Gemini Deep Think hits new benchmark records as OpenAI ships ultra-low-latency Codex Spark

AI News Digest

2026-02-13

Gemini Deep Think hits new benchmark records as OpenAI ships ultra-low-latency Codex Spark

By AI News Digest • February 13, 2026

Google's Gemini 3 Deep Think upgrade reports new highs on ARC-AGI-2, Humanity's Last Exam, and Codeforces, alongside expanded availability. OpenAI launches GPT-5.3-Codex-Spark for ultra-low-latency coding and begins sunsetting older ChatGPT models, while Anthropic announces a $30B round and steps up policy engagement amid a new debate on ads inside ChatGPT.

Gemini Deep Think sets new records; Codex Spark makes coding feel instant

Google upgrades Gemini 3 Deep Think (new benchmark highs + broader rollout)

Google DeepMind says it has **upgraded Gemini 3 Deep Think**, refining it with scientists and researchers to tackle "tough, real-world challenges" [1, 2]. Reported results include **84.6% on ARC-AGI-2**, **48.4% on Humanity's Last Exam (without tools)**, and **3455 Elo on Codeforces** [1, 3]. The upgraded mode is rolling out to **Google AI Ultra** subscribers in the Gemini app, with **early access API** availability for select researchers and enterprises (including a Vertex AI early access program) [4, 5].

Why it matters: This is a concrete "frontier reasoning" jump with hard numbers on the most-discussed benchmarks—and it's tied to real deployments for researchers, not just a lab demo [6, 2].

OpenAI launches GPT-5.3-Codex-Spark (ultra-low-latency coding)

OpenAI released **GPT-5.3-Codex-Spark** in **research preview**, positioning it as a way to "just build things—faster" [7]. Sam Altman highlighted **"more**

than 1000 tokens per second" and noted there are "limitations at launch" with rapid improvements planned [8]. It's rolling out to **ChatGPT Pro** users via the Codex app, CLI, and IDE extension (with Codex users on the Pro plan called out specifically) [9, 8].

Why it matters: The product framing here is speed as a first-class feature for software creation—explicitly pushing token-throughput as UX, not just a benchmark stat [8, 10].

Real-world signal: building complex apps with Codex 5.3

Martin Casado described **Codex 5.3** as "another level of sophistication," citing progress on a distributed world-building app (permissions/policy, portals, deployment) and noting the difficulty of shared mutable state with optimistic client-side updates [11, 12]. He emphasized it still takes iteration and testing to tune performance, but called the result "really impressive" [12].

Why it matters: Agentic coding is increasingly being judged on whether it can handle *systems* problems (state, permissions, deployment), not just generate clean functions [12].

OpenAI to deprecate multiple "legacy" ChatGPT models

OpenAI says "legacy models" (including **GPT-5**, **GPT-4o**, **GPT-4.1**, **GPT-4.1 mini**, and **OpenAI o4-mini**) will be **deprecated in ChatGPT** at 10am PT the next day [13]. Separate commentary noted the rapid release cadence (GPT-5.1, GPT-5.2, and GPT-5.3 variants arriving within months/weeks) and @swyx pointed to GPU tradeoffs and limited research access as a frustration for serving older models [14, 15].

Why it matters: Model churn is becoming an operational constraint for teams who build workflows around specific model behaviors—and it's explicitly being linked to GPU scarcity/tradeoffs [15, 13].

Anthropic: massive financing + policy posture

Anthropic raises $30B at a $380B post-money valuation; cites $14B run-rate

Anthropic announced it has raised **$30B** in funding at a **$380B post-money valuation** [16]. The company said the investment will support research, product innovation, infrastructure expansion, and making **Claude** available "everywhere" customers are [16]. In a separate post, Anthropic stated **$14B run-rate revenue**, claiming it has grown **10x** in each of the past three years, driven by enterprise and developer adoption [17].

Why it matters: The scale of capital and claimed revenue growth underscore how quickly frontier labs are turning into infrastructure-heavy businesses—and how directly product availability is tied to compute expansion [16, 17].

Anthropic commits $20M to a new bipartisan AI policy org

Anthropic announced a **$20M** contribution to **Public First Action**, describing it as a new bipartisan organization aimed at mobilizing people and politicians around AI policy [18].

