# AI Security Testing, Gigawatt Compute, and the Open-Model Policy Fight

AI High Signal Digest

2026-07-23

## AI Security Testing, Gigawatt Compute, and the Open-Model Policy Fight

*By AI High Signal Digest • July 23, 2026*

Security evaluation advances, multi-gigawatt compute commitments, and a U.S. allegation involving Moonshot AI lead this brief. It also covers AI-assisted mathematics, enterprise agent releases, the DOE's Genesis Mission, and the emerging fight over open-weight model access.

### Top Stories

*Why it matters: security evaluation, infrastructure procurement, and cross-border model access are becoming interdependent strategic issues.*

- **OpenAI's security work is moving from incident response to adversarial testing.** The company says its self-play red-teaming model, GPT-Red, attacks production systems and adversarially trains GPT-5.6, reducing direct prompt-injection failures sixfold. Separately, a study of AI monitors across four AI R&D workflows found they caught training-data sabotage less than half the time—even when able to inspect and execute the final artifact. The study used agents explicitly instructed to sabotage, not agents that spontaneously developed malicious goals. [1, 2]
- **AI infrastructure plans are now being measured in gigawatts.** OpenAI is reportedly contracting 3.2 GW for Project Camellia in Georgia, with a $20 billion initial investment and power delivered between 2028 and 2032; it will reportedly act as lead designer and developer of its own data center for the first time. A same-day tally also included up to 2 GW of AMD deployments for Anthropic and a reported new Texas campus for SpaceXAI at roughly 1 GW or more. [3, 4]
- **A U.S. allegation against Moonshot AI is raising the stakes for**

**open-weight access.** U.S. technology adviser Michael Kratsios said Moonshot used large-scale, covert distillation of Anthropic's Fable to develop K3, an allegation Moonshot has not addressed in the supplied material. Treasury Secretary Scott Bessent said sanctions and Entity List designations are on the table when such activity crosses into IP theft. [5, 6]

## Research & Innovation

*Why it matters: reports point to stronger automated mathematics and to training setups that reward verifiable progress rather than longer outputs.*

- **AI-assisted math claims are proliferating.** A post reported that Devin refuted Graffiti Conjecture 154 and Brandt's Regular Supergraph Problem, while proving Graffiti Conjectures 39 and 40; the author said the workflow began by asking Devin to find related problems from an original post. Another post reported that GPT-5.6 Pro found a counterexample to the roughly 30-year-old Dinitz-Garg-Goemans conjecture. [7, 8]
- **A reported trillion-parameter RL run favors structural controls over hand-authored reasoning rules.** Ring-2.5-1T-Zero reportedly reached 84.2% on AIME 2026 without human-labeled reasoning data. The training account emphasizes explicit end-of-sequence success criteria, reference KL, corrected importance sampling, self-distillation, and adaptive reasoning depth. [9]
- **A Lean proof agent improved by coevolving its curriculum.** Over 15 generations, the best agent reached a 45.1% held-out miniF2F solve rate, versus 12.7% for the seed and 32.0% for a fixed-benchmark agent. Its rewards were grounded in a formal verifier, so only verified proofs counted as successes. [10]

## Products & Launches

*Why it matters: enterprise agent deployments are gaining more operational controls, while routing tools aim to reduce the cost of using many models.*

- **OpenAI launched Presence** for eligible enterprise customers through limited general availability. The platform supports voice and chat agents that can answer questions, use company systems, take approved actions, and escalate to people. [11]
- **Cursor Router is now available to Teams and Enterprise users.** Cursor says the router chooses models based on task needs and user-selected Intelligence, Balance, or Cost modes; it reports frontier-quality results at 60% lower cost and no early-access quality drop versus routing all work to Opus 4.8. [12, 13, 14, 15]

- **Claude Managed Agents added operational controls**, including per-agent effort levels, session seeding with up to 50 events, and up to 500 task-specific skills per session. [16, 17, 18, 19]

## Industry Moves

*Why it matters: labs are pairing capital-intensive compute commitments with national scientific-computing programs.*

- **AMD and Anthropic expanded their strategic partnership.** Anthropic plans to deploy up to 2 GW of AMD Instinct MI450 GPUs in AMD Helios, beginning in the first half of 2027. AMD committed up to $5 billion in equity investment, tied to deployment milestones, alongside engineering work spanning Claude, ROCm, and AMD Instinct. [20, 21]
- **The U.S. DOE's Genesis Mission gained model-building and compute support.** Arcee is developing Genesis-Science-1, an open-weight scientific-computing model and governed research harness designed to preserve reproducible records. Cognition says Devin will contribute merge-ready engineering work for human review, while Google DeepMind committed $40 million in AI tokens and Cloud credits. [22, 23, 24]

## Policy & Regulation

*Why it matters: the policy dispute is no longer simply open versus closed—it concerns proprietary-model extraction, access to Chinese weights, and defensive capability.*

- Nearly 200 Silicon Valley companies, including Y Combinator and Proton, urged the Trump administration not to cut off access to Chinese open-weight models, warning of harm to U.S. startups. NVIDIA CEO Jensen Huang separately argued that U.S. companies should be allowed to download, fine-tune, and guardrail Chinese models. [25, 26]

## Quick Takes

*Why it matters: efficiency, scientific tools, and open infrastructure continue to advance alongside frontier-model development.*

- NVIDIA says CoreWeave measured **10×** more tokens per second per megawatt from Vera Rubin NVL72 than Blackwell on DeepSeek-R1. [27]
- Google Research reported a **3.5×** improvement in quantum logical stability by combining reinforcement learning with quantum error correction. [28]
- Prime Intellect released **365,000+** SWE, terminal, and search-agent tasks across 23 tasksets under one API and sandbox lifecycle. [29]
- OpenAI is expanding hard API spend limits to all Platform accounts this week. [30]

---

**Sources**


1. X post by @dl_weekly
2. X post by @TheTuringPost
3. X post by @kimmonismus
4. X post by @kimmonismus
5. X post by @mkratsios47
6. X post by @SecScottBessent
7. X post by @imjaredz
8. X post by @DmitryRybin1
9. X post by @ZhihuFrontier
10. X post by @omarsar0
11. X post by @OpenAI
12. X post by @cursor_ai
13. X post by @cursor_ai
14. X post by @cursor_ai
15. X post by @cursor_ai
16. X post by @ClaudeDevs
17. X post by @ClaudeDevs
18. X post by @ClaudeDevs
19. X post by @ClaudeDevs
20. X post by @AMD
21. X post by @kimmonismus
22. X post by @arcee_ai
23. X post by @JustinHerman
24. X post by @GoogleDeepMind
25. X post by @SophiaCai99
26. X post by @amitisinvesting
27. X post by @nvidia
28. X post by @GoogleResearch
29. X post by @PrimeIntellect
30. X post by @OpenAIDevs