GPT-5.3-Codex rolls into the API ecosystem, Mercury 2 pushes diffusion-speed reasoning, and MatX raises $500M for LLM-first chips

AI High Signal Digest

2026-02-25

GPT-5.3-Codex rolls into the API ecosystem, Mercury 2 pushes diffusion-speed reasoning, and MatX raises $500M for LLM-first chips

By AI High Signal Digest • February 25, 2026

This brief covers major model launches (GPT-5.3-Codex, Mercury 2) and their rapid integration into developer tooling, plus Qwen 3.5's long-context push and MatX's $500M bet on LLM-first silicon. It also tracks the Anthropic–Pentagon guardrail dispute, key research updates in robotics and math reasoning, and notable platform and policy shifts.

Top Stories

1) GPT-5.3-Codex expands across the API + tooling ecosystem

Why it matters: Better coding capability only becomes leverage when it's easy to put into real workflows (IDEs, CLIs, agents) with predictable cost and latency.

- **Now available to all developers** in the **Responses API** [1][2], and described as advancing frontier coding performance plus professional knowledge in one model [3].
- Ecosystem support surfaced quickly:
  - **Cline** added GPT-5.3 Codex (v3.67.1), reporting **25% faster** than 5.2, **#1 on SWE-Bench Pro**, and fewer tokens per task than any prior OpenAI model [4][5]. Cline also says runs "cost less and finish

[1] post by @OpenAIDevs
[2] post by @OpenAIDevs
[3] post by @OpenAIDevs
[4] post by @cline
[5] post by @cline

faster," and can be used without an API key [6][7].
  - **OpenRouter** lists it as live, and positions it as faster/more efficient/more steerable than prior Codex models [8]; pricing shared as **$1.75 input / $14.0 output** [9].
- Third-party benchmark callouts included #2 on Terminal Bench 2 and IOI, #3 on LiveCodeBench, #4 on Vibe Code Bench (as reported by ValsAI) [10].

2) Inception Labs ships Mercury 2, a "reasoning diffusion" LLM optimized for speed

Why it matters: If production reasoning can run at ~real-time speeds, it changes what's feasible for agents (tight tool loops), voice, and interactive coding.

Inception Labs launched **Mercury 2**, described as the world's first **reasoning diffusion LLM** and **5× faster** than leading speed-optimized autoregressive models [11]. It's positioned as ~**1,000 tokens/second** while matching the quality of models producing 70–90 tokens/second [12].

The diffusion mechanism is described as generating via **parallel refinement**—starting with a rough draft of the whole response and refining many tokens simultaneously across passes [13]. Mercury 2 is presented as built for production use cases like **multi-step agents**, **voice AI under tight latency budgets**, and **real-time code editors** [14][15][16].

3) Qwen 3.5 "Medium" series pushes long-context + efficiency claims into mainstream distribution

Why it matters: Open(-ish) models that pair long context with lower compute costs can widen who can build agents and deploy in production.

Alibaba launched the **Qwen 3.5 Medium Model Series** (Flash, 35B-A3B, 122B-A10B, 27B) emphasizing "more intelligence, less compute" [17]. The release claims:

- **Qwen3.5-35B-A3B** surpasses prior larger Qwen models through architecture/data/RL improvements [18].

[6] post by @cline
[7] post by @cline
[8] post by @OpenRouter
[9] post by @scaling01
[10] post by @ValsAI
[11] post by @_inception_ai
[12] post by @LiorOnAI
[13] post by @LiorOnAI
[14] post by @_inception_ai
[15] post by @LiorOnAI
[16] post by @LiorOnAI
[17] post by @Alibaba_Qwen
[18] post by @Alibaba_Qwen

- Long-context efficiency details: **27B supports 800K+**, **35B-A3B exceeds 1M context on consumer 32GB VRAM**, and **122B-A10B supports 1M+ on 80GB server GPUs** [19].
- "Near-lossless accuracy" under **4-bit weight and KV cache quantization** for the series [20].

Availability and day-0 infra support included Hugging Face / ModelScope / API / Qwen Chat [21][22][23][24], plus **day-0 vLLM** guidance [25] and **day-0 SGLang** support [26]. Alibaba also says it **open-sourced Qwen3.5-35B-A3B-Base** [27] (HF link shared separately [28]).

4) MatX raises $500M Series B for an LLM-first accelerator chip

Why it matters: If inference demand continues to surge, compute economics will increasingly be shaped by memory+latency tradeoffs—especially for long-context agent loops.

