Anthropic flags large-scale model distillation as OpenAI sunsets a key coding benchmark

AI News Digest

2026-02-24

Anthropic flags large-scale model distillation as OpenAI sunsets a key coding benchmark

By AI News Digest • February 24, 2026

A major security story leads: Anthropic alleges industrial-scale distillation attacks against Claude. Also: OpenAI sunsets SWE-Bench Verified amid contamination and test-quality issues, multiple labs ship new models and real-time voice updates, and prominent researchers sharpen the safety/governance debate—backed by new benchmarks and crisis simulations.

Security & model protection: Anthropic alleges large-scale distillation of Claude

Anthropic: "industrial-scale distillation attacks" tied to DeepSeek, Moonshot AI, and MiniMax

Anthropic says it identified distillation attacks on its models by **DeepSeek**, **Moonshot AI**, and **MiniMax**, involving **24,000+ fraudulent accounts** and **16M+ exchanges** with Claude used to "extract its capabilities" for training other models [1]. Anthropic also emphasized that distillation can be legitimate (e.g., making smaller/cheaper models), but warned that **illicit distillation** can remove safeguards and feed capabilities into military, intelligence, and surveillance systems [2].

Why it matters: This is a concrete, quantified claim of large-scale capability extraction—and a signal that model access controls are becoming a first-order competitive and national-security issue [3][4].

[1] post by @AnthropicAI
[2] post by @AnthropicAI
[3] post by @AnthropicAI
[4] post by @AnthropicAI

> "more industrial strength thieves complaining about having been ripped off " [5]

Anthropic says attacks are "growing in intensity and sophistication" and calls for "rapid, coordinated action" across industry, policymakers, and the broader AI community [6]. More details: https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks [7].

Benchmarks reset: OpenAI sunsets SWE-Bench Verified

OpenAI: stop reporting SWE-Bench Verified; recommend SWE-Bench Pro

OpenAI says it will no longer report **SWE-Bench Verified**, recommending **SWE-Bench Pro** instead, citing benchmark **saturation** and evidence of **contamination** from public repositories [8]. In OpenAI-linked discussion, "every single frontier model" is described as able to regurgitate evaluation data and solutions—sometimes from the **Task ID alone** [9][10].

Why it matters: SWE-Bench Verified has functioned as a "north star" coding benchmark; OpenAI's deprecation is a public admission that headline coding-eval progress can become misleading once contamination and test issues dominate [11][12].

What OpenAI says went wrong (two separate failure modes)

- **Bad / unfair tests:** In a review, OpenAI's team describes many cases where tests expected unspecified implementation details (e.g., exact naming) or even additional features not in the problem description [13]. In a separate summary, OpenAI's deeper analysis is described as finding **>60%** of remaining problems unsolvable, including **49** tests "too narrowly defined" and **26** tests "too wide" (requiring unspecified features) [14].
- **Training-on-test contamination:** SWE-Bench tasks draw from popular open-source repos (no "canary strings"), which makes leakage hard

[5] post by @GaryMarcus
[6] post by @AnthropicAI
[7] post by @AnthropicAI
[8] post by @OpenAIDevs
[9] post by @latentspacepod
[10] post by @swyx
[11] SWE-Bench Verified is Contaminated: What Comes Next — with OpenAI Frontier Evals team
[12] post by @OpenAIDevs
[13] SWE-Bench Verified is Contaminated: What Comes Next — with OpenAI Frontier Evals team
[14] The End of SWE-Bench Verified — Mia Glaese & Olivia Watkins, OpenAI Frontier Evals & Human Data

to prevent; OpenAI describes examples where a model's reasoning referenced repository specifics needed to pass a test that was otherwise "pretty impossible" [15].

Where evals are headed next (per OpenAI Frontier Evals discussions)

OpenAI-associated commentary points to future coding evals that better capture **long-horizon work**, **open-ended design decisions**, **code quality/maintainability**, **end-to-end product building**, and **real-world usage metrics** [16][17].

Model and product updates (signals from major labs)

Google: Gemini 3.1 Pro announced

Google announced **Gemini 3.1 Pro**, positioned to power consumer apps like **Gemini** and **NotebookLM** plus enterprise products, and claimed "more than double the reasoning performance" versus the prior Gemini model [18]. Google also gave examples of advanced reasoning tasks (e.g., code-based animations and more advanced web design), and named early enterprise testers including **JetBrains**, **Databricks**, **Cartwheel**, and **Hostinger Horizons** [19][20].

Why it matters: The pitch explicitly bundles *reasoning gains* with *agentic workflow* positioning (data synthesis, long context, multi-step tools), which is increasingly how frontier model launches are being framed [21][22].

