Inkling 2 Mini explained: a looped mixture-of-experts model built for local setups

23 November 2026 · 7 min read · blogs.iictp.org/inkling-2-mini-looped-transformer-local-model-explained

Short answer: Inkling 2 Mini is a multimodal mixture-of-experts model from Thinking Machines Lab, introduced in a Hugging Face post on 20 November 2026, with about 26B total and 3B active parameters per token and a 1M-token context. Its distinguishing design is a looped core: 12 middle layers run three times, so 24 physical layers behave like 48 logical ones. Hugging Face's tests are described as vibe tests, and the benchmark claims come from the model card.

TL;DR The short version

- Inkling 2 Mini is a multimodal mixture-of-experts model from Thinking Machines Lab with about 26B total and 3B active parameters and a 1M-token context.
- Its looped core runs 12 middle layers three times, so 24 physical layers behave like 48, passing messages between loops.
- Benchmark claims come from the vendor and Hugging Face, so test it on your own prompts against your current local model.

Read online, listen to it, or change the text size: https://blogs.iictp.org/inkling-2-mini-looped-transformer-local-model-explained

What is Inkling 2 Mini?

It is a smaller sibling of Inkling 2 Small and is designed for local AI setups. It accepts images and text, and ships in a mixed-precision format with NVFP4 and MXFP8 expert weights plus a DSpark drafter for speed-ups at a small memory cost, per the post.

How the looped architecture works

- The decoder runs layers 0 to 5 once, layers 6 to 17 three times with the same weights, and layers 18 to 23 once, giving 48 layer visits per token from 24 stored layers.
- Between passes, **cross-pass messages** carry information: each looped layer reads a message from the previous pass using its own learned query, and a small gating controller decides how much of the current input to forget and how much of the message to add.
- Each six-layer block has five sliding-window layers and one full-attention layer, with a 1,024-token local window.
- The router picks 8 of 128 experts per token; hidden size is 2,304.

What is new compared with earlier approaches

The post says looping has been explored before, for example in Nanbeige, but Inkling 2 Mini passes information between loops explicitly. A research variant, TML-d24-varloop-research, was trained to accept a variable number of passes, so you can set one or two passes for faster, possibly lower-quality generation. The post says it works but has not been evaluated in depth.

Results reported by the source

- Hugging Face says Inkling 2 Mini is state of the art in its class on the Artificial Analysis Intelligence Index v5, and leads on 32 of 48 benchmarks against Gemma 4 31B, Muse Glimmer, Qwen3.8-27B and Nemotron 3.5 Super VL.
- It reports a lead on FrontierCode v2.0, a tie with Qwen3.8-27B on Terminal-Bench 4.0 and a deficit on AutomationBench-AA v1.0.6.
- The authors add that, in vibe tests, it is at least comparable with Qwen 3.8 27B and that more validation is needed.

Limitations and open questions

- Benchmark claims are the vendor’s and Hugging Face’s; check the model card and replicate on your tasks.
- We read the blog post, not the model card or weights.
- Looped models still execute all 48 visits, so compute per token does not drop; only the stored distinct layers do.

How to try it

1. Load the fixed-loop model through Transformers (`AutoModelForMultimodalLM`), llama.cpp, MLX, or SGLang for datacentre serving.
2. Test your own prompts against your current local model.
3. Try the variable-loop research checkpoint with 1 to 3 passes if latency matters.
4. Measure memory, speed and quality, not just the leaderboard.

What it means for local-LLM practitioners

Looped weights are a way to buy capacity without storing more distinct parameters, which suits memory-limited devices. See LLM quantisation explained and edge AI with small models.

Frequently asked questions

- **How many parameters does it have?** About 26B physical total and 3B active per token, per the post.
- **What is the context window?** 1M tokens.
- **Does it run in llama.cpp?** The post says Llama App (llama.cpp), Transformers and MLX are supported, with SGLang for production.

Related reading

- LLM quantisation explained: GGUF and AWQ
- Edge AI inference with small models
- What is an LLM context window?
- Local LLM Practitioner

Sources and currency

Sources were read on 23 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

> Related credential: Local LLM Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Hugging Face. Welcome Inkling 2 Mini, Cutting-edge Performance for your Local Setup. 20 November 2026. huggingface.co/blog/inkling2mini.

Published by IICTP, an independent technical credential body, on 23 November 2026. The current version of this post is at https://blogs.iictp.org/inkling-2-mini-looped-transformer-local-model-explained. More guides: blogs.iictp.org.