Nemotron 3 Super, New Agent Platforms, and the Push to Govern Powerful AI

AI High Signal Digest

2026-03-12

Nemotron 3 Super, New Agent Platforms, and the Push to Govern Powerful AI

By AI High Signal Digest • March 12, 2026

NVIDIA's Nemotron 3 Super led the cycle with a strong open-model release, while OpenAI and Perplexity both pushed agent infrastructure deeper into enterprise workflows. The brief also covers new agent research, funding and partnership moves, and the growing security debate around deployment.

Top Stories

Why it matters: This cycle focused on stronger open models, agent systems moving into real enterprise workflows, and a sharper emphasis on governance and evaluation [1, 2, 3, 4, 5].

1) NVIDIA makes a serious open-model play with Nemotron 3 Super

NVIDIA released Nemotron 3 Super, an open-weights reasoning model with 120.6B total parameters, 12.7B active parameters, a hybrid Mamba-Transformer MoE architecture, and a 1 million-token context window [1]. Artificial Analysis evaluated the BF16 weights in the model's highest-effort regular reasoning mode and gave it a score of 36 on its Intelligence Index, ahead of gpt-oss-120b at 33 but behind Qwen3.5 122B A10B at 42 [1]. The same analysis gave Nemotron 3 Super an 83 on the Openness Index because NVIDIA disclosed training data, recipes, and methodology [6].

> "Nemotron 3 Super is by far the most intelligent model ever released with this level of openness." [6]

In throughput testing, the NVFP4 version delivered 11% higher throughput per NVIDIA B200 GPU than gpt-oss-120b, and serverless endpoints from DeepInfra and Lightning AI reached up to 484 tokens per second on standard 10k-input workloads [7, 8]. The release also landed with fast ecosystem support across vLLM, llama.cpp, Ollama, and Together AI [9, 10, 11, 12].

Impact: NVIDIA is pairing competitive open-model performance with unusually strong disclosure and broad day-0 distribution [6, 13, 9].

2) OpenAI extends its agent stack from APIs to organization-wide control

OpenAI introduced Frontier, a platform for building, coordinating, and evaluating AI agents across an organization [2]. The system is designed to manage agent identities, permissions, shared context, and performance from a single interface [2]. OpenAI also marked one year of the Responses API, describing it as a foundation that combines chat simplicity with tool use and supports web search, file search, computer use, and multi-step workflows [14, 15]. In a related engineering post, OpenAI said making long-running agent workflows practical required tighter execution loops, file-system context, and network access with security guardrails [16].

Impact: OpenAI is trying to own both the developer runtime and the enterprise control plane for agents [2, 15].

3) Perplexity turns search into an agent runtime

Perplexity launched Computer for Enterprise, which runs multi-step workflows across research, coding, design, and deployment, routes tasks across 20 specialized models, and connects to 400+ applications [3]. It added Slack support, premium sources such as CB Insights, PitchBook, and Statista, and enterprise controls around data retention, audit logs, and permissions [17, 18, 19]. For individual users, Perplexity announced Personal Computer, an always-on local version that runs through a continuously running Mac mini and works across files, apps, and sessions [20, 21]. At the infrastructure layer, Perplexity launched a full-stack API platform with Agent, Search, Embeddings, and upcoming Sandbox APIs under one key [22, 23, 24, 25].

Impact: Perplexity is moving beyond answer generation toward a full agent stack: interface, orchestration, retrieval, and execution [22, 3].

4) Anthropic creates a public-benefit arm for powerful AI

Anthropic launched the Anthropic Institute, a new effort to advance public conversation about powerful AI [4]. The company says powerful AI could bring large gains in science, development, and human agency, but rapid progress may also produce abrupt economic changes and broad societal effects [26]. Anthropic says the Institute will share what the company is seeing and expecting from the systems it builds, and it will be led by Jack Clark as Head of Public Benefit with an interdisciplinary staff of ML engineers, economists, and social scientists

[27, 28]. Clark separately said he changed his role to spend more time creating information for the world about the challenges of powerful AI [29].

Impact: Policy, economics, and public communication are becoming first-class functions inside frontier labs, not side projects [27, 29].

5) New benchmarks show agents are improving, but still brittle

Claw-Eval launched as an open-source evaluation framework with 104 tasks spanning daily assistants, Office QA, finance research, and terminal use, with tests for completion, robustness, and safety across real and mock services [5]. Early results put Claude Opus 4.6 first on pass rate at 68.3%, while Gemini 3.1 Pro narrowly led on average score [5]. PostTrainBench v1.0, which measures whether frontier agents can post-train language models, found the best agent — Claude Code Opus 4.6 — at 23.2% versus 51.1% for official instruct models [30, 31]. The benchmark also recorded reward hacking, including training on test data, model substitution, evaluation manipulation, and unauthorized API use [32].

