Gemini Live Goes Global as Codex Plugins and Open Audio Models Expand AI Workflows

AI High Signal Digest

2026-03-27

Gemini Live Goes Global as Codex Plugins and Open Audio Models Expand AI Workflows

By AI High Signal Digest • March 27, 2026

Google pushed Gemini 3.1 Flash Live across Search, Gemini, and developer channels, while OpenAI broadened Codex with open-source plugins. The brief also covers open audio models, new research systems, industry partnerships, and the latest safety and compliance signals.

Top Stories

Why it matters: The biggest developments today pushed AI deeper into real-time interaction, connected workflow automation, open audio infrastructure, and operational safety.

Google turned Gemini 3.1 Flash Live into a broad real-time platform

Google rolled out Gemini 3.1 Flash Live across Gemini Live, Search Live, Google AI Studio, and Google Cloud, positioning it as a production-ready realtime model for voice and vision agents [1, 2, 3]. Google said it improved quality, reliability, latency, conversation memory, and instruction-following, while Search Live is now available in more than 200 countries and territories with multilingual support [4, 5, 6, 7, 8]. Independent benchmarking also showed a clear speed/quality tradeoff: 95.9% on Big Bench Audio at the high thinking setting with 2.98s time-to-first-audio, versus 70.5% and 0.96s on minimal thinking [9, 10].

Impact: Google is not just shipping a model. It is distributing one live audio stack across consumer search, the Gemini app, developer tooling, and enterprise channels.

OpenAI expanded Codex from coding assistant to connected work surface

OpenAI is rolling out plugins in Codex so it can work with tools like Slack, Figma, Notion, Gmail, and Google Drive, including Docs, Sheets, and Slides [11, 12]. OpenAI said plugins extend Codex into planning, research, coordination, and post-coding workflows; they are available in the Codex app, CLI, and IDE extensions [12, 13]. OpenAI also said users will be able to build and share their own plugins, and that today's plugins are open source [14, 15].

Impact: This moves Codex closer to a general work agent that operates inside the tools teams already use, not just inside a code editor.

Open speech models got stronger on both input and output

Cohere launched Cohere Transcribe, its first audio model, under Apache 2.0. The company said it is state of the art in open-source speech recognition, ranks #1 on the Open ASR leaderboard, supports 14 languages, and reached 5.42% English word error rate in human evaluation [16, 17, 18]. Mistral released Voxtral TTS as an open-weight text-to-speech model with low latency, emotional expressiveness, and support for 9 languages; the company published weights and a technical report [19, 20].

Impact: The open audio stack is improving at both ends: transcription on the way in, expressive speech generation on the way out.

Safety work became more operational

Google DeepMind published new research on harmful manipulation based on studies with more than 10,000 people, finding high influence in finance but lower influence in health where existing guardrails blocked false medical advice [21, 22]. Separately, METR said it spent three weeks red-teaming Anthropic's internal monitoring and security systems, found several new vulnerabilities, and produced artifacts to improve future monitoring, while saying none of the findings severely undermined major claims in Anthropic's sabotage risk report [23, 24, 25, 26].

Impact: Frontier labs are moving from abstract safety principles toward live testing, measurement, and third-party scrutiny.

Research & Innovation

Why it matters: The strongest technical work today focused on specialized systems: brain modeling, search agents, self-modifying agents, and automated security research.

- **Meta FAIR's TRIBE v2:** Meta introduced a foundation model trained on 500+ hours of fMRI recordings from 700+ people to predict how the

human brain responds to sights and sounds. Meta says it supports zero-shot predictions for new subjects, languages, and tasks, improves 2-3x over prior methods on movies and audiobooks, and is being released with code, paper, and demo [27, 28].

