OpenAI's $110B raise, U.S. defense deployment deals, and DeepSeek V4 signals

AI High Signal Digest

2026-02-28

OpenAI's $110B raise, U.S. defense deployment deals, and DeepSeek V4 signals

By AI High Signal Digest • February 28, 2026

OpenAI announces a $110B funding round and expanded infrastructure partnerships, then signs a classified-network deployment agreement with the Department of War as Anthropic faces a supply-chain risk designation and prepares a court challenge. Also: DeepSeek V4 timing signals and systems commits, plus notable advances in video generation, KV-cache efficiency, and privacy-preserving inference wrappers.

Top Stories

1) OpenAI's $110B funding round reshapes the infra/partner map

Why it matters: This round ties OpenAI's growth directly to **specific cloud + chip roadmaps** (AWS/Trainium, NVIDIA systems, Azure API exclusivity), and signals how competitive advantage is increasingly negotiated through infrastructure access and distribution.

OpenAI CEO Sam Altman said OpenAI raised a **$110B** round from **Amazon, NVIDIA, and SoftBank** [1]. Reporting in the same cycle pegged the round at a **$730B pre-money valuation**, with amounts broken out as **$50B (Amazon), $30B (SoftBank), $30B (NVIDIA)** [2]. OpenAI's own messaging framed this as scaling infrastructure "to bring AI to everyone," supported by those partners [3].

Partnership details highlighted publicly include: - **Amazon/AWS**: New enterprise products including a **stateful runtime environment** and use of **Trainium** [4]. A separate OpenAI-partner post described co-building a **Stateful Runtime** for agentic apps on **Bedrock**, scaling with **2GW of Trainium compute**, and creating **custom models for Amazon apps** [5]. - **Microsoft**:

OpenAI said its **stateless API will remain exclusive to Azure**, alongside plans to build more capacity with Microsoft [6]. - **NVIDIA**: OpenAI described NVIDIA chips as foundational, and said it's excited to run NVIDIA systems in **AWS** [7]. NVIDIA also said it's entering a "next phase" with OpenAI to deploy **5GW on Vera Rubin** for training and inference [8, 9].

On finances, Epoch AI noted the round "nearly triples" OpenAI's total raised so far and cited a projection (attributed to *The Information*) of **$157B cash burn through 2028**, saying this round plus **$40B cash on hand** roughly matches that projection [10].

2) OpenAI says it reached a classified-network agreement with the U.S. Department of War

Why it matters: Frontier-lab defense deployments are becoming **contract + control-system design problems**: what gets prohibited, who enforces it, and what technical safeguards accompany access.

Altman said OpenAI reached an agreement with the **Department of War (DoW)** to deploy models in its **classified network** [11]. He said the agreement embeds OpenAI's safety principles—**prohibiting domestic mass surveillance** and requiring **human responsibility for use of force** (including autonomous weapon systems)—and that DoW agrees and reflects these in law/policy and the agreement [11].

OpenAI also described additional deployment measures: - **Technical safeguards** to ensure model behavior [11] - **Field Deployed Engineers (FDEs)** to help with the models and safety [11] - **Cloud networks only** [11]

Altman further said OpenAI is asking DoW to offer the "same terms to all AI companies," and expressed a desire to de-escalate away from legal/government actions toward "reasonable agreements" [11].

A DoW official account characterized the OpenAI contract as flowing from a touchstone of **"all lawful use"**, referencing legal authorities and mutually agreed safety mechanisms (described as a compromise Anthropic was offered and rejected) [12].

3) Anthropic faces a U.S. government crackdown; says it will fight a "supply chain risk" label in court

Why it matters: "Supply chain risk" designations and procurement rules can reshape the AI market indirectly—by forcing contractors and cloud ecosystems to pick sides.

Secretary of War Pete Hegseth's account said the DoW is designating Anthropic a **"Supply-Chain Risk to National Security"**, and that "no contractor, supplier, or partner that does business with the United States military may conduct any commercial activity with Anthropic," while allowing Anthropic to

provide services for **no more than six months** for transition [13]. The same post tied this to the President's directive for the federal government to **cease all use of Anthropic's technology** [13].

Anthropic subsequently said it will **challenge any supply chain risk designation in court**, arguing the DoW cannot restrict customers' use of Claude outside of DoW contract work [14].