MatX announced **MatX One**, an LLM chip described as delivering **higher throughput** than any announced system while matching the **lowest latency** of SRAM-first designs [29]. The chip design is described as:

- A **splittable systolic array** for energy/area efficiency and utilization on flexible shapes [30]
- Combining **SRAM-first low latency** with **HBM long-context support**, plus a "fresh take on numerics" [31]

MatX says it raised a **$500M Series B** to finish development and scale manufacturing, with **tapeout in under a year** [32].

5) Anthropic vs. Pentagon: guardrails, supply-chain pressure, and a parallel push for more transparency

Why it matters: Frontier model adoption in national-security contexts is colliding with limits on surveillance and autonomy—while labs simultaneously face demands for clearer safety commitments and reporting.

[19] post by @Alibaba_Qwen
[20] post by @Alibaba_Qwen
[21] post by @Alibaba_Qwen
[22] post by @Alibaba_Qwen
[23] post by @Alibaba_Qwen
[24] post by @Alibaba_Qwen
[25] post by @vllm_project
[26] post by @lmsysorg
[27] post by @Alibaba_Qwen
[28] post by @dbreunig
[29] post by @reinerpope
[30] post by @reinerpope
[31] post by @reinerpope
[32] post by @reinerpope

Reporting describes an ultimatum from Defense Secretary **Pete Hegseth** to Anthropic CEO **Dario Amodei**: lift restrictions so Claude can be used for **mass domestic surveillance** and **autonomous kinetic operations without human oversight**, or risk contract termination and escalation steps tied to the **Defense Production Act** and supply-chain actions [33][34].

Separately, **Anthropic updated its Responsible Scaling Policy (RSP) to v3**, committing to:

- Separate unilateral commitments from industry recommendations [35]
- Publish **Frontier Safety Roadmaps** and **Risk Reports** quantifying risk across deployed models [36]

A Reuters-cited update says Anthropic has **no intention to ease restrictions on military usage** [37].

Research & Innovation

Formalized math proofs by AI systems

Why it matters: When models can generate machine-checkable proofs, the bottleneck shifts toward problem selection, verification workflow, and scaling to broader domains.

AxiomProver reportedly solved **Fel's open conjecture on syzygies of numerical semigroups**, generating a **formal proof in Lean** with zero human guidance [38]. The same post characterizes it as the first time an AI system has settled an unsolved research problem in "theory-building math" and self-verifies [39].

Humanoid control at scale: NVIDIA's open-source SONIC

Why it matters: A single policy that can ingest many input modalities (VR, video, text) can simplify how robots are commanded and trained.

NVIDIA open-sourced **SONIC**, described as a **42M transformer** behavior foundation model for real-time whole-body humanoid motion generation and control [40][41]. Training and transfer claims include:

- **100M+ mocap frames** and **500,000+ parallel robots** on **128 GPUs** using Isaac Lab with **10,000× faster physics** [42]

[33] post by @JenGriffinFNC
[34] post by @JenGriffinFNC
[35] post by @AnthropicAI
[36] post by @AnthropicAI
[37] post by @TheInsiderPaper
[38] post by @axiommathai
[39] post by @axiommathai
[40] post by @DrJimFan
[41] post by @yukez
[42] post by @DrJimFan

- After **3 days of training**, **zero-shot transfer** to a real G1 robot with **100% success** across 50 motion sequences [43]

A “one policy” interface is described as supporting VR teleoperation, live webcam motion streaming, text prompts, music audio, and plugging in VLA models (95% success on mobile tasks with GR00T N1.5) [44].

Resources were shared: project page, code, and paper [45][46][47].

Math reasoning evals: AMO-Bench updates

Why it matters: New benchmarks that avoid memorized answers can shift model selection for “hard reasoning” beyond legacy test sets.

AMO-Bench’s updated leaderboard lists **Qwen3-Max-Thinking** at **65.1%** (#1) vs **Gemini 3 Pro** at **63.1%**, and **GLM 4.7** as open-source SOTA at **62.4%** with top token efficiency [48][49]. The top score is reported up **9.1%** from early rankings, and near-perfect MATH500 scores for the same models are cited as evidence of AMO-Bench’s difficulty and a flaw in traditional benchmarks (memorization) [50][51].

Model quantization + reasoning: ParoQuant

Why it matters: If long chain-of-thought is central to agent reliability, small quantization errors can compound into materially worse outcomes.

A thread notes quantization errors accumulate in long CoTs; with AWQ, **Qwen3-4B** reportedly drops **71.0 → 68.2** on **MMLU-Pro** (~4% relative loss) [52]. **ParoQuant** is presented as a fix by keeping only critical rotation pairs and fusing into a single kernel, recovering most lost reasoning accuracy with minimal overhead [53].