- **Chroma Context-1:** Chroma launched a 20B search agent it says pushes the pareto frontier of agentic search and is an order of magnitude faster and cheaper [29]. The model was trained with SFT + RL on 8,000+ synthetic multi-hop tasks across web, SEC filings, patent law, and email, and Chroma open-sourced both the weights and the task-generation codebase [30, 31].
- **Hyperagents and DGM-H:** Hyperagents are presented as self-modifying AI systems that can rewrite both the task-solving and self-improvement parts of the agent. In the DGM-H setup, reported performance improved across coding, paper review, and robotics, with gains accumulating across runs [32].
- **Autoresearch for jailbreaking:** A new paper used Claude Code in an autoresearch loop to discover novel jailbreaking algorithms that reportedly beat 30+ existing GCG-like attacks and generalized better to unseen models than prior work. The authors said this suggests some incremental safety and security research can now be automated [33, 34].

Products & Launches

Why it matters: Product launches kept reducing friction around memory, provisioning, orchestration, and domain-specific deployment.

- **Gemini import tools:** Gemini is rolling out memory import and chat history import, letting users bring preferences and prior chats from other AI apps into Gemini on desktop, with mobile coming later [35, 36, 37, 38].
- **Stripe Projects:** Stripe launched Projects in developer preview so agents can provision third-party services from the CLI. Stripe's example command creates a PostHog account, gets an API key, and sets up billing without leaving the terminal [39].
- **Cline Kanban:** Cline launched a free, open-source standalone app for CLI-agnostic multi-agent orchestration, compatible with Claude, Codex, and Cline. Tasks run in worktrees, can be linked into dependency chains, and include built-in git views [40, 41, 42].
- **Glass Developer API:** Glass Health made its Developer API self-serve inside its web app. The API supports clinical question answering, differential diagnosis, treatment planning, and documentation, with structured JSON, in-text citations, and HIPAA compliance with BAA [43, 44, 45, 46].
- **Ollama in VS Code:** Visual Studio Code can now use local or cloud Ollama models through GitHub Copilot if Ollama is installed [47, 48].

Industry Moves

Why it matters: Partnerships and financing are showing where companies think AI value will concentrate: manufacturing, multi-agent systems, and new revenue lines.

- **Sakana x Mitsubishi Electric:** Sakana AI announced a strategic partnership and investment from Mitsubishi Electric. The two companies said they will combine Mitsubishi's manufacturing data and domain knowledge with Sakana's AI systems, and Sakana framed manufacturing and physical AI as its third major pillar after finance and defense [49, 50].
- **OpenAI backs Isara:** Isara raised $94 million at a $650 million valuation. Posts describing the company say it coordinates thousands of AI agents to solve complex problems, used roughly 2,000 agents to forecast gold prices, and plans to sell predictive modeling tools to finance firms first [51, 52].
- **OpenAI ads pilot:** Reporting shared on X said OpenAI's ads pilot surpassed $100 million in ARR six weeks after launch, expanded to more than 600 advertisers, and plans self-serve advertiser access in April [53].
- **Anthropic IPO talk:** A post linking *The Information* said Anthropic has discussed going public as soon as the fourth quarter and that bankers pitching the company think an IPO could raise more than $60 billion [54, 55].

Policy & Regulation

Why it matters: The clearest policy signals today were around safety governance, privacy, and compliance rather than formal rulemaking.

- **Third-party red-teaming:** METR said Anthropic gave an external researcher substantial access to internal monitoring and security systems for a three-week exercise, and METR said some vulnerabilities found during the exercise have already been patched [56, 24].

  > "This kind of adversarial testing by external researchers is valuable for discovering vulnerabilities, as well as for developing best practices for embedding third party evaluators inside frontier AI companies." [57]

- **Manipulation measurement:** Google DeepMind said it built a first-of-its-kind empirically validated toolkit to measure real-world AI manipulation, based on nine studies involving more than 10,000 participants across three countries [58, 59, 60].
- **OpenAI put an erotic chatbot plan on hold:** Posts citing the *Financial Times* said OpenAI indefinitely shelved a planned adult-mode chatbot amid concerns about risks to minors, unhealthy emotional attachments, and the difficulty of filtering illegal material while generating explicit content [61, 62].
- **Encrypted inference:** Chutes said its end-to-end encrypted AI inference

keeps user data encrypted until it reaches a GPU inside a trusted execution environment, and uses ML-KEM-768 with fresh ephemeral keypairs for forward secrecy and post-quantum resistance [63].