Several posts highlighted second-order implications: Anthropic serves models via cloud providers including **AWS (primary), Google Cloud, and Azure** [15], and another post argued that since those providers do business with the U.S. military, a literal interpretation could block Anthropic from serving via them [15].

4) DeepSeek V4 countdown: release timeline + chip collaborations + major systems work

Why it matters: DeepSeek's next release is being framed as both a **capability event** and a **hardware-optimization event**, alongside continued investment in the systems stack.

Multiple posts citing the *Financial Times* said DeepSeek is set to release **DeepSeek-V4 next week** [16, 17], and is working with Chinese AI chipmakers **Huawei** and **Cambricon** to optimize V4 for their latest products [16]. The same reporting said a **brief technical note** will accompany the release, followed by a more comprehensive report about a month later [16].

Separately, DeepSeek made a "major commit" to **DeepGEMM**, adding **mHC integration**, early support for **NVIDIA Blackwell (SM100)**, and **FP4 ultra-low precision computing** [18].

There were also signals of a web update: a post claimed DeepSeek was updated with **knowledge cutoff May 2025** and **1M token context length**, "likely V4 (or V4 lite)," pushed to the web [19, 20].

5) Kling 3.0 claims the top Text-to-Video spot (with and without audio)

Why it matters: Video generation is rapidly stratifying around **quality tiers, audio integration, and leaderboard-driven iteration**, with pricing now comparable across leading tools.

Artificial Analysis reported **Kling 3.0 (1080p Pro)** took **#1 in Text-to-Video** across both With Audio and Without Audio leaderboards, surpassing Grok Imagine, Runway Gen-4.5, and Veo 3.1 [21]. The release supports up to **15-second generations** and **native audio**, with 1080p (Pro) and 720p (Standard) tiers [21].

Kling also released **Kling 3.0 Omni**, a unified multimodal model supporting image/video inputs, editing, and generation; Omni 1080p (Pro) placed **#2** in

Text-to-Video With Audio and **#4** in No Audio [21]. Pricing cited: ~**$13/min** (1080p Pro, no audio) and ~**$20/min** (with audio); 720p Standard ~**$10/min** (no audio) and ~**$15/min** (with audio) [21].

Research & Innovation

Why it matters: Several releases this period target practical bottlenecks—**long-context cost**, **KV-cache memory**, and **stable post-training**—which increasingly determine what "agentic" systems can do in production.

- **Instant model customization via Doc-to-LoRA / Text-to-LoRA (Sakana AI)**: Sakana introduced hypernetwork methods that generate task- or document-specific LoRA adapters "on the fly," turning customization into a single forward pass rather than fine-tuning or long prompts [22]. Reported results include near-perfect needle-in-a-haystack performance on instances **5× longer** than the base model's context window [22] and **sub-second latency** for rapid experimentation [22]. A separate summary emphasized Doc-to-LoRA compressing long documents into adapters to avoid repeated context re-reading, improving memory/update latency and serving cost for long-document agents [23].
- **Self-managed KV cache (NVIDIA SideQuest)**: SideQuest has the reasoning model decide which tokens remain useful and "garbage collect" the rest, running this management as an auxiliary task so it doesn't pollute the main context [24]. Trained with **215 samples**, it reduced peak token usage by up to **65%** with minimal accuracy loss [24].
- **Off-policy RL for reasoning (Databricks OAPL)**: Databricks said its OAPL approach shows you don't need strict on-policy training to improve reasoning [25]. Reported metrics: matches/beats GRPO, remains stable with large policy lag, and uses ~**3× fewer** training generations [25].
- **Agentic inference systems (DeepSeek DualPath)**: A DeepSeek/THU/PKU paper summary described DualPath pooling otherwise-mismatched NIC bandwidth between prefill and decode to move KV cache more efficiently [26]. Reported results: up to **1.87× speedup** on DS-660B offline inference [26] and positioning for higher concurrency/lower cost in multi-agent systems with repeated long-context KV-cache access [26].
- **Physics-aware image editing (PhysicEdit)**: PhysicEdit reframes editing as a **physical state transition** and distills transition priors from videos into a latent representation for more physically plausible edits [27, 28]. It introduced the **PhysicTran38K** dataset (38K video trajectories with reasoning traces) [29] and reported benchmark improvements over prior approaches [30].
- **Long-term coherence eval (YC Bench)**: YC Bench simulates "running a startup" for **three years** to test long-horizon agent coherence [31]. It reported **GPT-5.2** (and sometimes Sonnet 4.6) "quickly goes bankrupt"

and fails to beat a sub-optimal greedy baseline [31], while **Gemini-3-Flash** was described as matching the baseline via multi-stage strategy in the provided scratchpad [31].

