

Mira Murati Releases Inkling: Thinking Machines Lab's First Open-Weight Model

MANAGEMENT SUMMARY & CORE TAKEAWAY

Mira Murati, former CTO of OpenAI, has released Inkling through her startup Thinking Machines Lab. The multimodal mixture-of-experts system features 975 billion parameters (41 billion active), operates under an Apache 2.0 license, and enables Swiss enterprises to fully self-host, customize, and optimize a high-performance AI model for specialized domains—without vendor dependency and with controllable thinking effort balancing cost efficiency and maximum performance.

CONTEXT & EXECUTIVE RELEVANCE

Mira Murati shaped the development of ChatGPT and DALL-E as Chief Technology Officer at OpenAI. In September 2024, she served as interim CEO during the Altman crisis. In February 2025, she founded Thinking Machines Lab alongside John Schulman (OpenAI co-founder, ChatGPT lead) and Lilian Weng (former VP Safety and Robotics at OpenAI). Her vision: create AI systems that extend human will and judgment—collaborative, human-centered, and customizable.

On July 15, 2026, Thinking Machines Lab released Inkling, the startup's first full-scale open-weight model. With 975 billion parameters (41 billion active), a mixture-of-experts architecture, and an Apache 2.0 license, Inkling sets new standards for customizable AI systems. Unlike closed models, you can fully self-host, modify, and tailor Inkling to your specific domain—a decisive advantage for Swiss enterprises with strict compliance requirements.

CORE ACTION ITEMS

- 01 Text input: Up to 1 million tokens, ideal for long documents or multi-turn dialogues
- 02 Image input: 40×40-pixel patches processed via hierarchical MLP without separate vision encoder
- 03 Audio input: dMel spectrograms for speech recognition, transcription, or multimodal analysis
- 04 Text output: Code, structured data, styled artifacts—no image or audio generation

SWISS COMPLIANCE & ARCHITECTURE GUIDELINES

- ☑ Full compliance with revised Swiss Data Protection Act (revDSG).
- ☑ Zero vendor lock-in on closed proprietary models.
- ☑ Seamless integration into Abacus, Bexio and M365 ecosystems.

Non-Binding Process Analysis

Let us review in 20 minutes where AI agents create immediate measurable efficiency in your workflows.

ki-outsourcing.ch