Codex-Spark hits 1,000 tok/s as Gemini Deep Think, open coding models, and mega-funding reshape the landscape

AI High Signal Digest

2026-02-13

Codex-Spark hits 1,000 tok/s as Gemini Deep Think, open coding models, and mega-funding reshape the landscape

By AI High Signal Digest • February 13, 2026

OpenAI launches a Cerebras-powered Codex-Spark for real-time coding speed, Google rolls out a major Gemini 3 Deep Think upgrade with strong ARC/HLE/Codeforces results, and open-source coding models (MiniMax M2.5, GLM-5) keep tightening the cost/performance race. Also: Anthropic's $30B raise, new research on long-horizon agents and theorem proving, and a key US legal ruling on AI chats and privilege.

Top Stories

1) OpenAI ships GPT-5.3-Codex-Spark: ultra-fast, low-latency coding in Pro (Cerebras-powered)

Why it matters: Latency is becoming a first-class product differentiator for coding agents—Spark is positioned as a dedicated "fast tier" that complements GPU serving for real-time development workflows. [1, 2]

- **Launch + availability:** OpenAI introduced GPT-5.3-Codex-Spark as a research preview for **ChatGPT Pro** users in the **Codex app, Codex CLI, and IDE extension**. [1, 3]
- **Performance emphasis:** The model is framed as a major speed upgrade for coding and the "first in a family of ultra-fast models" for real-time development. [4]
- **Partnership + infra:** Spark is described as the **first milestone** in OpenAI's partnership with **Cerebras**, providing a faster tier for workloads where low latency is critical. [2]

- **Current limitations + roadmap:** Spark is **text-only** with a **128k context window**, with plans to add larger models, longer context, and multimodal inputs as deployment learns accumulate. [5]
- **Codex-wide speedups coming:** OpenAI says Codex will continue to get faster via improved response streaming, session initialization, and inference stack rewrites rolling out across all Codex models. [6]

2) Google rolls out a major Gemini 3 Deep Think upgrade (benchmarks + practical science/engineering use)

Why it matters: Deep Think is being positioned not just as "better reasoning," but as a specialized mode intended to move R&D work forward (with early deployments in research and engineering workflows). [7, 8]

- **Headline benchmarks:** Google reports **84.6% on ARC-AGI-2** (verified by ARC Prize Foundation), **48.4% on Humanity's Last Exam (without tools)**, and **3455 Elo on Codeforces**. [9]
- **Availability:** Deep Think is available now in the **Gemini app** for **Google AI Ultra** subscribers, and via the **Gemini API** to select researchers/engineers/enterprises through early access. [10, 11]
- **Practical applications highlighted:**
  - Sketch → **3D-printable file** generation by analyzing a drawing, building the shape, and generating an output file. [8]
  - Early testers report spotting subtle flaws in technical math papers and optimizing semiconductor crystal growth (including a recipe for thin films larger than **100 m**). [12, 13]

3) Open-source coding models tighten the gap: MiniMax M2.5 and GLM-5 push cost/perf and agentic workflows

Why it matters: Multiple releases point to a fast-moving open model layer that's increasingly competitive for coding/agent tasks—often paired with integrations that make switching costs lower. [14, 15, 16]

- **MiniMax M2.5 positioning:** MiniMax describes **M2.5** as an open-source frontier model for "real-world productivity," with **SWE-Bench Verified 80.2%**, **BrowseComp 76.3%**, and **BFCL 76.8%** (agentic tool-calling). [14]
- **Cost/speed claims in developer tooling:** Cline reports M2.5 at **100 tokens/s** and **$0.06/M blended cost** (with caching), alongside benchmark comparisons vs Opus 4.6. [17, 18, 19]
- **Access expands quickly:** Ollama partnered with MiniMax for **free usage** of M2.5 for a couple of days (cloud model), with CLI "launch" integrations for multiple coding tools. [16]
- **GLM-5 scale + intent:** Zai_org positions GLM-5 for complex systems engineering and long-horizon agentic tasks, scaling to **744B params (40B active)** and **28.5T** pretraining tokens. [15]

4) Anthropic announces $30B Series G at $380B post-money, citing $14B run-rate revenue

Why it matters: The funding round is framed explicitly as fuel for research, product, and infrastructure scale—plus wider distribution of Claude. [20]

