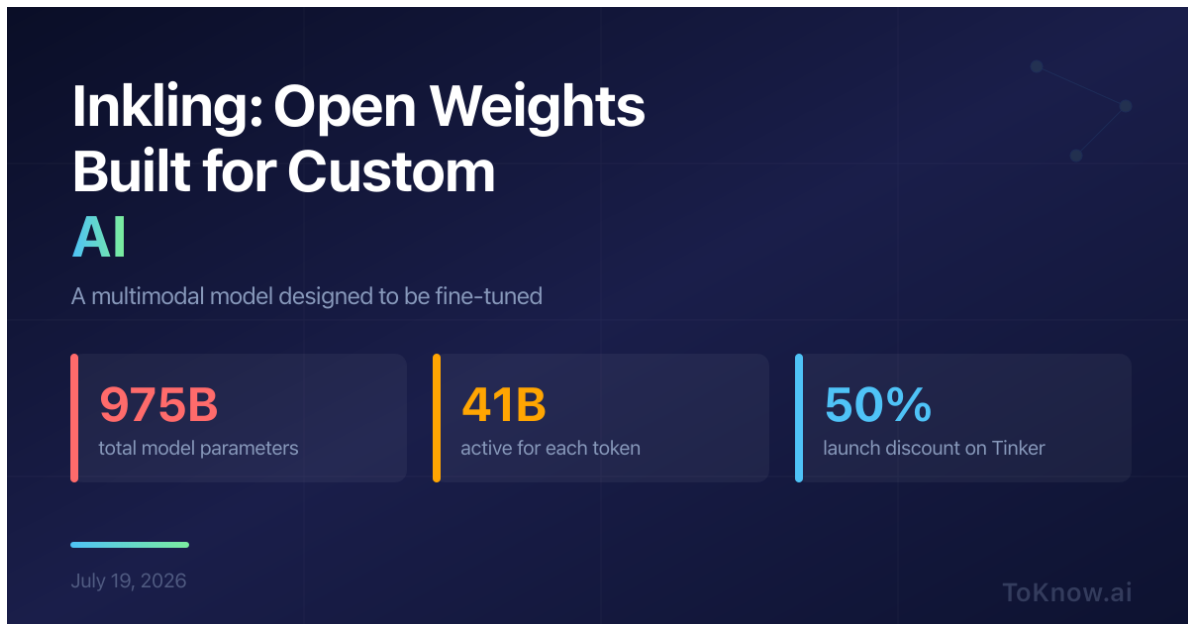


Inkling: Open Weights Built for Custom AI

Kabui, Charles

2026-07-19

[Read at ToKnow.ai](#)



The banner features a dark blue background with a subtle grid pattern. In the top right corner, there is a small, faint graphic of a neural network with three nodes and connecting lines. The main title 'Inkling: Open Weights Built for Custom AI' is prominently displayed in white and light blue. Below the title, a subtitle reads 'A multimodal model designed to be fine-tuned'. Three key statistics are highlighted in separate boxes: '975B total model parameters' with a red vertical bar, '41B active for each token' with an orange vertical bar, and '50% launch discount on Tinker' with a light blue vertical bar. At the bottom left, the date 'July 19, 2026' is shown next to a small horizontal bar with a green-to-blue gradient. The 'ToKnow.ai' logo is positioned in the bottom right corner.

**Inkling: Open Weights
Built for Custom
AI**

A multimodal model designed to be fine-tuned

975B total model parameters	41B active for each token	50% launch discount on Tinker
---------------------------------------	-------------------------------------	---

July 19, 2026

ToKnow.ai

Mira Murati's Thinking Machines Lab released [Inkling](#), its first model trained from scratch. It is a mixture-of-experts model with 975 billion total parameters but activates 41 billion per token, lowering the work needed for each response. Inkling accepts text, images, and audio, can handle up to one million tokens of context, and lets developers control how much computation it spends reasoning. The full weights are downloadable under the [Apache 2.0 license](#), subject to a separate [Model Acceptable Use Policy](#), including a smaller NVFP4 version for Nvidia hardware. Thinking Machines also previewed Inkling-Small, with 276 billion total and 12

billion active parameters. The company says Inkling is not the strongest model overall. Its pitch is the combination of multimodal input, efficient reasoning, and weights that developers can alter.

The unusual offering is how the model connects to customization. Developers can fine-tune Inkling through [Tinker](#), download the resulting weights, and run them elsewhere. Tinker limits Inkling training to 64K or 256K context windows. After a limited 50% discount, its 64K tier costs \$1.87 per million input tokens, \$4.68 per million generated tokens, and \$5.61 per million training tokens. An Inkling Playground is also available in the Tinker console. In the company's demo, Inkling wrote and ran its own fine-tuning job in about 27 minutes, then loaded the new checkpoint. That turns customization from a dedicated GPU project into an API job, although Tinker supports low-rank adaptation rather than changing every model weight.

Inkling separates access from practicality. Its estimated 1.9 TB BF16 checkpoint is genuinely downloadable, but most users will still need a cloud host. Open weights give companies control over deployment and tuning, while Tinker, hosted APIs, and hardware partners provide the expensive path from files to a usable product.

Sources:

- [Thinking Machines Lab: Introducing Inkling](#)
- [Inkling model card and weights](#)
- [Tinker training API documentation](#)
- [Tinker models and pricing](#)

Disclaimer: For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. **Read more:** [/terms-of-service](#)