Products & Launches

Why it matters: The ecosystem continues shifting from chat to **systems that execute work**—with privacy wrappers, agents that run while you're away, and developer-grade infrastructure for evaluation and ranking.

- **Open Anonymity "unlinkable inference" (privacy wrapper for remote models)**: Open Anonymity described a "VPN for AI inference" layer that uses decentralized proxies and blind signatures to make requests hard to link back to users across time [32]. It emphasized ephemeral keys per session/request to combat longitudinal tracking [32] and shipped an open chat app, **oa-chat**, with local chat history and temporary keys for OpenAI calls [33]. Resources: https://chat.openanonymity.ai/ [32] and https://openanonymity.ai/blog/unlinkable-inference/ [32].
- **Hermes Agent (NousResearch) adds OCR/document extraction skill**: Hermes Agent is positioned as an open-source agent with multi-level memory and persistent dedicated machine access [34]. A recent update added broad OCR/document extraction (PDFs, ePubs, DocX, PowerPoint, etc.) [35].
- **Claude Code Remote Control**: A rollout to Claude Code Pro users enables "remote control," with instructions to update to v2.1.58+, log out/in for new flags, and use `/remote-control` [36].
- **Gemma on iOS via Google AI Edge Gallery**: A post said the Google AI Edge Gallery app brings fully offline, on-device AI to iOS (chat, image Q&A, audio transcription/translation, voice commands), with an App Store link [37, 38].
- **Perplexity embeddings open-sourced (bidirectional + context-aware variants)**: Perplexity open-sourced four bidirectional embedding models (0.6B and 4B parameters; standard and context-aware types) [39]. The "context-aware" version processes an entire document so chunks "know" the full document meaning [39]. Collection: https://huggingface.co/collections/perplexity-ai/pplx-embed [40].
- **Arena-Rank (open-source leaderboard construction)**: Arena released **Arena-Rank**, a Python package for statistically grounded, reproducible leaderboards using pairwise comparison data [41]. GitHub: https://github.com/lmarena/arena-rank [42].

Industry Moves

Why it matters: This week's biggest competitive moves were about **capital + distribution + compute**—and about who controls the "control plane" (cloud distribution, identity, and evaluation infrastructure).

- **AWS frames the OpenAI partnership as distribution + runtime + Trainium adoption**: Amazon's CEO described a stateful runtime environment on Amazon Bedrock powered by OpenAI intelligence for developers running OpenAI services on AWS [43]. He also said OpenAI is "going big" on Trainium, describing Trainium as **30–40% more price performant than comparable GPUs** [43], and said AWS will be the exclusive third-party cloud distribution provider for **OpenAI Frontier** (agent teams) [43].
- **Microsoft–OpenAI joint statement on AGI definition unchanged**: A Microsoft/OpenAI joint statement was shared alongside commentary that the contractual definition and determination process for AGI remains unchanged despite new funding and partnerships [44, 45]. The AGI definition quoted: a system that can perform "most economically valuable tasks better than humans," and is officially declared AGI by the OpenAI board [45].
- **Guidde raises $50M (agent training from screen-recordings)**: Guidde raised **$50M** to train AI agents on expert screen-recording videos rather than static documentation, claiming **41%** reduction in video creation time and **34%** fewer support tickets [46].
- **Taalas launches first product (models encoded into chips)**: Taalas said it launched its first product after **$30M** in development by **24** people, emphasizing specialization, speed, and power efficiency [47]. A separate summary described "Hardcore Models" chips that store weights on-chip (mask ROM) and can reach **16–17k tokens/sec** inference, with RAM for KV cache and small updates like LoRA [48].
- **OpenAI Codex usage growth**: A post said Codex added **600k users in three weeks**, moving from **1M WAU** (Feb 4) to **1.6M WAU** (Feb 27) [49].

Policy & Regulation

Why it matters: AI governance is being operationalized through **procurement, designations, and deployment constraints**, not just principles—and the effects spill into cloud ecosystems and enterprise buyers.

- **Anthropic refuses to enable domestic mass surveillance or fully autonomous weapons**: A post quoted Anthropic's position: "threats do not change our position: we cannot in good conscience accede to their re-

quest," framed as a moral line against enabling **mass domestic surveillance** and **fully autonomous weapons** [50].

