Claude extraction allegations, video-trained action agents, and faster real-time AI APIs

AI High Signal Digest

2026-02-24

Claude extraction allegations, video-trained action agents, and faster real-time AI APIs

By AI High Signal Digest • February 24, 2026

Anthropic says it uncovered industrial-scale Claude extraction campaigns, while new agent capabilities arrive via video-trained computer-action models (FDM-1) and OpenAI's lower-latency WebSockets + upgraded real-time voice. Also: a notable interpretability release (Steerling-8B), DeepSeek V4 signals amid export-ban reporting, and a broad set of product and enterprise moves.

Top Stories

1) Anthropic reports industrial-scale "distillation" campaigns targeting Claude

Why it matters: If model capabilities can be reconstructed through massive API querying, it changes the security perimeter from *compute access* to *API abuse detection*—with implications for safeguards, export controls, and the economics of frontier development.

Anthropic says it identified industrial-scale campaigns by **DeepSeek, Moonshot AI, and MiniMax** that used ~**24,000 fraudulent accounts** to generate **16M+ exchanges** with Claude, extracting capabilities to train or improve other models [1][2]. One breakdown circulating in the discussion lists query volume as ~**150k (DeepSeek), 3.4M (Moonshot), and 13M (MiniMax)** [3].

A widely shared explanation frames distillation as copying a stronger model's behavior by collecting millions of input/output examples—common internally for "smaller, cheaper models," but here characterized as an espionage-like operation

[1] post by @AnthropicAI
[2] post by @niubi
[3] post by @Bayesian0_0

via API calls [4][5]. Specific tactics cited include requesting step-by-step reasoning, targeting agent capabilities across many accounts, and rapidly pivoting to new model releases [6].

Several posts emphasized downstream risks: safety filters can be lost, and "frontier AI without safeguards" could be used in military/surveillance contexts [7][8][9]. Others argue this could undermine chip export controls if output-copying scales [10][11]. A recurring recommendation: **shared detection systems across frontier labs**, since attackers can rotate to the weakest defenses [12].

Notably, reaction is mixed: one commenter argues "distillation is not an attack" [13], another claims DeepSeek's use was "qualitatively different" than MiniMax's and that lumping them together may be unfair [14].

2) Standard Intelligence's FDM-1: computer action learning from raw internet video

Why it matters: If agents can learn UI actions directly from video at internet scale, the bottleneck shifts from human labeling to compute and data access—enabling longer-horizon, higher-precision computer use.

Standard Intelligence announced **FDM-1**, a foundation model for computer actions that learns from **video** (not screenshot datasets with human action labels) [15][16]. The reported approach uses:

- An **inverse dynamics model** to infer the action between frames [17]
- A **video encoder** that compresses nearly **2 hours** of high-res footage into the space other models use for **1 minute** [18]
- Auto-labeling to reach **11M hours** of screen recordings after training on **40k hours** of labeled data (described as **550,000×** larger than the biggest open dataset) [19]

Demonstrations include constructing a gear in Blender, finding software bugs, and driving a real car through San Francisco using arrow keys [20]. One claim

[4] post by @LiorOnAI
[5] post by @AnthropicAI
[6] post by @LiorOnAI
[7] post by @LiorOnAI
[8] post by @LiorOnAI
[9] post by @AnthropicAI
[10] post by @LiorOnAI
[11] post by @LiorOnAI
[12] post by @LiorOnAI
[13] post by @glennko
[14] post by @andersonbcdefg
[15] post by @si_pbc
[16] post by @LiorOnAI
[17] post by @LiorOnAI
[18] post by @LiorOnAI
[19] post by @LiorOnAI
[20] post by @si_pbc

highlights the driving result after **<1 hour** of training footage [21]. A separate framing: this takes computer action learning from "data-constrained" to "compute-constrained" [22].

3) OpenAI ships WebSockets for agents + upgrades real-time voice with gpt-realtime-1.5

Why it matters: Latency and tool-call overhead are becoming core constraints for agent products; reducing round trips can translate directly into faster, more reliable agentic workflows.

OpenAI introduced **WebSockets in the Responses API** for low-latency, long-running agents with heavy tool calling [23]. The mode keeps a persistent connection and sends only incremental inputs rather than resending full context each turn [24][25]. OpenAI says maintaining in-memory state can speed up runs with **20+ tool calls by 20%–40%** [26].

Early third-party results highlighted in posts:

- **Cline** reported ~**15% faster** on simple tasks and ~**39% faster** on complex multi-file workflows (best cases **50%**) vs the standard API, noting a small handshake overhead that amortizes on heavy tool use [27][28].
- **Cursor** said OpenAI models are now up to **30% faster** after upgrading users to WebSockets [29].