Why it matters: Labs are increasingly pairing technical scaling with explicit political organization-building, reflecting a belief that the "window to get policy right is closing" [18].

Monetization shift: ads inside ChatGPT (and the trust debate)

OpenAI begins testing ads in ChatGPT for free + Go tiers

A video summary reports OpenAI is **testing ads in ChatGPT** for logged-in adult users on the **free and Go** subscription tiers, with higher paid plans not receiving ads [19]. The same summary described ads as clearly labeled "sponsor" in a gray box at the bottom of responses (outside the main answer), alongside controls to turn off ad personalization, disable memory access for ads, or delete ad data [19].

Why it matters: This is a meaningful product/business-model change: ads introduce new incentives, and OpenAI is explicitly trying to separate ad placement from answer generation to preserve trust [19].

Competing philosophies: "space to think" vs. ad-supported access

In the same discussion, Sam Altman is quoted expressing an aesthetic dislike of ads and describing them as a "last resort," while also arguing subscription models help users trust answers aren't advertiser-influenced [19]. Anthropic's stance is presented as: ads inside a Claude conversation would compromise it as a "clear space to think and work," and would create incentives to optimize for engagement rather than genuine helpfulness [19].

Why it matters: The LLM UI is becoming a monetization battleground, and both camps are explicitly arguing about incentive gradients—not just UX preferences [19].

Critiques focus on ad targeting and long-run incentive drift

Ben Thompson criticized the initial approach as "banner ads" tied to the conversation context, arguing this can raise user suspicion about whether the answer

is influenced, and advocating more "Meta-style" user understanding for less conflicted targeting [20]. Separately, a former OpenAI researcher, Zoe Hitzig, warned that ads built on intimate chatbot conversations could enable manipulation and that the incentive system may pressure companies to erode early principles over time [19].

Why it matters: The hard problem isn't just ad insertion—it's what business incentives do to product boundaries over multiple iterations [19, 20].

Compute economics: token cost continues to fall (hardware + inference stack)

NVIDIA: up to 10x lower cost per token with Blackwell; Rubin platform next

NVIDIA highlighted that inference providers (Baseten, DeepInfra, Fireworks AI, Together AI) are seeing up to **10x cost-per-token reduction** using open source models on **NVIDIA Blackwell** versus Hopper [21]. It also pointed to GB200 NVL72 delivering a **10x reduction** in cost per token for reasoning MoE models versus Hopper, and teased the **Rubin** platform as targeting **10x performance and 10x lower token cost** over Blackwell [21]. NVIDIA additionally cited MIT research claiming inference costs for frontier-level performance can drop by up to **10x annually** due to infrastructure and algorithmic efficiencies [21].

Why it matters: "Tokenomics" is increasingly the strategic lever—shaping which products can be offered broadly and which agent workflows are economically viable [21].

DGX Spark: data-center-class AI on the desktop (universities + sensitive data)

NVIDIA described DGX Spark as a compact desktop system supporting models up to **200B parameters**, aimed at keeping sensitive workloads on premises while shortening iteration loops [22]. Examples included Stanford prototyping workflows locally (reporting ~**80 tokens/sec** on a 120B-parameter model at MXFP4 via Ollama) and NYU running agentic report evaluation for radiology without sending medical imaging data to the cloud [22].

Why it matters: Local "lab bench" compute changes who can iterate quickly (and privately) on large-model workflows—especially in regulated domains like healthcare [22].

Research & safety: steering models without breaking usefulness

Anthropic: "assistant axis" + activation capping to reduce personality drift

A research summary described "personality drift" in assistants—where models can be steered away from a helpful assistant persona (including via emotional or philosophical topics) and become unstable or jailbreak-prone [23]. The approach identifies an "assistant axis" (a geometric direction associated with the assistant persona) and applies **activation capping** that nudges the model back only when helpfulness drops below a threshold; reported results cut jailbreak rate roughly in half with little performance loss, and suggested the axis generalizes across different models [23].

Anthropic Found Why AIs Go Insane (3:52)

Why it matters: This frames jailbreak resistance as a targeted representational-control problem—potentially offering a more granular safety knob than broad refusals or heavy-handed fine-tuning [23].