- **Financing + valuation:** Anthropic says it raised **$30B** at a **$380B post-money valuation**. [20]
- **Business metrics disclosed:** Anthropic reports **$14B run-rate revenue**, with "over 10x" growth in each of the past three years. [21]
- **Use of proceeds:** The company says the investment will deepen research, innovate in products, and expand infrastructure as it makes **Claude** available "everywhere our customers are." [20]

5) ARC-AGI-2 heats up: Gemini 3 Deep Think at 84.6% and an agent-based 85.28% SOTA; Chollet outlines ARC's benchmark roadmap

Why it matters: ARC continues to function as a focal point for "reasoning + adaptation" claims, while its creator reiterates it's a research tool—not a proof-of-AGI milestone. [22, 9, 23]

- **Scores cited:** Gemini 3 Deep Think is reported at **84.6%** on ARC-AGI-2. [9]
- **New SOTA claim:** Agentica reports **85.28%** using an agent (~350 lines) that writes and runs code. [23]
- **Benchmark roadmap:** François Chollet says **ARC-4** is in the works (early 2027), **ARC-5** is planned, and a final ARC may be 6–7, aiming to keep making benchmarks until no "humans can do and AI can't" tasks remain; he also states "AGI ~2030." [24]

Research & Innovation

Smaller models + test-time compute: theorem proving and reasoning infrastructure

Why it matters: Several updates emphasize using scaffolds, RL, and test-time compute to make smaller or specialized models punch above their weight.

- **QED-Nano (4B) theorem prover:** Presented as the smallest theorem-proving model to date, matching much larger models on **IMO-ProofBench** in natural language (no Lean/external tools). It's post-trained with **RL using rubrics as rewards** and open-sourced on Hugging Face. [25, 26]
- **Agentica's recursive harness:** Commentary describes the ARC-AGI-2 result as using a deeply recursive RLM-style loop with a **stateful REPL** to manage long horizons beyond a single context window. [27, 28]

Data contamination + evaluation tooling: searching trillion-token corpora

Why it matters: As benchmark contamination concerns grow, faster "soft match" search for near-duplicates becomes a practical requirement for trustworthy evals.

- **SoftMatcha 2 (Sakana AI + collaborators):** A "fast and soft pattern matcher" that searches **trillion-scale corpora in under 0.3 seconds**, handling semantic variations (substitution/insertion/deletion) and identifying potential benchmark contamination missed by exact match. [29]

Long-context efficiency + architecture work

Why it matters: Attention and serving efficiency are repeatedly highlighted as the constraint for long-horizon agents and large-context models.

- **HySparse (Xiaomi MiMo):** A hybrid sparse attention architecture interleaving full attention with multiple sparse layers that derive token selection and KV caches from the preceding full layer. [30]
- **Hybrid linear attention claims:** A post describes a GQA overhaul mixing Multi-head Linear Attention (MLA) with Lightning Linear, claiming **3x+ throughput** at >32K context and **>10x** reduced memory access overhead. [31]

Products & Launches

Developer productivity: long-running agents and parallelism

Why it matters: New tooling is shifting from "single chat sessions" to sustained, parallel, and observable agent workflows.

- **Cursor long-running agents:** Cursor reports long-running coding agents peaking at **1,000+ commits/hour** across hundreds of agents in a week-long run, now available to Ultra/Teams/Enterprise. [32, 33]
- **Cline v3.58.0 subagents:** Adds native subagents that run parallel subtasks with their own contexts (experimental in VSCode/CLI), plus GLM-5 support. [34, 35]

Search and realtime interaction building blocks

Why it matters: Agents doing multi-step work become limited by tool latency—especially web search.

- **Exa Instant:** Launched as a sub-200ms search engine, described as custom-built for realtime AI products like chat and voice. [36]

Speech + translation

Why it matters: Open-source realtime speech translation expands what can run locally or be integrated without closed, hosted constraints.

- **Kyutai Hibiki-Zero:** An open-source real-time multilingual speech translation model (French/Spanish/Portuguese/German → English), emphasizing low latency, high audio quality, and voice transfer. [37]

Industry Moves

New companies and funding

Why it matters: Capital continues to cluster around "new interfaces" (simulation, agents) and infrastructure that turns models into usable systems.