- **DoW designates Anthropic a supply-chain risk; broad procurement restrictions announced**: DoW's directive said Anthropic will be designated a **Supply-Chain Risk to National Security**, and barred contractors/suppliers/partners doing business with the U.S. military from commercial activity with Anthropic, with a 6-month transition window [13]. Anthropic says it will challenge the designation in court [14].
- **OpenAI–DoW agreement highlights "no domestic mass surveillance" and "human responsibility for force"**: OpenAI's agreement announcement reiterated these core principles as incorporated into the contract and reflected in DoW law/policy framing [11].
- **Pentagon cyber tooling (FT-cited) aims at mapping/exploiting vulnerabilities in Chinese infrastructure**: A post citing the FT said the Pentagon is developing AI-powered cyber tools to map and exploit vulnerabilities in Chinese infrastructure (e.g., power grids and sensitive networks), automating reconnaissance and speeding targeting [51].

Quick Takes

Why it matters: These smaller items collectively show where momentum is compounding: usage scale, agent reliability, model-serving efficiency, and fast-moving leaderboards.

- **ChatGPT scale**: ChatGPT crossed **900M weekly users** and **50M paying subscribers** [52].
- **ChatGPT Android**: The Android app (v1.2026.055) mentions a "**Naughty chats**" setting for 18+ users [53].
- **GPT-5.3-Codex cost/throughput notes**: Reported as **28% cheaper** than GPT-5.2 (xhigh) on Artificial Analysis [54], with a post also calling it more token efficient than 5.2 [54]. Another post cited **400k context** and "extra high thinking" in settings [55].
- **Open models: Feb Text Arena**: Arena's Top 3 open models were **GLM-5 (1455)**, **Qwen-3.5 397B A17B (1454)**, and **Kimi-K2.5 Thinking (1452)** [56].
- **vLLM on AMD GPUs**: vLLM described ROCm attention backends delivering up to **4.4× decode throughput** on AMD GPUs, with model-specific benchmarks (e.g., Qwen3-235B MHA 2.7–4.4× TPS) and a one-env-var enablement path [57, 58, 59].
- **UI-agent click accuracy fix**: Tzafon claimed scaling positional embeddings **3×** improved click accuracy from **40% to 80%** with no retraining [60].
- **RF-DETR on Apple MLX**: A post said RF-DETR on MLX runs at **100+ FPS** on an M4 Pro Mac [61].

Sources

1. X post by @sama
2. X post by @TheRundownAI
3. X post by @OpenAI
4. X post by @sama
5. X post by @snsf
6. X post by @sama
7. X post by @sama
8. X post by @nvidianewsroom
9. X post by @kimmonismus
10. X post by @EpochAIResearch
11. X post by @sama
12. X post by @UnderSecretaryF
13. X post by @SecWar
14. X post by @iScienceLuvr
15. X post by @deredleritt3r
16. X post by @jukan05
17. X post by @scaling01
18. X post by @chetaslua
19. X post by @teortaxesTex
20. X post by @teortaxesTex
21. X post by @ArtificialAnlys
22. X post by @SakanaAILabs
23. X post by @omarsar0
24. X post by @dair_ai
25. X post by @DbrxMosaicAI
26. X post by @ZhihuFrontier
27. X post by @RisingSayak
28. X post by @RisingSayak
29. X post by @RisingSayak
30. X post by @RisingSayak
31. X post by @HeMuyu0327
32. X post by @kenziyuliu
33. X post by @percyliang
34. X post by @NousResearch
35. X post by @Teknium
36. X post by @noahzweben
37. X post by @_philschmid
38. X post by @_philschmid
39. X post by @LiorOnAI
40. X post by @LiorOnAI
41. X post by @arena
42. X post by @arena

43. X post by @ajassy
44. X post by @MSFTnews
45. X post by @kimmonismus
46. X post by @dl_weekly
47. X post by @taalas_inc
48. X post by @TheTuringPost
49. X post by @swyx
50. X post by @mmitchell_ai
51. X post by @clashreport
52. X post by @nickaturley
53. X post by @btibor91
54. X post by @JasonBotterill
55. X post by @pierceboggan
56. X post by @arena
57. X post by @vllm_project
58. X post by @vllm_project
59. X post by @vllm_project
60. X post by @tzafon_company
61. X post by @skalskip92