Jeff Dean (Google): distillation, energy bottlenecks, and "trillions of tokens" via retrieval

In a conversation with Latent Space, Jeff Dean emphasized a strategy of pairing frontier models with **lower-latency "Flash" models** produced via **dis-**

tillation, enabling widespread deployment while keeping frontier capability for deep reasoning [24, 25]. He also framed energy as a core constraint: moving data across a chip can cost far more (in picojoules) than a multiply, which helps explain why batching matters [24]. On scaling context, he argued that attending to trillions of tokens won't come from simply scaling quadratic attention, but from systems that create the *illusion* of that scale via staged retrieval and refinement [24, 25].

Why it matters: This is a "systems" roadmap—distillation + retrieval + hardware co-design—aimed at making agentic workloads affordable and interactive, not just smarter in isolation [24].

Benchmarks, AGI narratives, and labor signals

Chollet: ARC is a research tool (not AGI proof); ARC-4 planned for early 2027

François Chollet argued ARC was never meant as proof of AGI and remains a tool to steer research toward "fluid intelligence," noting that base LLMs (no test-time adaptation) still perform "abysmally low" despite massive scaleups [26]. He said ARC-4 is "in the works" for **early 2027**, with ARC-5 also planned, aiming to keep producing tests that humans can do and AI can't [27].

Why it matters: As benchmark scores accelerate, Chollet is explicitly pushing the community toward moving-target evaluation—and away from declaring victory based on a single saturated test [27, 28].

Labor canary: call centers, not "tech Twitter vibes"

Chollet proposed call centers as a canary for AI-caused job loss, citing a projection of ~**2.75M US call center jobs in 2026** and suggesting a **-50% employment drop** would indicate broader disruption [29]. He also said he generally doesn't expect AI-caused mass unemployment in the next five years, with a plausible scenario being changing job nature and higher throughput with stable or slightly lower employment [30].

Why it matters: This is a concrete, measurable lens on "AI job displacement" that avoids both hype and denial—by anchoring on an industry where automation pressure is intuitive and trackable [29, 30].

Quick hits

- **Alibaba open-sources Zvec**, an embedded vector database pitched as "SQLite-like" for on-device RAG (repo link included) [31].

- **Neuphonic releases NeuTTS Nano multilingual** (German/French/Spanish), 120M-parameter on-device TTS models with real-time CPU inference via llama.cpp and ~3s zero-shot voice cloning [32].
- **GLM-5 weights are out**, with architecture notes highlighting more experts plus multi-head latent attention and DeepSeek Sparse Attention [33].
- **New A2A (Agent2Agent) course** from deeplearning.ai (with Google Cloud and IBM Research) aims to standardize agent discovery/communication across frameworks [34].

Sources

1. X post by @sundarpichai
2. X post by @GoogleDeepMind
3. X post by @demishassabis
4. X post by @sundarpichai
5. X post by @GoogleDeepMind
6. X post by @demishassabis
7. X post by @OpenAI
8. X post by @sama
9. X post by @OpenAI
10. X post by @gdb
11. X post by @martin_casado
12. X post by @martin_casado
13. X post by @OpenAINewsroom
14. X post by @iScienceLuvr
15. X post by @swyx
16. X post by @AnthropicAI
17. X post by @AnthropicAI
18. X post by @AnthropicAI
19. ChatGPT Ads Are Here. It Gets Worse.
20. Ben Thompson from Stratechery on AI ads, the end of SaaS, and the future of media
21. Leading Inference Providers Cut AI Costs by up to 10x With Open Source Models on NVIDIA Blackwell
22. NVIDIA DGX Spark Powers Big Projects in Higher Education
23. Anthropic Found Why AIs Go Insane
24. Owning the AI Pareto Frontier — Jeff Dean
25. Owning the AI Pareto Frontier — Jeff Dean
26. X post by @fchollet
27. X post by @fchollet
28. X post by @fchollet
29. X post by @fchollet
30. X post by @fchollet

31. r/LocalLLM post by u/techlatest_net
32. r/LocalLLM post by u/TeamNeuphonic
33. X post by @rasbt
34. X post by @AndrewYNg