On voice, OpenAI released **gpt-realtime-1.5** in the Realtime API with improved instruction following, tool calling, and multilingual accuracy [30]. OpenAI also reported internal eval lifts: **+5%** on Big Bench Audio, **+10.23%** on alphanumeric transcription, and **+7%** on instruction following [31]. Partners shared deployment signals, including Genspark's phone-call alpha test reporting a **66% human connection rate** (up from 43.7%) and a reduced "problem case rate" (2.1% vs 4.2%) [32].

4) Guide Labs releases Steerling-8B, positioning interpretability as a first-class model feature

Why it matters: Claims of built-in traceability and memorization suppression—if they hold up in practice—target two recurring barriers to high-stakes deploy-

[21] post by @LiorOnAI
[22] post by @Hangsiin
[23] post by @OpenAIDevs
[24] post by @OpenAIDevs
[25] post by @cline
[26] post by @OpenAIDevs
[27] post by @cline
[28] post by @cline
[29] post by @leerob
[30] post by @OpenAIDevs
[31] post by @OpenAIDevs
[32] post by @genspark_ai

ment: understanding *why* outputs occur and controlling training-data leakage.

Guide Labs announced **Steerling-8B**, described as the first and largest "large-scale inherently interpretable" language model [33]. It is presented as tracing each generated token back to **input context, training data, and human-understandable concepts** [34]. The team also claims it can **self-monitor memorized content and suppress it at inference time without retraining** [35].

Release links: Guide Labs post, GitHub, and Hugging Face [36][37][38].

5) DeepSeek V4 signals intensify alongside claims of training on NVIDIA Blackwell despite U.S. export restrictions

Why it matters: Reports of cutting-edge GPU access despite bans, combined with ongoing model-copying allegations, sharpen the question of what policy levers (compute vs access vs outputs) can realistically constrain capability diffusion.

Reuters reporting (as shared in posts) cites a senior U.S. official saying DeepSeek's new model—described as **imminent**—was trained using **NVIDIA Blackwell GPUs** despite the export ban [39][40]. Separately, posts suggested DeepSeek V4's release is imminent, with one claim pointing to a pre-release polish pattern (merging PRs) [41].

Research & Innovation

Monitoring reasoning traces: when chain-of-thought (CoT) helps—and when it doesn't

Why it matters: As CoT is used for oversight, the question isn't just "does the model show its work," but whether monitors can reliably extract the right signals.

A paper summary shared by DAIR.AI formalizes CoT "monitorability" using information theory: mutual information between CoT and output is **necessary but not sufficient** for effective monitoring [42]. It identifies two failure modes—**information gap** and **elicitation error**—and proposes two training approaches (oracle-based rewards for transparency, and a label-free conditional

[33] post by @guidelabsai
[34] post by @guidelabsai
[35] post by @guidelabsai
[36] post by @guidelabsai
[37] post by @guidelabsai
[38] post by @guidelabsai
[39] post by @AndrewCurran_
[40] post by @niubi
[41] post by @intheworldofai
[42] post by @dair_ai

mutual information objective) that improve monitor performance without degrading reasoning traces [43][44]. Paper link: https://arxiv.org/abs/2602.18297 [45].

Synthetic reasoning structure (ByteDance): "semantic isomers" and Mole-Syn

Why it matters: Long CoT training can become unstable if the model learns incompatible "reasoning structures," even when surface-level solutions look similar.

ByteDance research is summarized as treating strong long CoT as having a **molecular-like internal structure** with three behaviors: deep reasoning, self-reflection, and self-exploration [46][47][48][49]. The summary warns that simply copying reasoning traces can fail—mixing traces from different models can destabilize training due to incompatible structures ("semantic isomers") [50]. A proposed method, **Mole-Syn**, extracts transition patterns (deep reasoning → reflection → exploration) and generates new structured synthetic data without verbatim copying [51]. Paper link: https://arxiv.org/abs/2601.06002 [52].

Speech NER under real-world diversity: SF Streets benchmark + a small-sample fix

Why it matters: Navigation and emergency dispatch failures can be driven by named-entity transcription errors, especially across diverse linguistic backgrounds.

Together Research introduced **SF Streets**, a benchmark for named entity recognition in speech across **15 models** [53]. Reported metrics include **39%** average error rate on street names, **18%** lower accuracy for non-English speakers, and mis-transcriptions landing you **2.4 miles** off target [54]. A proposed fix—cross-lingual style transfer with **<1,000** synthetic samples—yielded a **60% relative improvement** on Whisper-Large [55]. SF Streets and US Streets datasets are said to be releasing publicly [56].

