Anthropic Files for IPO as OpenAI Lands on Bedrock and Perplexity Rewrites Search

AI High Signal Digest

2026-06-02

Anthropic Files for IPO as OpenAI Lands on Bedrock and Perplexity Rewrites Search

By AI High Signal Digest • June 2, 2026

Anthropic's confidential IPO filing, OpenAI's Bedrock expansion, and Perplexity's code-driven search architecture led the day. The brief also covers ARC-AGI progress, new agent tooling from Qwen and Google, and major capital and infrastructure bets from Alphabet, OpenAI, and NVIDIA.

Top Stories

Why it matters: distribution, capital markets, and agent architecture all shifted meaningfully today.

- **OpenAI expanded onto AWS.** GPT-5.5, GPT-5.4, and Codex are now generally available on Amazon Bedrock, giving enterprises access through AWS security, compliance, and governance workflows they already use [1, 2]. AWS said Bedrock adds automatic scaling, and Codex now supports CLI, desktop, and IDE workflows on Bedrock with AWS-native auth/IAM [2, 3]. OpenAI said this is the start of a broader AWS expansion, including future cybersecurity tools like Daybreak [1].
- **Anthropic took a formal step toward public markets.** The company said it confidentially submitted a draft S-1 to the SEC, which gives it the option to pursue an IPO pending SEC review [4]. That makes Anthropic the latest frontier lab moving from private funding toward public-market preparation.
- **Perplexity changed how its agents search.** Its new Search as Code system has models write Python that calls search primitives directly instead of looping through one function call at a time [5, 6]. Perplexity said the design cuts latency and context pollution, and reported wins or ties

across DSQA, BrowseComp, HLE, WideSearch, and WANDR, including 0.871 on DSQA versus Anthropic's 0.815 at nearly half the cost per task [7, 8, 9].

Research & Innovation

Why it matters: benchmark behavior, scaling theory, and reasoning efficiency all advanced.

- **Claude Opus 4.8 posted a notable ARC-AGI-3 jump.** A shared result said it tripled GPT-5.5's score and reached 1.5% human efficiency [10]. ARC-specific notes said 4.8 worked at a higher abstraction level than 4.7—seeing objects instead of just pixels—and held onto hypotheses longer before resetting [11].
- **A new scaling paper argued larger models are bottlenecked by data competition.** Goodfire and collaborators traced better task learning to data-induced competition for neuron resources, using formal analysis, idealized tasks, and real pretraining [12, 13].
- **Reasoning in Memory proposed latent reasoning without visible thought tokens.** The claim: a dedicated latent workspace can preserve reasoning quality while making inference much faster by avoiding explicit reasoning-token generation [14].

Products & Launches

Why it matters: new launches kept moving agent capabilities into APIs and safer execution environments.

- **Qwen3.7-Plus** is a new multimodal agent model that unifies vision and language for GUI and CLI tasks, coding, visual reasoning, grounding, and search-augmented QA [15]. Alibaba said it is available via API on Model Studio, and shared results describing competitive text performance plus broad multimodal gains in understanding, tool use, and task execution [15, 16, 17].
- **Managed Agents in the Gemini API** lets developers spin up an agent with a single API call that can reason, write and run code, and manage files inside a hosted Linux sandbox [18].
- **LangSmith Sandboxes** are now generally available for agents that need to execute code safely, with runtime isolation, network controls, persistent state, and snapshot/restore [19, 20].

Industry Moves

Why it matters: the biggest companies kept committing more capital, infrastructure, and device surface area to AI.

- **Alphabet is raising more capital for AI buildout.** The Wall Street Journal said Alphabet plans to issue $80 billion in equity to finance AI-related capital expenditures [21]. A cited Google press statement said demand for its AI products from enterprises and consumers is exceeding available supply [22].
- **OpenAI broke ground on Stargate Michigan.** The 1GW data center uses closed-loop cooling, is expected to create thousands of union jobs, and comes with more than $40 million in free Codex credits for Michigan college, community college, and trade school students [23]. OpenAI-linked commentary framed the site as part of making AI more useful, reliable, and affordable over time [24].
- **NVIDIA and Microsoft pushed harder on local AI PCs.** Microsoft highlighted next-generation Windows PCs powered by RTX Spark with 1 petaflop of AI performance and up to 128GB unified memory [25]. NVIDIA also introduced DGX Station for Windows, with up to 748GB coherent memory and support for running models up to 1 trillion parameters locally [26, 27].

Policy & Regulation

Why it matters: policy debate is expanding from safety rules to ownership and governance.

- **Bernie Sanders said he will introduce the American AI Sovereign Wealth Fund Act.** The proposal would impose a one-time 50% tax paid in stock by OpenAI, Anthropic, and xAI; the government would receive voting shares and equal board seats, with the stock funding a public citizen-owned fund [28, 29].

Quick Takes

Why it matters: a few smaller updates still add signal on safety, open models, and benchmarks.

- OpenAI Foundation said it has more than **$130M** in initial grants underway for bio-resilience, cyber-resilience, AI model safety, and AI's impact on young people [30, 31].
- JetBrains released **Mellum2**, an open-source **12B MoE** for natural language and code with **128K** context, ultra-low-latency inference, and day-0 vLLM support [32, 33].
- Artificial Analysis launched **AA-WER Streaming** for speech-to-text agents; Cartesia Ink-2 and ElevenLabs Scribe v2 lead the accuracy-latency frontier, while Deepgram Flux is fastest [34].
- **Cosmos3-Super** is now live on fal for text-to-image and image-to-video workflows [35, 36].

Sources

1. X post by @OpenAI
2. X post by @awscloud
3. X post by @reach_vb
4. X post by @AnthropicAI
5. X post by @perplexity_ai
6. X post by @perplexity_ai
7. X post by @perplexity_ai
8. X post by @perplexity_ai
9. X post by @perplexity_ai
10. X post by @scaling01
11. X post by @GregKamradt
12. X post by @ChrisGPotts
13. X post by @AndrewLampinen
14. X post by @HochreiterSepp
15. X post by @Alibaba_Qwen
16. X post by @Alibaba_Qwen
17. X post by @Alibaba_Qwen
18. X post by @_philschmid
19. X post by @LangChain
20. X post by @hwchase17
21. X post by @WSJTech
22. X post by @jaminball
23. X post by @OpenAINewsroom
24. X post by @sk7037
25. X post by @yusuf_i_mehdi
26. X post by @nvidianewsroom
27. X post by @kimmonismus
28. X post by @TheRundownAI
29. X post by @SenSanders
30. X post by @FoundationOAI
31. X post by @woj_zaremba
32. X post by @jetbrains
33. X post by @vllm_project
34. X post by @ArtificialAnlys
35. X post by @fal
36. X post by @fal