Anthropic: Claude Sonnet 4.6 + 1M-token context (beta)

Anthropic debuted **Claude Sonnet 4.6**, saying it reaches capabilities similar to its larger **Opus 4.6** at lower pricing, and introduced a **1M token context window** in beta—positioned as large enough for full "codebases, lengthy contracts or dozens of research papers" in one request [23]. Claimed upgrades include improved agentic tools, "computer use" (operating software "like a human would"), and stronger long-context reasoning for business tasks like financial and document analysis [24].

[15] SWE-Bench Verified is Contaminated: What Comes Next — with OpenAI Frontier Evals team
[16] SWE-Bench Verified is Contaminated: What Comes Next — with OpenAI Frontier Evals team
[17] SWE-Bench Verified is Contaminated: What Comes Next — with OpenAI Frontier Evals team
[18] NVIDIA's Big Week, New AI Models, Social Media On Trial
[19] NVIDIA's Big Week, New AI Models, Social Media On Trial
[20] NVIDIA's Big Week, New AI Models, Social Media On Trial
[21] NVIDIA's Big Week, New AI Models, Social Media On Trial
[22] NVIDIA's Big Week, New AI Models, Social Media On Trial
[23] NVIDIA's Big Week, New AI Models, Social Media On Trial
[24] NVIDIA's Big Week, New AI Models, Social Media On Trial

Why it matters: The combination of long context + agentic "computer use" continues the trend toward assistants that can act across tools and documents, not just chat [25][26].

OpenAI: gpt-realtime-1.5 (realtime API)

OpenAI released **gpt-realtime-1.5**, described as improving "intelligence, instruction following, and voice quality," with a public demo link and a phone number to try it [27]. Greg Brockman also pointed to an "Improved realtime API" announcement [28].

Why it matters: Realtime voice quality and instruction following are key friction points for voice-first agents; shipping a new realtime model suggests continued iteration toward production-grade conversational interfaces [29].

OpenAI Codex: ex-Cursor hire to pursue an "Agent Development Environment"

Rohan Varma (ex-Cursor) said he's joining **OpenAI Codex** to build the "future of agentic development," arguing the next step isn't "a better IDE" but an **Agent Development Environment (ADE)** for orchestrating agents and reasoning over their outputs [30]. In the same thread, he points to Codex shipping models for agentic coding (e.g., **gpt-5.3-codex**) and to the new "Codex App" as a glimpse of direction [31].

Why it matters: This is an explicit strategic framing: developer tooling as orchestration and supervision of multiple agents, not just code completion [32].

Safety, governance, and "what happens next" (research + expert signals)

Bengio: hardware controls and international safety guidelines (nuclear analogy)

Yoshua Bengio warned that human-level AI could arrive "in a few years" to "20 years," citing how systems like ChatGPT surprised researchers [33]. He argued for international safety guidelines (drawing parallels to post–WWII nuclear gov-

[25] NVIDIA's Big Week, New AI Models, Social Media On Trial
[26] NVIDIA's Big Week, New AI Models, Social Media On Trial
[27] post by @juberti
[28] post by @gdb
[29] post by @juberti
[30] post by @rohanvarma
[31] post by @rohanvarma
[32] post by @rohanvarma
[33] Quebec AI expert warns that the technology could be disastrous without regulation

ernance) and advocated controls around AI hardware—describing GPUs as a bottleneck that could be registered and licensed with guardrails [34].

Quebec AI expert warns that the technology could be disastrous without regulation (3:30)

Hinton: most experts expect superintelligence within ~20 years; calls for cross-country alignment research sharing

Geoffrey Hinton said "most neural net experts believe" superintelligent AI will arrive within ~20 years (offering his own range of "very likely" more than five years and possibly up to 20) and noted other prominent timelines he's heard (e.g., Demis Hassabis ~10 years; Ilya Sutskever sooner than 10; Dario Amodei 3 years) [35]. He also suggested creating research institutes in different countries that test how to make national "super smart AI" systems "care more about people than about itself," sharing alignment techniques internationally while not sharing capability-advancing methods [36].

[34] Quebec AI expert warns that the technology could be disastrous without regulation
[35] Ep. 2 - Five Decades of Neural Networks with Geoffrey Hinton
[36] Ep. 2 - Five Decades of Neural Networks with Geoffrey Hinton

Simulated nuclear crises: LLMs escalated and never chose de-escalation options

A King's College London researcher ran simulated nuclear crisis games across **21 matches** (over 300 turns) with **GPT-5.2**, **Claude Sonnet 4**, and **Gemini 3 Flash**, finding models used nuclear weapons "more often and earlier than humans," and that **no model selected any de-escalatory option** in the action distribution [37]. The paper reported **95%** of games reached tactical nuclear use and **76%** reached strategic nuclear threats [38].

