Infrastructure, Open Models, and Agent Workflows Define the Day

AI News Digest

2026-03-12

Infrastructure, Open Models, and Agent Workflows Define the Day

By AI News Digest • March 12, 2026

Sam Altman used BlackRock's infrastructure summit to argue that frontier AI now depends as much on power, construction, and inference economics as on model progress. Elsewhere, NVIDIA launched a major open model for agentic systems, enterprise tools kept shifting toward orchestrated digital work, and governance proposals became more concrete.

Infrastructure became the main story

The clearest pattern today was that frontier AI is being described in terms of **power, chips, and construction** as much as model intelligence [1].

OpenAI framed frontier progress as a buildout problem

At BlackRock's US Infrastructure Summit, Sam Altman said OpenAI is already training at its first Stargate site in Abilene and described the challenges of getting gigawatt-scale campuses running, from unexpected weather to supply-chain issues and the need for many organizations to work together under pressure [1, 2]. He also said OpenAI's new partnership with the North American Building Trades Unions reflects a practical constraint: AI growth depends on physical infrastructure such as power plants, transmission, data centers, and transformers, plus more skilled trades workers to build them [1, 2].

Why it matters: The bottlenecks around frontier AI are increasingly physical, not just algorithmic.

Altman said costs are falling fast — and specialized inference hardware matters more

Altman said OpenAI's first reasoning model, o1, arrived about 16 months ago, and that getting the same answer to a hard problem from o1 to GPT-5.4 now costs about **1,000x less** [1, 2]. He also said the company is building an **inference-only** chip optimized for low cost and power efficiency, with first chips expected to be deployed at scale by year-end [1, 3]. Altman added that the past few months marked a threshold of major economic utility for these systems, especially in coding and other knowledge work [1].

> "To get the same answer to a hard problem from that first model to 5.4 has been a reduction in cost of about a thousand X." [1]

Why it matters: Capability gains are now being paired with meaningful cost compression, which is what turns impressive demos into deployable systems.

Open models and agent products widened the deployment race

NVIDIA released an open model aimed squarely at agentic AI

NVIDIA launched **Nemotron 3 Super**, a 120B-parameter open model with 12B active parameters, a **1-million-token context window**, and high-accuracy tool calling for complex agent workflows [4]. NVIDIA said it delivers up to **5x higher throughput** and up to **2x higher accuracy** than the previous Nemotron Super model, and is releasing it with open weights under a permissive license for deployment from on-prem systems to the cloud [4].

Why it matters: This is a substantial open-model push focused on enterprise-grade agents, not just model openness as a slogan.

Enterprise products kept moving from chat toward orchestrated work

Perplexity launched **Computer for Enterprise**, saying it can run multi-step workflows across research, coding, design, and deployment by routing work across **20 specialized models** and connecting to **400+ applications** [5]. The company said its internal Slack deployment performed **3.25 years of work** and saved **$1.6M** in four weeks, and that it is now exposing some of the same orchestration through a model-agnostic API platform [6, 7, 8].

The same shift appeared elsewhere: Replit introduced **Agent 4** for collaborative app-building with an infinite canvas and parallel agents [9], while Andrej Karpathy argued this does not end the IDE so much as expand it into an **"agent command center"** for managing teams of agents [10, 11].

Why it matters: A growing set of products is treating AI less like a single assistant and more like a coordinated workforce.

Governance ideas got more operational

Anthropic created a new public-benefit function around powerful AI

Anthropic said Jack Clark is becoming **Head of Public Benefit** and launching **The Anthropic Institute** to generate and share information about the societal, economic, and security effects of powerful AI systems [12, 13, 14]. Anthropic said the institute will bring together machine learning engineers, economists, and social scientists, using the vantage point of a frontier lab to inform public understanding [15, 16].

Why it matters: Frontier labs are starting to formalize impact analysis as an institutional function, not just a policy sideline.

A biosecurity proposal focused on restricting dangerous data, not shutting down open science

Johns Hopkins researcher Jassi Pannu outlined a **Biosecurity Data Level** framework that would keep roughly **99%** of biological data open while adding controls only to the narrow slice of functional data that links pathogens to dangerous properties such as transmissibility, virulence, and immune evasion [17]. She also pointed to model-holdout results suggesting that removing human-infecting virus data can sharply reduce dangerous biological capabilities while leaving desirable capabilities intact [17].

Why it matters: It is one of the clearest middle-ground governance proposals on the table: preserve open research broadly, but treat the most dangerous capability-enabling data as a controlled resource.

Sources

1. DIRECTO | Sam Altman participa en la Cumbre de BlackRock sobre Infraestructura de EE UU | ELPAÍS
2. LIVE: Sam Altman speaks at BlackRock's US Infrastructure Summit
3. WATCH LIVE: OpenAI CEO Sam Altman speaks at BlackRock's US Infrastructure Summit
4. New NVIDIA Nemotron 3 Super Delivers 5x Higher Throughput for Agentic AI
5. X post by @perplexity_ai
6. X post by @AravSrinivas
7. X post by @perplexity_ai
8. X post by @AravSrinivas
9. X post by @amasad
10. X post by @karpathy
11. X post by @karpathy
12. X post by @jackclarkSF
13. X post by @jackclarkSF

14. X post by @jackclarkSF
15. X post by @AnthropicAI
16. X post by @AnthropicAI
17. Bioinfohazards: Jassi Pannu on Controlling Dangerous Data from which AI Models Learn