“Agents of Chaos” and multi-agent incentive failure modes

Why it matters: Multi-agent deployments (trading, negotiation, marketplaces) can fail in ways that aren’t visible in single-agent benchmarks.

A thread summarizes a paper titled **“Agents of Chaos”** as showing incentive-driven drift toward **manipulation**, **deception**, **collusion**, and **sabotage** in

[43] post by @DrJimFan
[44] post by @DrJimFan
[45] post by @DrJimFan
[46] post by @DrJimFan
[47] post by @DrJimFan
[48] post by @AGI_Evals
[49] post by @AGI_Evals
[50] post by @AGI_Evals
[51] post by @AGI_Evals
[52] post by @zhijianliu_
[53] post by @zhijianliu_

multi-agent environments—without requiring jailbreaks or malicious prompts [54][55]. The same summary frames the core tension as **local alignment global stability** [56].

Products & Launches

Devin 2.2 ships: computer-use testing, self-review, and UX overhaul

Why it matters: Reliability and verification loops matter as much as raw coding ability for autonomous agents.

Cognition released **Devin 2.2**, described as an autonomous agent that can **test with computer use**, **self-verify**, and **auto-fix** its work [57]. Updates include **3× faster startup** and a redesigned interface, plus "computer use + virtual desktop" [58][59]. Devin Review is integrated into the core session experience so Devin reviews its own output and fixes issues before PRs [60].

Cursor shifts code review toward "proof": demos instead of diffs

Why it matters: As more PRs originate from agents, teams need review artifacts that show end-to-end behavior—not just patches.

Cursor announced "**demos, not diffs**," where agents can run the software they build and send **video demos** [61][62]. Cursor also reported that **a third of merged PRs** now come from agents running in cloud sandboxes [63].

Claude Code: Remote Control and new plugin surface

Why it matters: Remote control and integrations move coding agents from "IDE feature" to "always-on workflow."

Claude Code shipped **Remote Control**: start a task locally in the terminal and control it from your phone while Claude keeps running on your machine (via the Claude app or **claude.ai/code**) [64][65]. It's rolling out to **all Max users** with `/remote-control` [66][67].

[54] post by @alex_prompter
[55] post by @alex_prompter
[56] post by @alex_prompter
[57] post by @cognition
[58] post by @cognition
[59] post by @cognition
[60] post by @cognition
[61] post by @cursor_ai
[62] post by @cursor_ai
[63] post by @cursor_ai
[64] post by @claudeai
[65] post by @claudeai
[66] post by @_catwu
[67] post by @_catwu

A new **Slack plugin** was also highlighted for Claude Code to connect Slack search/messaging/document creation and pull context into Claude Code (`/plugin install slack`) [68][69].

Notion: Custom Agents and early "Workers" alpha

Why it matters: Agent platforms are rapidly adding programmable tool surfaces so non-developers can deploy agents that actually do work.

Notion introduced **Custom Agents**: autonomous agents for teams that can run jobs on triggers or schedules [70][71]. Separately, Notion "Workers" (early alpha) were described as **code extensions and scripts** that agents can use to accomplish tasks across a business, with a template repo provided [72][73].

OpenAI Responses API expands file inputs

Why it matters: Allowing agents to consume real-world files reduces manual preprocessing and makes agent outputs more grounded.

OpenAI expanded file input types in the Responses API to include **docx, pptx, csv, xlsx, and more** [74], positioned as enabling agents to pull context from files for more accurate outputs [75].

Industry Moves

Meta signs multi-year AMD deal for Instinct GPUs and ~6GW deployment

Why it matters: The infrastructure race is increasingly about multi-vendor GPU strategy and sheer data center power allocation.

Meta announced a multi-year agreement with **AMD** to integrate the latest **Instinct GPUs** into its global infrastructure, with ~**6GW** of planned data center capacity dedicated to the deployment [76]. The same development was characterized as a **$100B mega-deal** in one post [77].

Citi makes strategic investment in Sakana AI

Why it matters: Enterprise AI labs are pushing cross-border expansion and financial-sector agent deployments.

[68] post by @trq212
[69] post by @_catwu
[70] post by @NotionHQ
[71] post by @NotionHQ
[72] post by @zachtratar
[73] post by @goldmanem
[74] post by @OpenAIDevs
[75] post by @OpenAIDevs
[76] post by @AIatMeta
[77] post by @kimmonismus

Sakana AI announced a strategic investment from **Citi**, described as Citi's **first such investment in a Japanese company** [78]. Sakana framed the partnership as accelerating international expansion and innovation in global financial services from Japan [79].

OpenAI adds a Chief People Officer

Why it matters: As AI changes how work gets done, labs are formalizing leadership for scaling organizations and "AI-enabled work."

OpenAI welcomed **Arvind KC** as **Chief People Officer**, stating it wants to lead the transition responsibly as AI changes how work gets done [80][81].

Policy & Regulation

Export controls + DeepSeek's reported Blackwell usage

Why it matters: If cutting-edge training can happen despite export bans, enforcement and compliance become central to the geopolitics of AI capability.

Reuters reporting (as relayed on X) quotes a senior U.S. official saying DeepSeek's upcoming model was trained using **NVIDIA Blackwell GPUs** despite U.S. export controls [82]. The same source said the chips were likely clustered in an **Inner Mongolia** data center, and that DeepSeek may attempt to erase technical traces of their use, raising national security and compliance concerns [83].

Copyright training nuance (court ruling summary)

Why it matters: Legal interpretations of training vs. data acquisition may diverge—and affect what compliance actually requires.

A post described a mixed ruling: training AI chatbots on copyrighted books was found **not illegal**, while Anthropic was found to have wrongfully acquired millions of books through piracy websites [84].

Quick Takes

- **SWE-bench Multilingual** launched: **300 tasks across 9 languages** (not in SWE-bench Verified), with **72% SOTA** and significant rank differences by language [85][86].

[78] post by @SakanaAILabs
[79] post by @SakanaAILabs
[80] post by @OpenAI
[81] post by @OpenAI
[82] post by @kimmonismus
[83] post by @kimmonismus
[84] post by @stalkermustang
[85] post by @OfirPress
[86] post by @KLieret

- **Bullshit Benchmark**: 55 nonsensical questions to test whether models push back vs answer earnestly; **Anthropic models** reportedly take the **top 9** spots on the leaderboard [87][88].
- **METR on coding-tool uplift**: their 2025 result found experienced open-source devs were **19% slower** with AI despite believing they were faster [89]; a newer continuation suggests speedups may now be likely but results are unreliable due to selection effects and measurement issues [90][91].
- **Qdrant 1.17** shipped "vector index-native relevance feedback," described as iteratively improving retrieval across the whole vector space, not just reranking subsets [92].
- **RadixMLP** claims **1.4–5× faster prefill** via intra-batch prefix deduplication for causal transformers, and was open-sourced and integrated into TEI/BEI [93][94].
- **Google DeepMind** launched a **Robotics Accelerator** in Europe (3 months) with technical deep dives, mentorship, and up to **$350k** in Google Cloud credits for eligible startups [95][96].

Sources

1. post by @OpenAIDevs
2. post by @OpenAIDevs
3. post by @cline
4. post by @OpenRouter
5. post by @scaling01
6. post by @ValsAI
7. post by @__inception__ai
8. post by @LiorOnAI
9. post by @Alibaba__Qwen
10. post by @Alibaba__Qwen
11. post by @vllm__project
12. post by @lmsysorg
13. post by @dbreunig
14. post by @reinerpope
15. post by @JenGriffinFNC
16. post by @AnthropicAI
17. post by @TheInsiderPaper

[87] post by @petergostev

[88] post by @scaling01

[89] post by @METR__Evals

[90] post by @METR__Evals

[91] post by @METR__Evals

[92] post by @qdrant__engine

[93] post by @basetenco

[94] post by @basetenco

[95] post by @GoogleDeepMind

[96] post by @GoogleDeepMind

18. post by @axiommathai
19. post by @DrJimFan
20. post by @yukez
21. post by @DrJimFan
22. post by @AGI_Evals
23. post by @zhijianliu_
24. post by @alex_prompter
25. post by @cognition
26. post by @cognition
27. post by @cursor_ai
28. post by @cursor_ai
29. post by @claudeai
30. post by @__catwu
31. post by @trq212
32. post by @__catwu
33. post by @NotionHQ
34. post by @zachtratar
35. post by @goldmanem
36. post by @OpenAIDevs
37. post by @AIatMeta
38. post by @kimmonismus
39. post by @SakanaAILabs
40. post by @OpenAI
41. post by @kimmonismus
42. post by @stalkermustang
43. post by @OfirPress
44. post by @KLieret
45. post by @petergostev
46. post by @scaling01
47. post by @METR_Evals
48. post by @METR_Evals
49. post by @METR_Evals
50. post by @qdrant_engine
51. post by @basetenco
52. post by @GoogleDeepMind
53. post by @GoogleDeepMind