- **Simile raises $100M:** Positioned around simulating human behavior; funding cited from Index, Hanabi, A* BCV, and angels including Karpathy, Fei-Fei Li, Adam D'Angelo, and others. [38]
- **Meta infrastructure buildout:** Meta is breaking ground on a **1GW** data center in Lebanon, Indiana, described as over **$10B** in infrastructure investment. [39, 40]

Data and ecosystem partnerships

Why it matters: Data access (and who pays for it) is becoming a key constraint—and business model—for model development.

- **Wikimedia high-speed API program:** Wikimedia partnered with AI firms (including Amazon, Meta, Microsoft, Mistral AI, Perplexity) to provide high-speed API access to Wikipedia and related datasets, aiming to support developers while reducing infrastructure strain from crawlers. [41]

Policy & Regulation

Legal risk: AI chats and privilege

Why it matters: Courts are now directly addressing whether AI-generated materials are privileged—raising immediate compliance and workflow questions for legal teams.

- **SDNY ruling (Judge Jed Rakoff):** 31 documents generated using an AI tool (Claude) and later shared with defense attorneys were ruled **not protected** by attorney-client privilege or work product doctrine. Reasons cited include that AI is not an attorney and that the provider's terms disclaimed an attorney-client relationship; forwarding documents later does not retroactively make them privileged. [42]

Platform churn: model deprecations

Why it matters: Rapid iteration cycles increasingly force product teams to plan for upgrades, regressions, and continuity.

- **OpenAI deprecations in ChatGPT:** OpenAI says legacy models (GPT-5, GPT-4o, GPT-4.1, GPT-4.1 mini, o4-mini) will be deprecated in ChatGPT at **10am PT** the next day. [43]

Political engagement on AI policy

Why it matters: Leading labs are funding policy engagement while warning that the policy window is tightening.

- **Anthropic donation:** Anthropic says AI is being adopted faster than any technology in history and the policy window is closing; it is contributing **$20M** to Public First Action, described as a new bipartisan organization. [44]

Quick Takes

Why it matters: These are smaller updates that may still become default tools, constraints, or reference points.

- **Karpathy's microGPT:** A minimal "train + inference GPT" implementation in **243 lines** of dependency-free Python, later simplified to **200 lines** by returning local gradients per op and letting `backward()` chain them. [45, 46]
- **Soft shift in software workflows:** A firm says it's rethinking a banker take-home test because their technical cofounder (no IB experience) can now "one-shot" it using **Opus 4.6**. [47]
- **ColGrep / LateOn-Code:** Introduced as lightweight local code retrieval that "wins 70% vs grep" and uses "15.7% fewer tokens," with Claude Code integration mentioned. [48, 49]
- **UN scientific panel on AI:** The UN General Assembly appointed **40 experts** to an Independent International Scientific Panel on AI, described as providing evidence-based scientific assessments to inform international deliberations. [50]

Sources

1. X post by @OpenAIDevs
2. X post by @OpenAIDevs
3. X post by @OpenAI
4. X post by @TheRundownAI

5. X post by @OpenAIDevs
6. X post by @OpenAIDevs
7. X post by @Google
8. X post by @GeminiApp
9. X post by @Google
10. X post by @Google
11. X post by @_philschmid
12. X post by @Google
13. X post by @Google
14. X post by @MiniMax_AI
15. X post by @Zai_org
16. X post by @ollama
17. X post by @cline
18. X post by @cline
19. X post by @cline
20. X post by @AnthropicAI
21. X post by @AnthropicAI
22. X post by @fchollet
23. X post by @agenticasdk
24. X post by @fchollet
25. X post by @_lewtun
26. X post by @_lewtun
27. X post by @lateinteraction
28. X post by @lateinteraction
29. X post by @SakanaAILabs
30. X post by @XiaomiMiMo
31. X post by @AntLingAGI
32. X post by @cursor_ai
33. X post by @cursor_ai
34. X post by @cline
35. X post by @cline
36. X post by @ExaAILabs
37. X post by @kyutai_labs
38. X post by @joon_s_pk
39. X post by @fb_engineering
40. X post by @kimmonismus
41. X post by @DeepLearningAI
42. X post by @mpeltz
43. X post by @OpenAINewsroom
44. X post by @AnthropicAI
45. X post by @karpathy
46. X post by @karpathy
47. X post by @leveredvlad
48. X post by @antoine_chaffin
49. X post by @tonywu_71
50. X post by @ODET_UN