

Mira Murati's Thinking Machines releases Inkling, its first model, with open weights

Source copy kept by wheresthe.ai for the story "Mira Murati's Thinking Machines releases Inkling, its first model, with open weights" (<https://wheresthe.ai/d/us-thinking-machines-inkling-open-2026-07/>).

Source: <https://techcrunch.com/2026/07/15/thinking-machines-amps-up-its-bet-against-one-size-fits-all-ai-with-its-first-open-model-inkling/>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 58a093388a4c3d854b200852ac02285cd08666b32edaecb3b785bf95603a17e0

Thinking Machines Lab, the AI startup founded by former OpenAI CTO Mira Murati, released its first in-house AI model Wednesday morning, called Inkling. And unlike the flagship models from OpenAI, Anthropic, or Google, it's open-weight, meaning outside developers and companies can download it and modify it directly.

Inkling is a mixture-of-experts system with 975 billion total parameters, though it only draws on a fraction of that — about 41 billion — for any given task, a common design that keeps very large models faster and cheaper to run. It was trained on 45 trillion tokens of text, image, audio, and video, and reasons natively across all four, according to the company's own release materials. For now, though, its outputs are limited to text, including code, styled artifacts, and structured data.

The model is Thinking Machines Lab's first public proof point after a year and a half spent building AI infrastructure largely out of public view. Some of that work had already surfaced in a May research preview of "interaction models" — AI designed to listen and speak (and even interrupt) instead of stop and wait as with typical chatbots. It's also a test of the central bet behind the startup, which is that AI that organizations can adapt for themselves will outperform the one-size-fits-all models the biggest labs currently sell.

Inkling is designed to give calibrated answers, including flagging uncertainty rather than guessing, and lets users dial "thinking effort" up or down when they want to trade for speed. On one benchmark, the company says, Inkling uses a third as many tokens as Nvidia's Nemotron 3 Ultra — its latest generation open-weight model — to hit the same coding performance.

Thinking Machines doesn't claim Inkling is best-in-class. Its newest blog post states explicitly that Inkling is "not the strongest overall model available today, open or closed." What it's evidently going for instead is well-rounded performance.

That raises the question of who, within the enterprise market it's targeting, this product is really for. Thinking Machines is, for now, marketing Inkling less as a finished product than as a starting point, something for organizations to fine-tune themselves through Tinker, the company's model-customization platform. This also means customers, not Thinking Machines, are responsible for making sure their customizations are safe, for example. (Fine-tuning requires serious machine-learning talent.)

OpenAI, Anthropic, and Google have all taken a very different approach with ChatGPT, Claude, and Gemini, respectively, which were all built to compete as general-purpose chatbots first, with agentic, autonomous features layered on top.

A post published by Thinking Machines last week was clearly meant as the backdrop for this release. AI that's trained centrally by one company and then set in stone, the company argued in that post, underperforms AI that organizations shape themselves because so much expertise is specific to the people who hold it.

Other arguments against closed models are gaining steam. In a blog post published Sunday, Microsoft CEO Satya Nadella — whose company has invested billions in both OpenAI and Anthropic — warned that enterprises using proprietary AI models effectively pay twice: once in subscription costs and again by handing over business knowledge embedded in their prompts and corrections, which can be absorbed into future model versions.

Hugging Face CEO Clem Delangue made a similar prediction in conversation with TechCrunch last week. Frontier models, he said, will increasingly be reserved for experimentation and high-value tasks, while most production AI work shifts to private or open source alternatives — the exact split Thinking Machines is building around.

The clearest argument for Thinking Machines' approach came from a recent project with Bridgewater Associates,

the world's largest hedge fund (which is not, for what it's worth, a Thinking Machines investor). Researchers from both companies took an existing open source model and trained it further on Bridgewater's own financial expertise. The result was said to score 84.7% on financial reasoning tests, beating top proprietary AI models, while costing roughly a fourteenth as much to run — though those results come from the two companies' own evaluation, not an independent one.