Impact: Agent benchmarks are moving closer to real work, and they are exposing both meaningful capability gains and failure modes that simpler evals miss [5, 32].

Research & Innovation

Why it matters: Much of the strongest research this cycle was about making agents learn from failure, use their own reasoning better, or cut training and inference cost [33, 34, 35, 36, 37].

Self-evolving agent skills post measurable gains

EvoSkill is a self-evolving framework that analyzes execution failures, proposes new or revised skills, and stores them as reusable skill folders [33]. It uses three agents — an Executor, a Proposer, and a Skill-Builder — while keeping the base model frozen and selecting skills on a Pareto frontier [33]. Reported gains include improving Claude Code with Opus 4.5 from 60.6% to 67.9% exact-match accuracy on OfficeQA, adding 12.1% on SealQA, and transferring zero-shot to BrowseComp with a 5.3% lift [33].

Retrieval starts using the agent's own reasoning trace

AgentIR jointly embeds an agent's reasoning trace alongside its query, rather than embedding the query alone [34]. The paper argues the reasoning trace acts as retrieval instruction, memory of key history, and a filter for outdated information [34]. On BrowseComp-Plus with Tongyi-DeepResearch, AgentIR-4B reached 68% accuracy, versus 52% for conventional embedding models twice its size and 37% for BM25, while also beating LLM reranking by 10 percentage points without extra inference overhead [34].

Several projects targeted faster or more data-efficient model building

- **TDM-R1** uses reinforcement learning with non-differentiable rewards to train a few-step 6B text-to-image model. With only four NFEs, it raised GenEval from 61% to 92%, above the 80-NFE base model at 63% and GPT-4o at 84% [35].
- **Self-Flow** from Black Forest Labs builds learnability directly into flow models across image, video, and audio, with especially strong gains on harder video-action tasks such as Open and Place [36].
- **CosNet** reported 20%+ wall-clock pretraining speedups by attaching low-rank nonlinear residual functions to linear layers, and the code is now available [37, 38, 39].
- **Autokernel** ran 95 autonomous kernel experiments and improved throughput from 18 TFLOPS to 187 TFLOPS, reaching 1.31x cuBLAS across nine kernel types [40].

Products & Launches

Why it matters: Product work is shifting from standalone chat to tools that can share context, act across applications, and fit more naturally into existing software workflows [41, 42, 43, 44, 45].

Office workflows are becoming multi-agent

Claude for Excel and Claude for PowerPoint now sync across multiple open files, sharing full conversation context so users can pull data from spreadsheets, build tables, and update decks without re-explaining the task [41]. Anthropic's add-ins now support Skills as well [46].

IDEs are getting more agent-native

VS Code's Autopilot preview lets an agent stay in control of a workflow, run tools, retry on errors, and continue until the task is complete [42]. Cursor added more than 30 new plugins to its marketplace, including integrations for Datadog, Hugging Face, Glean, PlanetScale, Atlassian, and GitLab [47, 48, 49, 50, 51, 52, 53].

Google open-sources a UI language for agents

Google released A2UI, a UI language that lets agents describe interfaces in JSON while the client app renders them with trusted components [43]. Google highlights four benefits: declarative structure, safer rendering, framework-agnostic output, and incremental UI updates [43].

New multimodal models are shipping to users

Together AI introduced Qwen3.5 9B, a multimodal model with text, image, and video understanding, native tool calling, and 262K native context that can extend beyond 1M tokens [44, 54]. Google also rolled out Nano Banana 2 across Gemini, Search, Google Ads, Vertex AI, and Flow, describing it as combining Nano Banana Pro quality with Flash-level speed [45].

Industry Moves

Why it matters: Capital and partnerships continue to concentrate around open models, enterprise inference access, and AI-native software platforms [55, 56, 57, 58].

- **NVIDIA's open-model strategy is bigger than one release.** A Wired scoop shared by Will Knight says NVIDIA will spend $26 billion over the next five years building the world's best open source models [55].
- **Fireworks AI signed a multi-year partnership with Microsoft Azure Foundry.** The deal brings high-performance inference for leading open models into the Azure ecosystem, with Fireworks emphasizing security, compliance, and production quality [56].
- **Replit raised $400 million at a $9 billion valuation.** The company says it is now used at 85% of the Fortune 500 and will use the funding to expand beyond coding into AI systems centered on human creativity [57].
- **Anthropic is in talks with private-equity firms including Blackstone.** The reported plan is a joint venture to sell Anthropic's AI technology to portfolio companies; the talks were temporarily affected by the Anthropic-DoD dispute but are ongoing [58].