Quick Takes

Why it matters: These were smaller updates, but they point to where tooling, creator software, and AI operations are moving next.

- **Moondream Photon** claims 46ms end-to-end VLM inference and 60+ fps on a single H100, from edge devices to servers [64].
- **Runway's Multi-Shot App** turns a prompt or image into a scene with dialogue, sound effects, cuts, pacing, and cinematic framing [65].
- **Google's Lyria 3 Pro** can generate music tracks up to three minutes with structure-aware sections such as intros, verses, choruses, and bridges [66].
- **Stanford NLP's sycophancy study** reported that sycophantic LLMs can make users more self-centered, increase confidence that they are right, and reduce willingness to repair interpersonal conflicts, even while users prefer and trust those systems more [67, 68].
- **Anthropic tightened peak-hour Claude limits**, while OpenAI responded by offering temporary 2x Codex rate limits across ChatGPT subscriptions [69, 70].
- **AxiomMath open-sourced Axplorer**, a tool for searching interesting or optimal mathematical objects under constraints; the company said it matched state of the art on several combinatorics problems with much less compute and time [71, 72].

Sources

1. X post by @Google
2. X post by @kimmonismus
3. X post by @NoamShazeer
4. X post by @OfficialLoganK
5. X post by @Google
6. X post by @Google
7. X post by @Google
8. X post by @Google
9. X post by @ArtificialAnlys
10. X post by @ArtificialAnlys
11. X post by @OpenAIDevs
12. X post by @OpenAIDevs
13. X post by @OpenAIDevs
14. X post by @OpenAIDevs
15. X post by @reach_vb

16. X post by @cohere
17. X post by @aidangomez
18. X post by @cohere
19. X post by @MistralAI
20. X post by @GuillaumeLample
21. X post by @GoogleDeepMind
22. X post by @GoogleDeepMind
23. X post by @METR_Evals
24. X post by @METR_Evals
25. X post by @METR_Evals
26. X post by @METR_Evals
27. X post by @AIatMeta
28. X post by @AIatMeta
29. X post by @trychroma
30. X post by @trychroma
31. X post by @trychroma
32. X post by @TheTuringPost
33. X post by @kotekjedi_ml
34. X post by @jonasgeiping
35. X post by @GeminiApp
36. X post by @GeminiApp
37. X post by @GeminiApp
38. X post by @GeminiApp
39. X post by @patrickc
40. X post by @cline
41. X post by @cline
42. X post by @cline
43. X post by @GlassHealthHQ
44. X post by @GlassHealthHQ
45. X post by @GlassHealthHQ
46. X post by @GlassHealthHQ
47. X post by @ollama
48. X post by @ollama
49. X post by @hardmaru
50. X post by @SakanaAILabs
51. X post by @kimmonismus
52. X post by @WSJ
53. X post by @steph_palazzolo
54. X post by @srimuppidi
55. X post by @srimuppidi
56. X post by @METR_Evals
57. X post by @METR_Evals
58. X post by @GoogleDeepMind
59. X post by @_philschmid
60. X post by @_philschmid
61. X post by @kimmonismus

62. X post by @FT
63. X post by @chutes_ai
64. X post by @moondreamai
65. X post by @runwayml
66. X post by @dl_weekly
67. X post by @chengmyra1
68. X post by @stanfordnlp
69. X post by @trq212
70. X post by @reach_vb
71. X post by @axiommathai
72. X post by @TheTuringPost