[43] post by @dair_ai
[44] post by @dair_ai
[45] post by @dair_ai
[46] post by @TheTuringPost
[47] post by @TheTuringPost
[48] post by @TheTuringPost
[49] post by @TheTuringPost
[50] post by @TheTuringPost
[51] post by @TheTuringPost
[52] post by @TheTuringPost
[53] post by @togethercompute
[54] post by @togethercompute
[55] post by @togethercompute
[56] post by @togethercompute

Evaluations: OpenAI stops reporting SWE-bench Verified

Why it matters: If a benchmark is contaminated or broken, "leaderboard progress" can diverge from real capability—and distort model selection.

OpenAI says SWE-bench Verified is saturated due to test-design issues and contamination from public repositories, and recommends reporting **SWE-bench Pro** instead [57][58]. A separate audit summary shared in posts claims that after reviewing **27.6%** of frequently failed tasks, at least **59.4%** had flawed tests that reject correct solutions [59].

Products & Launches

Pip-installable vector search: Alibaba open-sources Zvec

Why it matters: Making vector search a library (not a server) lowers adoption friction for local RAG, edge retrieval, and offline-first apps.

Alibaba open-sourced **Zvec**, described as a vector database you can **pip install** with no servers or Docker [60]. Performance claims shared include **8,000+ QPS** on **10M vectors** and "2×" the previous leader on VectorDBBench [61][62]. Repo: https://github.com/alibaba/zvec [63].

LlamaIndex: LlamaAgents Builder adds file uploads for document workflows

Why it matters: Example documents as context can make "natural language workflow building" more grounded—especially for schema inference and validation.

LlamaIndex added **file upload support** to LlamaAgents Builder, letting users upload example docs so the agent can infer schema, validation rules, and pre/post-processing logic [64][65]. The tool is positioned for scalable extraction with citations over complex documents, with user review before approval [66][67]. Walkthrough + signup links were shared [68][69].

[57] post by @OpenAIDevs
[58] post by @OpenAIDevs
[59] post by @rasbt
[60] post by @LiorOnAI
[61] post by @LiorOnAI
[62] post by @LiorOnAI
[63] post by @LiorOnAI
[64] post by @llama_index
[65] post by @jerryjliu0
[66] post by @jerryjliu0
[67] post by @jerryjliu0
[68] post by @jerryjliu0
[69] post by @llama_index

Image generation: Reve V1.5 reaches the top of Image Arena with 4K output

Why it matters: Arena performance combined with higher-resolution output can influence which models become default choices for commercial design workflows.

Reve launched **Reve V1.5**, a text-to-image model with output up to **4K resolution** [70]. It ranked **top 3** in Image Arena behind GPT-Image-1.5 and Nano Banana Pro variants [71]. Detailed scores: https://arena.ai/leaderboard/text-to-image [72].

Developer tooling highlights

- **Devin Review**: an AI-powered interface for understanding complex PRs; now supports fixing PRs inline by asking Devin to propose changes and applying them with one click [73]. Try: http://devinreview.com [74].
- **LangSmith**: shipped native tracing for **Google ADK** agents [75]. Docs: https://docs.langchain.com/langsmith/trace-with-google-adk [76].
- **OpenRouter**: launched "Effective Pricing," estimating average provider costs based on cache pricing and cache hit rates [77].

Industry Moves

OpenAI expands enterprise deployment via "Frontier Alliances"

Why it matters: Enterprise adoption often hinges on integration and change management, not just model quality.

OpenAI announced **Frontier Alliances** with **BCG, McKinsey, Accenture, and Capgemini** to deploy "OpenAI Frontier" to enterprises globally [78][79]. The partnerships emphasize strategy, workflow redesign, system integration, and change management [80]. Announcement: https://openai.com/index/frontier-alliance-partners [81].

A separate report link shared on X says OpenAI is hiring **hundreds of AI consultants** to boost enterprise sales [82][83].

[70] post by @reve
[71] post by @arena
[72] post by @arena
[73] post by @cognition
[74] post by @cognition
[75] post by @LangChain
[76] post by @LangChain
[77] post by @OpenRouter
[78] post by @bradlightcap
[79] post by @kimmonismus
[80] post by @kimmonismus
[81] post by @bradlightcap
[82] post by @srimuppidi
[83] post by @srimuppidi

Salesforce Ventures invests in Sakana AI

Why it matters: Enterprise-focused AI labs are positioning around vertical credibility and deployment readiness.

Sakana AI announced an investment from Salesforce Ventures and said Salesforce will evaluate integrating Sakana's enterprise technology into Salesforce's global platform offerings [84].

Anthropic hiring: interpretability research engineers

Why it matters: As models become more central to critical workflows, internal understanding of model behavior is treated as infrastructure work.