Why it matters: If AI systems become routine "advisors" in high-stakes settings, this suggests model choice could materially change crisis dynamics—and that evaluation needs to capture strategic behavior, not just static QA [39][40].

Measuring progress (beyond product demos): new benchmarks and attribution tools

"Concept Influence": training-data attribution via interpretable vectors

A new approach called **Concept Influence** proposes attributing model behavior to interpretable vectors (e.g., probes, SAE features) rather than to individual examples, described as **20× faster** than influence functions and "more semantically meaningful" [41][42][43]. Reported results include outperforming influence functions on emergent misalignment attribution, and a finding on OASST1 that using only **5%** of the data maintained full capability while reducing harm **3×** [44][45].

Why it matters: If robust, this is a practical bridge between interpretability artifacts (probes/SAEs) and governance-oriented questions like "which data causes this behavior?" [46].

LABBench2: 1,900-task benchmark for AI systems doing biology research work

Researchers from Edison Scientific, UC Berkeley, FutureHouse, and the Broad Institute released **LABBench2** (1,900 tasks) spanning literature retrieval, pro-

[37] Import AI 446: Nuclear LLMs; China's big AI benchmark; measurement and AI policy
[38] Import AI 446: Nuclear LLMs; China's big AI benchmark; measurement and AI policy
[39] Import AI 446: Nuclear LLMs; China's big AI benchmark; measurement and AI policy
[40] Import AI 446: Nuclear LLMs; China's big AI benchmark; measurement and AI policy
[41] r/MachineLearning post by u/KellinPelrine
[42] r/MachineLearning post by u/KellinPelrine
[43] r/MachineLearning post by u/KellinPelrine
[44] r/MachineLearning post by u/KellinPelrine
[45] r/MachineLearning post by u/KellinPelrine
[46] r/MachineLearning post by u/KellinPelrine

tocol troubleshooting, molecular biology assistance, and experiment planning [47]. The benchmark highlights weaknesses like poor cross-referencing across biological databases and difficulty interpreting figures/tables, while noting that tool access can improve performance [48].

Why it matters: It's an example of evals shifting toward "can the system do real scientific work," not just answer exam-style questions—while explicitly showing where current frontier models still break down [49].

Workforce and ecosystem signals

Software jobs vs. "AI kills coding jobs" narratives

François Chollet highlighted a data point that **software development jobs grew 10% over the last year** while the overall market declined **5.8%** [50]. In related posts, he argued that if AI makes software engineers more productive, **demand can rise** (invoking Jevons paradox), and suggested latent demand for software is "orders of magnitude" larger than what's deployed today [51][52].

Andrew Ng: "X Engineer" roles and more software builders

Andrew Ng argued that even if developers become "10x more productive," demand for custom software "has no practical ceiling," and predicted growth in "X Engineer" jobs (e.g., Recruiting Engineer, Marketing Engineer) embedded in business functions to create software for that function [53].

Hugging Face: competition framing

Clement Delangue warned that the AI ecosystem needs "more competition and innovation spreading," arguing that otherwise "a few companies" could control the world in a "very scary" way [54].

Sources

1. post by @AnthropicAI
2. post by @AnthropicAI
3. post by @GaryMarcus

[47] Import AI 446: Nuclear LLMs; China's big AI benchmark; measurement and AI policy
[48] Import AI 446: Nuclear LLMs; China's big AI benchmark; measurement and AI policy
[49] Import AI 446: Nuclear LLMs; China's big AI benchmark; measurement and AI policy
[50] post by @perborgen
[51] post by @fchollet
[52] post by @fchollet
[53] post by @AndrewYNg
[54] post by @ClementDelangue

4. post by @AnthropicAI
5. post by @OpenAIDevs
6. post by @latentspacepod
7. post by @swyx
8. SWE-Bench Verified is Contaminated: What Comes Next — with OpenAI Frontier Evals team
9. The End of SWE-Bench Verified — Mia Glaese & Olivia Watkins, OpenAI Frontier Evals & Human Data
10. NVIDIA's Big Week, New AI Models, Social Media On Trial
11. post by @juberti
12. post by @gdb
13. post by @rohanvarma
14. Quebec AI expert warns that the technology could be disastrous without regulation
15. Ep. 2 - Five Decades of Neural Networks with Geoffrey Hinton
16. Import AI 446: Nuclear LLMs; China's big AI benchmark; measurement and AI policy
17. r/MachineLearning post by u/KellinPelrine
18. post by @perborgen
19. post by @fchollet
20. post by @fchollet
21. post by @AndrewYNg
22. post by @ClementDelangue