliverables. That run used 27 million tokens, on a new benchmark called PLE.

33. Damra Vol 00:11:58

Half, at 27 million tokens. Alongside 66 commits, that's one complaint from two directions: lots of motion, not much convergence.

34. Lenar Kess 00:12:08

Addy Osmani's advice is to give agents a way to verify their own work, with video verification and end-to-end tests, and he cites Warp's approach. JetBrains also launched ThinkRail, a tool built around a graph of specifications. It had 13 points on Hacker News and one comment, which disputes that the agent actually reads the graph. That's one reader's view.

35. Damra Vol 00:12:29

Osmani's advice reads like the repair for Mikov's 66 commits. If the agent can't see the feature working, it keeps committing until something looks done.

36. Lenar Kess 00:12:38

Now the contrast. Joseph Kim, who leads ChatGPT for financial services at OpenAI, with Morgan Stanley and Evercore as design partners, said Codex adoption in finance went from near zero to 90 percent by June. His explanation is that the February desktop app drove it, not new model weights. That's self-reported. He also described a threshold: usefulness jumps once the time you spend reviewing the output drops below the time it saved you.

37. Damra Vol 00:13:05

That threshold idea explains the 66 commits better than anything else in the pile. Review overhead above the time saved means the agent is a net cost, however good it is. And it explains why the same model can look useless in one workflow and indispensable in another.

38. Lenar Kess 00:13:21

Arca is an example of crossing it. Its head of AI described agents that own multi-day workflows end to end. One case was a client's \$100,000 tax liability. An advisor analyzes the situation and assets get sold. Then the workflow waits out settlement delays, and an ACH transfer closes it. A signal-driven harness feeds the agents emails, market moves, and custodian events, and agents subscribe to milestones. When a human completes a task or changes a calendar entry, that generates a signal

that wakes the relevant agent. He says agents pre-process about 90 percent of tasks and ask a human at decision points, like routing a liquidation to a trader.

39. Damra Vol 00:14:00

Waking on events rather than polling is how a good human assistant works. You don't check on the settlement every hour. You get tapped when it clears. I'd like to see how they handle two signals that contradict each other.

40. Lenar Kess 00:14:12

Contour, a hedge fund with about \$6 billion under management, measures this differently. Its head of AI said profit and loss takes months or years to show up, so they run live A/B tests on analyst workflows across three-month earnings cycles and track what they call cost per useful insight. And they count the engineering overhead: downtime from broken model calls or connectors, and the hours to fix them. Separately, an OpenAI customer clip from Shield, an IT-services company, says Astra is 35 to 40 percent faster on its tickets. That one is a testimonial.

41. Damra Vol 00:14:46

Counting your own connector downtime in the return calculation is rare. Most ROI slides assume the connectors work. A fund that bills the broken connector against the agent will know whether it's winning.

42. Lenar Kess 00:14:58

CNBC reports, via TLDR, that OpenAI is nearing a round of up to \$30 billion at a \$1.25 trillion valuation, led by investors from the United Arab Emirates. It's described as a bridge ahead of a possible IPO, with signatures targeted for the end of next week, and the final size is still being set. It isn't signed.

43. Damra Vol 00:15:19

A bridge round at over a trillion dollars is a strange phrase. A bridge implies the destination is the IPO and the money is for getting there. Either they want cash before the window opens, or they expect the public price to be higher. We covered Anthropic's draft prospectus numbers on Wednesday, so both big labs are now showing their finances.

44. Lenar Kess 00:15:38

Separately, Jeff Davies tweeted a structure in which Nvidia puts \$250 billion of equity into a special purpose vehicle that syndicates \$750 billion of debt for power and neocloud buildout. That's one person's hypothetical, not a reported deal.

45. Damra Vol 00:15:54

Even as a sketch, look at who owns the equity: the chip vendor, and the debt buys the chips. That loop has been described before, and it's still a loop.

46. Lenar Kess 00:16:03

Two AI Engineer talks show how GPU capacity changes hands. Afaan from Jump Trading calls the market an unlit dark pool, where big deals get done in Slack and Telegram. The contracts are built from master service agreements, service-level agreements, and order forms. Startups typically face 15 to 25 percent upfront prepayment on multi-year contracts, which ties compute purchases to fundraising cycles. Carmen Li runs Compute Exchange, a marketplace that doesn't own hardware, and Silicon Data, whose indices CME uses to settle futures. She says reserved GPU contracts moved from backwardation to contango, and that B200s occasionally trade above list price.

47. Damra Vol 00:16:44

Contango means the futures price sits above the spot price, so the market is saying GPU time will cost more over the next three years. Everyone keeps being told tokens get cheaper, and the forward curve says the hardware underneath gets more expensive. The Jump speaker's answer is Jevons: cheaper tasks mean more tasks. Li also says Nvidia cited Silicon Data's appraisal method in a September investor deck, which is her claim.

48. Lenar Kess 00:17:10

A short one on chips. Nvidia stated that court cases found less than one half of one percent of export-controlled chips were diverted to China. Chris McGuire called that "beyond absurd." The Financial Times, via Trevor Noren, reports China is building data centers fast across its energy-rich interior while chips stay scarce. I haven't seen Nvidia's full statement, so that's the claim as McGuire quotes it.

49. Damra Vol 00:17:34

Court cases only count diversions that got prosecuted. So that number is a floor on what happened, and it can't be a rate. That's presumably McGuire's whole objection.

50. Lenar Kess 00:17:44

A few small drops. BottleCap reports a version of Gemma-4-26B-A4B that uses 39 to 69 percent less thinking at the same accuracy, which is the company's own claim. NVIDIA released Nemotron-3.5-Bolt-E4B-4GB, a small model post-trained from Gemma-4-E4B-it for low-memory use. A LocalLLaMA user spotted a vLLM pull request that appears to reference Qwen4-27B, which is a

sighting rather than an announcement. And Jack Dorsey's Block published a voice model called Dessa.

51. Damra Vol 00:18:17

The Gemma number is the one with teeth. If thinking tokens fall by a third to two thirds at equal accuracy, that cuts latency and cost for anyone running a reasoning model. I'd wait for someone besides the company to reproduce it.

52. Lenar Kess 00:18:31

One security item. The Primetime reports that OpenCode suffered a distillation attack affecting 1.4 percent of its console users, with attackers using stolen API tokens to harvest model outputs. His figures on defense costs, \$1,000 to \$3,000 a day, come from a commentary video, and I haven't confirmed them against Anthropic's report. The piece I'd keep was a Bloomberg talk on agent security. Its rule of three says that when an operation combines untrusted input, sensitive data, and an external action, a human has to sign off.

53. Damra Vol 00:19:04

That rule works because you can check it in code. The three conditions are properties of a tool call, so a broker outside the model can enforce it without asking the model's opinion. After a day of agents breaking out of sandboxes, I like rules that don't depend on the agent cooperating.

54. Lenar Kess 00:19:20

November 4th is when Kimi's weights either ship with a license we can read or don't, and when independent Terminal-Bench runs show whether 65.7 percent holds up. This has been Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic