GPT-5.4 rolls out with native computer use; KARL and FlashAttention-4 reshape the agent stack

AI High Signal Digest

2026-03-06

GPT-5.4 rolls out with native computer use; KARL and FlashAttention-4 reshape the agent stack

By AI High Signal Digest • March 6, 2026

OpenAI's GPT-5.4 rollout dominates the cycle, bringing native computer use, tool-search efficiency, and 1M-token context (with real long-context caveats). Also: Databricks' RL-trained KARL knowledge agent, FlashAttention-4's push into mainstream frameworks, a major Anthropic–Pentagon escalation, and a developer-agent supply-chain security incident.

Top Stories

1) OpenAI rolls out GPT-5.4 (Thinking + Pro) with native computer use and 1M context

Why it matters: This is a consolidated "frontier model" push that pairs **agentic coding + tool use + computer control** with **very long context**, which changes what's practical in production workflows (especially multi-step, tool-heavy tasks). [1][2]

Key details (as announced across OpenAI + OpenAI DevRel):

- **Availability / SKUs**: GPT-5.4 is available now in the **API** and **Codex**, with **GPT-5.4 Thinking** and **GPT-5.4 Pro** rolling out in **ChatGPT**

[1] post by @OpenAI

[2] post by @OpenAIDevs

[3][4]. In the API, it's available as `gpt-5.4` and `gpt-5.4-pro` [5].

- **Core capability bundle**: Native computer-use capabilities; up to **1M tokens of context** (Codex + API); "best-in-class agentic coding for complex tasks"; scalable tool search; more efficient reasoning for long, tool-heavy workflows [6].
- **Computer use specifics**: OpenAI Devs says GPT-5.4 can write **Playwright** code, read screenshots, and issue keyboard/mouse actions to operate computers, with steerable behavior and configurable confirmation policies [7].
- **Benchmarks shared by OpenAI Devs**: **83.0%** on GDPval, **75.0%** on OSWorld-Verified, **57.7%** on SWE-Bench Pro (Public), **54.6%** on Toolathlon [8].
- **Efficiency + speed knobs in Codex**: `/fast` mode delivers up to **1.5×** faster performance across supported models (including GPT-5.4) [9]. Separately, a user report notes **1.5× speed at 2× credit consumption** [10].
- **Steering mid-response**: In ChatGPT, OpenAI says you can now interrupt GPT-5.4 Thinking mid-response to add instructions or adjust direction, with steering rolling out on Android and web (iOS "coming soon") [11][12].

Practical caveat on long context:

- Even with a **1M context window**, retrieval degrades at very large contexts. One reported MRCR v2 "needle-in-a-haystack" curve shows **97%** at 16–32K tokens, **57%** at 256–512K, and **36%** at 512K–1M—prompting recommendations to **compact regularly** [13][14].

Relevant links:

- GPT-5.4 announcement page: https://openai.com/index/introducing-gpt-5-4/ [15]
- Codex `/fast` details: https://developers.openai.com/codex/speed/ [16]

[3] post by @OpenAI
[4] post by @OpenAIDevs
[5] post by @OpenAIDevs
[6] post by @OpenAIDevs
[7] post by @OpenAIDevs
[8] post by @OpenAIDevs
[9] post by @OpenAIDevs
[10] post by @reach_vb
[11] post by @OpenAI
[12] post by @OpenAI
[13] post by @cline
[14] post by @cline
[15] post by @OpenAIDevs
[16] post by @reach_vb

2) Databricks releases KARL, an RL-trained "knowledge agent" aimed at grounded enterprise reasoning

Why it matters: KARL is a concrete example of applying RL to **non-verifiable enterprise knowledge tasks** (messy docs, long tool chains), and Databricks frames it as an **"assembly line"** for producing agents—important for teams trying to move beyond "RAG as a demo." [17][18]

What was announced:

- **What it is**: KARL (Knowledge Agents from Reinforcement Learning) is an RL-trained agent for document-centric grounded reasoning over complex questions, "millions of documents," "hundreds of tool calls," and repeated context compression [19].
- **Performance framing**: Databricks describes "frontier-level performance on complex knowledge workloads at a fraction of the cost and latency of leading proprietary models" [20].
- **Why RL here**: Databricks emphasizes these enterprise tasks "are not strictly verifiable" like unit-test-style RL wins [21].
- **Mechanics (high level)**: Off-policy RL with synthetic data (OAPL), multi-task RL that generalizes, and "parallel thinking" test-time compute to manage latency [22][23][24].
- **RAG++++ detail**: A VentureBeat summary highlights KARL matching frontier quality on messy enterprise data by running up to **200 vector searches per query** [25].

Links:

- Tech report PDF: https://www.databricks.com/sites/default/files/2026-03/karl.pdf [26]
- Databricks blog: http://databricks.com/blog/meet-karl-faster-agent-enterprise-knowledge-powered-custom-rl [27]

---

[17] post by @jefrankle
[18] post by @jefrankle
[19] post by @jefrankle
[20] post by @DbrxMosaicAI
[21] post by @TechJournalist
[22] post by @jefrankle
[23] post by @jefrankle
[24] post by @jefrankle
[25] post by @TechJournalist
[26] post by @jefrankle
[27] post by @jefrankle

3) FlashAttention-4 goes GA; PyTorch adds a FlashAttention-4 backend for FlexAttention

Why it matters: Attention kernels are a performance ceiling for both training and inference. FA4 is positioned as a Blackwell-era redesign that shifts bottlenecks away from softmax/SMEM limits, while PyTorch is trying to make these gains accessible for **custom attention variants** (not only a single "blessed" kernel). [28][29]

What's new:

- **FA4 GA**: "FlashAttention-4 is GA" [30].
- **Core performance claim**: FA4 reaches ~**1600 TFLOPs** attention on Blackwell GPUs and is described as "pretty much at matmul speed," by changing the algorithm/pipeline so softmax and shared memory bandwidth no longer dictate speed [31].
- **PyTorch integration**: PyTorch added a **FlashAttention-4 backend** to **FlexAttention** on Hopper and Blackwell GPUs; PyTorch now auto-generates CuTeDSL score/mask modifications and JIT-instantiates FA4 for custom attention variants [32]. PyTorch reports **1.2× to 3.2× speedups** over Triton on compute-bound workloads [33].
- **Transformers integration (in progress)**: A PR for FA4 integration into Hugging Face Transformers was shared (PR #42435) [34].

---

4) Anthropic–Pentagon escalation: "supply chain risk" designation + Amodei statement

Why it matters: This is a high-stakes governance signal: AI labs are increasingly treated as **critical suppliers** (and potential risks) in national-security procurement, with direct implications for enterprise adoption, contracts, and oversight.

Reported developments:

- **Designation**: A post claims the Pentagon formally notified Anthropic it's been deemed a **"supply chain risk"** [35].
- **Amodei response (as summarized)**: A memo-style summary says Amodei apologized for the tone of a leaked memo, said it was outdated/not his considered view, emphasized keeping warfighters equipped, and offered

[28] post by @tedzadouri

[29] post by @PyTorch

[30] post by @vipulved

[31] post by @tedzadouri

[32] post by @PyTorch

[33] post by @PyTorch

[34] post by @StasBekman

[35] post by @Polymarket

Claude to the military at nominal cost with forward-deployed engineer support [36].

- **Anthropic's statement link**: Anthropic shared a statement from Amodei: https://www.anthropic.com/news/where-stand-department-war [37].

> "Anthropic has much more in common with the Department of War than we have differences." [38]

---

5) Security incident report: "Clinejection" installs a separate agent (OpenClaw) without consent

Why it matters: Agentic dev tools run with broad local permissions; supply-chain style incidents can turn "developer convenience" into fleet-wide risk.

- A write-up alleges "every developer who installed or updated Cline got OpenClaw … installed globally on their machine without consent," describing it as "malicious agent injection" and noting OpenClaw has "full system access" [39].

Details: https://grith.ai/blog/clinejection-when-your-ai-tool-installs-another [40]

Research & Innovation

Why it matters: This week's research is converging on a few themes: **RL methods for messy tasks**, **hybrid architectures** for scaling efficiency, and **benchmarks** that better approximate real agent constraints (implicit rules, over/underthinking, interaction).

Open models + hybrid architectures

- **OLMo Hybrid (AI2)**: Allen AI released OLMo Hybrid, mixing transformer attention with linear RNN layers; the team claims hybrid models are "strictly more expressive" than either alone and that this translates to better scaling (49% fewer tokens to match OLMo 3 MMLU accuracy) [41].
- **Training "fully in the open"**: Lambda says OLMo Hybrid 7B was trained in the open with training logs/recovery metrics/weights, using **3T tokens**, **512 NVIDIA Blackwell GPUs**, over **7 days**, with **97% active training time** and median recovery under **4 minutes** [42].

[36] post by @cgtwts
[37] post by @AnthropicAI
[38] post by @theo
[39] post by @mmitchell_ai
[40] post by @mmitchell_ai
[41] post by @tyleraromero
[42] post by @LambdaAPI

RL + evaluation research (Meta FAIR ICLR set)

- Meta FAIR says its team co-authored **7 papers accepted to ICLR**, covering topics including joint safety agents (“Alignment Waltz”), judge RL (“J1”), experience synthesis for agent learning, and benchmarks for over/underthinking (“OptimalThinkingBench”) [43][44][45].

Data efficiency for language models

- **Semantic Tube Prediction (STP)**: STP (co-authored by Yann LeCun) is described as forcing hidden states into locally linear “semantic tubes,” matching baseline accuracy with **16× less training data** [46][47]. Paper: https://arxiv.org/abs/2602.22617 [48].

Benchmarks for agent “implicit constraints”

- **Implicit Intelligence**: Labelbox Applied ML Research introduced a benchmark testing whether agents respect **unstated constraints** across implicit reasoning, catastrophic risk, privacy/security, and accessibility [49]. Paper: https://arxiv.org/abs/2602.20424 [50].

Long-running agents: context compression as a core problem

- **Baseten KV-cache compression**: Baseten reports one-shot compaction preserves detailed information with **65–80% accuracy** at **2–5× compression** (outperforming text summarization) and explores what happens when you compress repeatedly for persistent agents [51][52].

Products & Launches

Why it matters: The biggest product shifts are around **agent scaffolding**: better computer-use interfaces, orchestration/automation, and cross-tool connectivity (so agents can actually act, not just chat).

GPT-5.4 distribution and integrations

- **GitHub Copilot**: GitHub says GPT-5.4 is now generally available and rolling out in Copilot; early testing highlights “enhanced logical reasoning

[43] post by @jaseweston
[44] post by @jaseweston
[45] post by @jaseweston
[46] post by @scaling01
[47] post by @scaling01
[48] post by @scaling01
[49] post by @TheTuringPost
[50] post by @TheTuringPost
[51] post by @basetenco
[52] post by @basetenco

and task execution" [53][54]. Changelog: https://github.blog/changelog/2026-03-05-gpt-5-4-is-generally-available-in-github-copilot/ [55].

- **Cursor**: Cursor says "GPT 5.4 is now available in Cursor," and they found it "more natural and assertive than previous models" [56][57].
- **Perplexity**: Perplexity announced GPT-5.4 and GPT-5.4 Thinking availability for Pro/Max subscribers [58].
- **Arena**: Arena reports GPT-5.4 variants in Text/Vision/Code arenas and publishes ranking highlights (e.g., GPT-5.4-high tied with Gemini-3-Pro in Text Arena) [59][60].

Codex tooling updates

- **Codex app on Windows**: OpenAI Devs announced Codex is now on Windows with a "native agent sandbox" and PowerShell support [61]. Landing page: https://developers.openai.com/wendows [62].

Always-on agent operations

- **Cursor Automations**: Cursor introduced Automations for always-on agents that run based on triggers and instructions you define [63]. Blog: http://cursor.com/blog/automations [64].

Office / finance workflow tooling

- **ChatGPT for Excel**: OpenAI launched "ChatGPT for Excel," positioning it as bringing ChatGPT into spreadsheet workflows ("where decisions get made") [65]. Link: https://openai.com/index/chatgpt-for-excel/ [66].

Video generation continues to split into "engines" vs "story tools"

- **Bing Video Creator**: Microsoft rolled out "Sora 2 generative video" in Bing Video Creator, adding audio integration and watermark + C2PA credentials [67][68].

[53] post by @github
[54] post by @github
[55] post by @github
[56] post by @cursor_ai
[57] post by @cursor_ai
[58] post by @perplexity_ai
[59] post by @arena
[60] post by @arena
[61] post by @OpenAIDevs
[62] post by @OpenAIDevs
[63] post by @cursor_ai
[64] post by @cursor_ai
[65] post by @angadsg
[66] post by @BorisMPower
[67] post by @JordiRib1
[68] post by @JordiRib1

- **PAI (Utopai Studios)**: Utopai says PAI is rolling out as a long-form cinematic model with **minutes-long** continuous generation, character/scene consistency, and natural-language editing [69].
- **LTX-2.3 on fal**: fal says LTX-2.3 is live with Pro (audio-to-video, retake, extend) and Fast modes plus sharper detail/cleaner audio/stronger motion [70].

Industry Moves

Why it matters: Distribution and enterprise positioning are starting to matter as much as raw model quality—especially for agents (where tool ecosystems + integrations decide what gets adopted).

- **Together AI fundraising (reported)**: Together AI is reportedly raising **$1B** at a **$7.5B** pre-money valuation, generating ~**$1B ARR**, with growth tied to moving from leasing GPUs to buying their own GPUs to rent out [71][72].
- **Codex user growth**: Codex surpassed **2M+ active users**, up **25% week-over-week** (noted as before Windows + GPT-5.4 launch) [73].
- **Claude adoption**: One post claims “more than a million people are now signing up for Claude every day” [74].
- **Sakana AI × MUFG**: Sakana AI and Mitsubishi UFJ Bank advanced their “AI Lending Expert” system from ~6-month PoC to **real-case verification phase** [75]. Link: https://sakana.ai/mufg-ai-lending [76].

Policy & Regulation

Why it matters: Export controls and professional-liability rules can become hard constraints on where AI can be deployed—and what assistants can legally do.

- **US AI chip export restrictions (reported)**: A post says the Trump Administration is preparing a rule to restrict AI chip shipments globally without US approval, requiring permission for “virtually all exports of AI chips,” with Nvidia and AMD heavily impacted [77][78][79].
- **New York bill targeting “substantive responses”**: A New York bill would ban AI from answering questions related to licensed professions (medicine, law, dentistry, nursing, psychology, social work, engineering,

[69] post by @UtopaiStudios
[70] post by @fal
[71] post by @steph_palazzolo
[72] post by @steph_palazzolo
[73] post by @thsottiaux
[74] post by @mikeyk
[75] post by @SakanaAILabs
[76] post by @SakanaAILabs
[77] post by @KobeissiLetter
[78] post by @KobeissiLetter
[79] post by @KobeissiLetter

and more), and companies would be liable if chatbots give "substantive responses" in these areas [80][81].

Quick Takes

Why it matters: Smaller releases often become "quiet defaults" inside stacks—especially around evaluation, routing, and on-device constraints.

- **OpenAI: Chain-of-Thought controllability**: OpenAI published a new evaluation suite/paper and says GPT-5.4 Thinking shows "low ability to obscure its reasoning," suggesting CoT monitoring remains a useful safety tool [82].
- **Gemini 3.1 Flash-Lite preview (pricing)**: Google launched Gemini 3.1 Flash-Lite in preview at **$0.25 / 1M input tokens** for high-volume developer workloads [83].
- **Perplexity "Model Council"**: Perplexity launched a mode that runs GPT-5.4, Claude Opus 4.6, and Gemini 3.1 Pro simultaneously and selects the best answer in one workflow [84].
- **OLMo Hybrid (distribution)**: AI2 released a family of OLMo Hybrid models (base/SFT/DPO) on Hugging Face [85][86].
- **FlashAttention-4 resources**: FA4 paper and code links shared (paper PDF + GitHub repo) [87][88].
- **LiquidAI on-device agent**: A 24B-parameter model (2.3B active per token) is reported to fit in **14.5GB** and run tool selection with **385ms** average latency (67 tools, 13 MCP servers) with "zero network calls" [89][90].
- **OpenHands Critic v1.0**: OpenHands released a "critic" model that scores coding agent traces to address the verification bottleneck, with real-time thumbs-up/down monitoring and support in SDK/CLI/Hugging Face [91][92].
- **LangChain skills evaluation**: LangChain released an evaluation benchmark for LangSmith/LangChain "skills," emphasizing variance across tasks for coding agents [93]. Repo: https://github.com/langchain-ai/skills-benchmarks [94].

[80] post by @MorePerfectUS
[81] post by @MorePerfectUS
[82] post by @OpenAI
[83] post by @dl_weekly
[84] post by @AskPerplexity
[85] post by @mervenoyann
[86] post by @mervenoyann
[87] post by @tedzadouri
[88] post by @tedzadouri
[89] post by @LiorOnAI
[90] post by @liquidai
[91] post by @xingyaow_
[92] post by @gneubig
[93] post by @LangChain
[94] post by @LangChain

- **GitHub AGENTS.md guidance**: GitHub's analysis of 2,500+ repos suggests effective AGENTS.md files stay brief and include persona, exact commands, boundaries, and good output examples [95][96].

Sources

1. post by @OpenAI
2. post by @OpenAIDevs
3. post by @OpenAIDevs
4. post by @OpenAIDevs
5. post by @OpenAIDevs
6. post by @OpenAIDevs
7. post by @reach_vb
8. post by @OpenAI
9. post by @cline
10. post by @jefrankle
11. post by @jefrankle
12. post by @jefrankle
13. post by @DbrxMosaicAI
14. post by @TechJournalist
15. post by @jefrankle
16. post by @jefrankle
17. post by @jefrankle
18. post by @jefrankle
19. post by @jefrankle
20. post by @tedzadouri
21. post by @PyTorch
22. post by @vipulved
23. post by @StasBekman
24. post by @Polymarket
25. post by @cgtwts
26. post by @AnthropicAI
27. post by @theo
28. post by @mmitchell_ai
29. post by @tyleraromero
30. post by @LambdaAPI
31. post by @jaseweston
32. post by @scaling01
33. post by @TheTuringPost
34. post by @TheTuringPost
35. post by @basetenco
36. post by @github

[95] post by @omarsar0

[96] post by @omarsar0

37. post by @cursor_ai
38. post by @perplexity_ai
39. post by @arena
40. post by @arena
41. post by @OpenAIDevs
42. post by @cursor_ai
43. post by @angadsg
44. post by @BorisMPower
45. post by @JordiRib1
46. post by @UtopaiStudios
47. post by @fal
48. post by @steph_palazzolo
49. post by @thsottiaux
50. post by @mikeyk
51. post by @SakanaAILabs
52. post by @KobeissiLetter
53. post by @MorePerfectUS
54. post by @OpenAI
55. post by @dl_weekly
56. post by @AskPerplexity
57. post by @mervenoyann
58. post by @mervenoyann
59. post by @tedzadouri
60. post by @LiorOnAI
61. post by @liquidai
62. post by @xingyaow_
63. post by @gneubig
64. post by @LangChain
65. post by @omarsar0