Either way, Thinking Machines is emphasizing how quickly it got here. OpenAI took roughly five years to bring its tech to market and show revenue, and Anthropic roughly three. Thinking Machines says it did the same in about nine months.

Some will wonder whether Inkling was trained on outputs from competitors' models, a practice known as “distillation” that has drawn scrutiny across the industry. The short answer, per the company's own materials, is partly. Thinking Machines pre-trained Inkling from scratch, but it says it used other open-weight models — including Moonshot AI's Kimi K2.5 — to help generate some of its early post-training data before large-scale reinforcement learning took over. The next model, the company insists, will use fully self-contained post-training instead.

On the cost side, Thinking Machines has been more guarded. It struck a partnership with Nvidia in March to deploy a gigawatt of Vera Rubin computing capacity and trained Inkling entirely on Nvidia's GB300 NVL72 systems — but hasn't said how it plans to cover those costs, and revenue, by most accounts, hasn't been a priority. (A reported \$50 billion fundraising round was said to be coming together last November but had stalled by January; the company has declined to talk about its funding picture since.)

A related question is whether Thinking Machines' spending will ever reach the scale of OpenAI's or Anthropic's, or whether its efficiency-driven approach means the economics look different. Put another way, the company's bet may be less that it will eventually spend like its larger rivals than that it won't need to at all — because once weights are public, nothing obligates anyone who downloads them to pay Thinking Machines to run them, unlike the metered access OpenAI and Anthropic sell. It's Tinker, not the model itself, where the company's revenue has to come from, via training, fine-tuning, and, now, a cut of the hosting ecosystem built around it.

Headcount, at least, looks more settled. Thinking Machines now employs roughly 200 people, up from levels reported after a wave of departures earlier this year, including two co-founders who left for OpenAI in January.

Thinking Machines, for its part, doesn't seem interested in playing up individual moves the way much of the industry does. According to a source inside the company, its culture, by design, favors continuity over reliance on any one personality. It makes sense: It's less of a setback when people change teams if they were never put on a pedestal to begin with. It's also a remarkable thing for a company to insist on, given how much of its own story is still associated with the name of its now-famous co-founder, whether she planned it or not.

Topics

AI , Mira Murati , thinking machine labs

When you purchase through links in our articles, we may earn a small commission . This doesn't affect our editorial independence.

Connie Loizos

Editor in Chief & General Manager

Loizos has been reporting on Silicon Valley since the late '90s, when she joined the original Red Herring magazine. Previously the Silicon Valley Editor of TechCrunch, she was named Editor in Chief and General Manager of TechCrunch in September 2023. She's also the founder of StrictlyVC, a daily e-newsletter and lecture series acquired by Yahoo in August 2023 and now operated as a sub brand of TechCrunch.

You can contact or verify outreach from Connie by emailing or , or via encrypted message at ConnieLoizos.53 on Signal.

View Bio

November 4

Boston

Get 50% off a second pass

Cover more ground at TechCrunch's ultimate founder bootcamp of the year by bringing your co-founder or partner. Save 50% off a second pass. Meet scaling and investment leaders, ask questions, find solutions, and more.

Last day to save Oct. 16

[Register Now](#)

Most Popular

Mark Wahlberg wants to take on the longevity-startup crowd with an affordable gym concept

Connie Loizos

Tesla reveals SpaceX-powered Roadster that flies and falls with style

Sean O'Kane

FBI arrests OpenAI employee accused of planting hidden recording device in conference room

Zack Whittaker

Bumble agrees to a \$1.3B acquisition deal as online dating faces a reset

Lauren Forristal

Just 5 years in, interactive chatbot platform Character.AI files for bankruptcy

Ivan Mehta

Meta releases its first frontier model, Muse Wonder

Russell Brandom

This AI startup says it can generate permit-ready blueprints in minutes

Kirsten Korosec