Anthropic's interpretability team is hiring ~**10 research engineers** (no prior interpretability experience required), targeting seasoned ML infrastructure engineers [85].

Policy & Regulation

Pentagon–Anthropic tensions over Claude safeguards

Why it matters: Government adoption of frontier models is colliding with limits on surveillance and autonomy—potentially shaping what "acceptable safeguards" look like in high-stakes deployments.

An Axios-sourced report shared on X says the Pentagon threatened to ban Claude from classified systems and planned a meeting with Anthropic CEO Dario Amodei involving an "ultimatum" [86][87]. The same reporting says Claude is described as the **only AI model available** in the military's classified systems and the **most capable** model for sensitive defense and intelligence work [88]. Anthropic is said to be willing to loosen restrictions, while still walling off mass surveillance of Americans and autonomous weapons that fire without human involvement [89].

Export controls vs capability diffusion

Why it matters: If frontier capability can be replicated via API outputs—or trained on restricted hardware anyway—policy focus may shift toward access, enforcement, and monitoring.

Posts summarizing Anthropic's position argue that illicit distillation can remove safeguards and feed capabilities into military/intelligence/surveillance systems

[84] post by @SakanaAILabs
[85] post by @ch402
[86] post by @JenGriffinFNC
[87] post by @AndrewCurran_
[88] post by @JenGriffinFNC
[89] post by @JenGriffinFNC

[90]. Separately, Reuters reporting (as cited in posts) claims DeepSeek trained on NVIDIA Blackwell GPUs despite U.S. export restrictions [91][92].

Quick Takes

- **Anthropic AI Fluency Index**: tracked **11 behaviors** across thousands of Claude.ai conversations; one finding shared says **85.7%** of conversations exhibited iteration and refinement [93][94].
- **Claude memorization extraction claim**: researchers reported extracting **95.8%** of *Harry Potter and the Sorcerer's Stone* from Claude Sonnet [95].
- **IBM volatility tied to AI modernization headlines**: a post said IBM stock fell **>10%** after claims that Claude can streamline COBOL code [96].
- **Gemini training for educators**: Google said it's making Gemini training available to **6 million** U.S. K–12 teachers and higher-ed faculty, with modular training and badges [97].
- **Veo 3.1 templates**: Google said templates are rolling out in the Gemini app to provide a "visual foundation" for video creation [98][99].
- **Qdrant 1.17**: announced features include relevance feedback queries, lower latency under heavy writes, and a cluster-wide telemetry API [100].
- **Wispr Flow**: launched on Android as an AI voice dictation app; Richard Socher said he's a "very happy user" [101][102].

Sources

1. post by @AnthropicAI
2. post by @niubi
3. post by @Bayesian0_0
4. post by @LiorOnAI
5. post by @AnthropicAI
6. post by @glennko
7. post by @andersonbcdefg

[90] post by @AnthropicAI
[91] post by @AndrewCurran_
[92] post by @niubi
[93] post by @AnthropicAI
[94] post by @dejavucoder
[95] post by @ivanfioravanti
[96] post by @KobeissiLetter
[97] post by @Google
[98] post by @Google
[99] post by @GeminiApp
[100] post by @qdrant_engine
[101] post by @tankots
[102] post by @RichardSocher

8. post by @si_pbc
9. post by @LiorOnAI
10. post by @Hangsiin
11. post by @OpenAIDevs
12. post by @OpenAIDevs
13. post by @cline
14. post by @leerob
15. post by @OpenAIDevs
16. post by @OpenAIDevs
17. post by @genspark_ai
18. post by @guidelabsai
19. post by @AndrewCurran_
20. post by @niubi
21. post by @intheworldofai
22. post by @dair_ai
23. post by @TheTuringPost
24. post by @TheTuringPost
25. post by @togethercompute
26. post by @OpenAIDevs
27. post by @rasbt
28. post by @LiorOnAI
29. post by @LiorOnAI
30. post by @llama_index
31. post by @jerryjliu0
32. post by @reve
33. post by @arena
34. post by @arena
35. post by @cognition
36. post by @cognition
37. post by @LangChain
38. post by @OpenRouter
39. post by @bradlightcap
40. post by @kimmonismus
41. post by @srimuppidi
42. post by @SakanaAILabs
43. post by @ch402
44. post by @JenGriffinFNC
45. post by @AndrewCurran_
46. post by @AnthropicAI
47. post by @dejavucoder
48. post by @ivanfioravanti
49. post by @KobeissiLetter
50. post by @Google
51. post by @Google
52. post by @GeminiApp
53. post by @qdrant_engine

54. post by @tankots
55. post by @RichardSocher