# Grok Speeds Up as Open Models Broaden and Compute Constraints Stay Central

AI News Digest

2026-06-29

## Grok Speeds Up as Open Models Broaden and Compute Constraints Stay Central

*By AI News Digest • June 29, 2026*

Musk outlined a fast-moving Grok roadmap, open-model releases kept broadening across the industry, and new data underscored both the scale of AI spending and the continued bottlenecks around compute, power, and regulatory structure.

### Model race

#### xAI lays out an aggressive Grok roadmap

Musk said Grok 4.5—built on a 1.5T V9 foundation model with supplemental Cursor data—is now in private beta at SpaceX and Tesla, with early evals showing performance close to or possibly above Opus. He also said RL is still driving meaningful gains, while separately clarifying that the v9 foundation model should be viewed as a solid workhorse in Opus territory rather than something dramatically beyond the field [1, 2].

Across several follow-up posts, Musk described a very fast roadmap: SpaceX releasing new models trained from scratch every month this year, a new 2T run finishing in late July for an August release, Cursor contributing engineering work to v9 SFT and RL, top Starlink and Starship engineers shifting time to AI, and a planned C/C++ rewrite of the training and inference stack mapped tightly to GB300 hardware [1, 3, 2, 4].

**Why it matters:** These posts tie model ambitions directly to release cadence, data partnerships, engineering staffing, and hardware-software co-design [3, 2, 4].

### Open-model breadth keeps expanding

Interconnects argues the open-model ecosystem is becoming more diverse after being dominated a year ago by a small set of mostly Chinese players, with momentum now coming from three camps: pure model makers, Big Tech, and product companies training specialized models [5]. The latest batch it highlights includes NVIDIA's Nemotron 3 Ultra under OpenMDW, Cohere's Apache 2.0 Command A+, GLM-5.2 from zai-org, Zyphra's AMD-trained ZAYA1-74B-preview, and Poolside's Apache 2.0 Laguna-M.1 plus an ongoing commitment to open releases [5].

Nathan Lambert separately pointed to 30 open-model releases across May and June from companies ranging from NVIDIA, Google, and Microsoft to JetBrains, Ideogram, Krea, and Photoroom, calling the breadth a reminder that a lot of open-model value sits outside the shadow of the biggest frontier labs [6, 7].

**Why it matters:** Open releases are now coming from a wider mix of company types and geographies than they were a year ago [5, 6].

## Economics, infrastructure, and policy

### The AI economy looks large already, but cost discipline is rising with it

New research shared by Azeem Azhar and highlighted by Thomas Wolf estimates that the genAI economy generated $110 billion in sales over the last 12 months and is already running above $175 billion on an annualized basis. The authors describe it as the first bottom-up, deduplicated measure of consumer and enterprise AI spending across the full stack [8, 9].

At the operating level, Brian Armstrong said his team is trying to keep AI spend flat while token usage grows by defaulting more work to open-weight models such as GLM 5.2 and Kimi 2.7, routing tasks to the best-fit model, and improving cache hit rates from 5% to 60%; he said those changes cut AI spend nearly in half [10]. Gary Marcus, looking at the broader market, argued it is hard to see how anyone makes much money from AI in the long run, comparing the economics to airlines with small margins and big expenses [11].

**Why it matters:** Growth is not removing cost pressure; model choice, routing, caching, and margin structure are all moving closer to the center of AI strategy [10, 11].

### Compute scarcity is still shaping who gets access

Compute bottlenecks remain a live competitive constraint. One widely shared report said Google limited Meta's use of Gemini because of a shortage of compute resources, with the blunt takeaway that compute remains power and AI's scarcest resource [12, 13].

Andreessen also amplified analysis arguing that AI and general automation could push electricity demand beyond demographic ceilings, that the world may need to double electricity production from 30,000TWh per year to 60,000TWh over the next 20 years, and that datacenter buildout is becoming a gating issue [14, 15].

**Why it matters:** AI capacity is still being shaped by chips, power, and datacenter construction—not just model quality [12, 14, 15].

### A sharper policy split is forming between frontier APIs and open weights

Clement Delangue argued it is rational to regulate frontier API models for government transparency without regulating open-source AI [16]. He framed frontier APIs as secretive black boxes controlled by a few profit-driven megacorps and distributed at massive scale, while arguing open-weight models are easier to analyze, currently less capable of misuse, and equally available to defenders and attackers [16].

Delangue also argued the cost-benefit is different: API regulation is relatively easy and mostly lands on large incumbents, while open-source regulation would hurt startups, researchers, universities, nonprofits, and competition [16]. In the same discussion, Nathan Lambert called the idea of getting regulated for being too dangerous a horrible consequence of "vibe regulation," echoing Delangue's point that danger labeling can function as enterprise marketing [17, 18].

**Why it matters:** This is a more specific governance split than generic calls for AI regulation: frontier API access and open-weight distribution are being treated as different policy problems [16].

## One hardware idea to watch

### Normal Computing says its thermodynamic chip has reached silicon

Machine Learning Street Talk featured Normal Computing's thermodynamic-chip approach, which uses inherent chip noise for probabilistic computation by implementing stochastic differential equations directly in hardware [19]. The company's first chip, CN101, has reached silicon and is aimed at narrow probabilistic workloads including Bayesian inference, MCMC, and diffusion models [19].

> The noise is the computation. [19]

The team also said it used swarms of AI agents to generate a Verilog simulator with more than 500,000 lines of code as an alternative to commercial EDA software that can cost about $10,000 per CPU core [19].


*The Thermodynamic AI Chip · Thomas Ahle (2:18)*

**Why it matters:** It is a concrete example of AI-adjacent hardware exploration aimed at probabilistic workloads rather than general-purpose model serving [19].

---

**Sources**

1. X post by @elonmusk
2. X post by @elonmusk
3. X post by @elonmusk
4. X post by @elonmusk
5. Latest open artifacts (#22): Zyphra, Cohere, and Poolside are expanding the breadth of the ecosystem
6. X post by @interconnectsai
7. X post by @natolambert
8. X post by @azeem
9. X post by @Thom_Wolf
10. X post by @brian_armstrong
11. X post by @GaryMarcus
12. X post by @jukan05
13. X post by @GaryMarcus
14. X post by @Object_Zero_
15. X post by @pmarca

16. X post by @ClementDelangue
17. X post by @natolambert
18. X post by @ClementDelangue
19. The Thermodynamic AI Chip · Thomas Ahle