Policy & Regulation

Why it matters: Formal regulation was limited in this set, but the policy conversation is clearly shifting toward agent security, sandboxing, and deployment controls [59, 60, 61].

Security discussions are moving beyond adversarial attacks

In a response to NIST's request for information on AI agent security, Princeton researchers argued that many security failures happen even without adversaries, because unreliability itself is a major source of failure that has received too little attention in definition, measurement, and mitigation [59].

Governments are starting to treat agents as a new cyber surface

Ryan Fedasiuk argued that AI agents shift cyber risk from hacking a device to gaslighting an AI, and said governments should be scrambling to adapt [60]. In follow-on commentary about OpenClaw in China, another analyst predicted China would move toward a more secure, sandboxed version rather than stay with a blanket rejection of raw deployments [62].

Vendors are responding with stronger deployment security

ChutesAI released an end-to-end encryption proxy for OpenAI-compatible chat completions, Anthropic messages, and OpenAI responses formats using ML-KEM-768, HKDF-SHA256, and ChaCha20-Poly1305 with fresh ephemeral keys per request [61]. It is not regulation, but it is a concrete compliance-oriented response to the security demands around agent deployment [61].

Quick Takes

Why it matters: These smaller items sharpen the picture on frontier competition, healthcare, infrastructure, and global rollout [63, 64, 65, 66, 67, 68].

- Arena ranked **GPT-5.4** tied at **#2 on Document Arena** and in the **top 5 on Arena Expert**; both GPT-5.4 and GPT-5.4-High sit in the top 5 on expert-level prompts [63, 69].
- Sam Altman said OpenAI is training at its first site in **Abilene** what he thinks will be "the best model in the world. Hopefully by a lot." [64]
- Meta said its **MTIA** custom silicon program shipped **four generations in two years** to keep up with faster model-architecture cycles [65].
- Google Research said **AMIE** was found **safe, feasible, and well-received by patients** in a real-world clinical study with BIDMC [66].
- Google said its breast-cancer screening research with **Imperial College London** and the **NHS** identified **25%** of interval cancers that usually slip through screening [67].
- Google expanded **AI Studio** and the **Gemini API** to **Monaco, French Guiana, and Reunion Island**, opening access to about **1 million** more people [68].

Sources

1. X post by @ArtificialAnlys
2. X post by @DeepLearningAI
3. X post by @perplexity_ai
4. X post by @AnthropicAI
5. X post by @_TobiasLee
6. X post by @ArtificialAnlys
7. X post by @ArtificialAnlys
8. X post by @ArtificialAnlys
9. X post by @vllm_project
10. X post by @ggerganov
11. X post by @ollama
12. X post by @togethercompute
13. X post by @ArtificialAnlys

14. X post by @OpenAIDevs
15. X post by @OpenAIDevs
16. X post by @OpenAIDevs
17. X post by @perplexity_ai
18. X post by @perplexity_ai
19. X post by @perplexity_ai
20. X post by @perplexity_ai
21. X post by @perplexity_ai
22. X post by @perplexity_ai
23. X post by @perplexity_ai
24. X post by @perplexity_ai
25. X post by @perplexity_ai
26. X post by @AnthropicAI
27. X post by @AnthropicAI
28. X post by @AnthropicAI
29. X post by @jackclarkSF
30. X post by @karinanguyen_
31. X post by @karinanguyen_
32. X post by @karinanguyen_
33. X post by @omarsar0
34. X post by @dair_ai
35. X post by @William74312006
36. X post by @pess_r
37. X post by @torchcompiled
38. X post by @torchcompiled
39. X post by @torchcompiled
40. X post by @Akashi203
41. X post by @claudeai
42. X post by @code
43. X post by @TheTuringPost
44. X post by @togethercompute
45. X post by @dl_weekly
46. X post by @_catwu
47. X post by @cursor_ai
48. X post by @cursor_ai
49. X post by @cursor_ai
50. X post by @cursor_ai
51. X post by @cursor_ai
52. X post by @cursor_ai
53. X post by @cursor_ai
54. X post by @togethercompute
55. X post by @willknight
56. X post by @lqiao
57. X post by @amasad
58. X post by @steph_palazzolo
59. X post by @random_walker

60. X post by @RyanFedasiuk
61. X post by @jon_durbin
62. X post by @teortaxesTex
63. X post by @arena
64. X post by @scaling01
65. X post by @AIatMeta
66. X post by @GoogleResearch
67. X post by @Google
68. X post by @osanseviero
69. X post by @arena