MODELS

LFM2.5-230M: Built to Run Anywhere

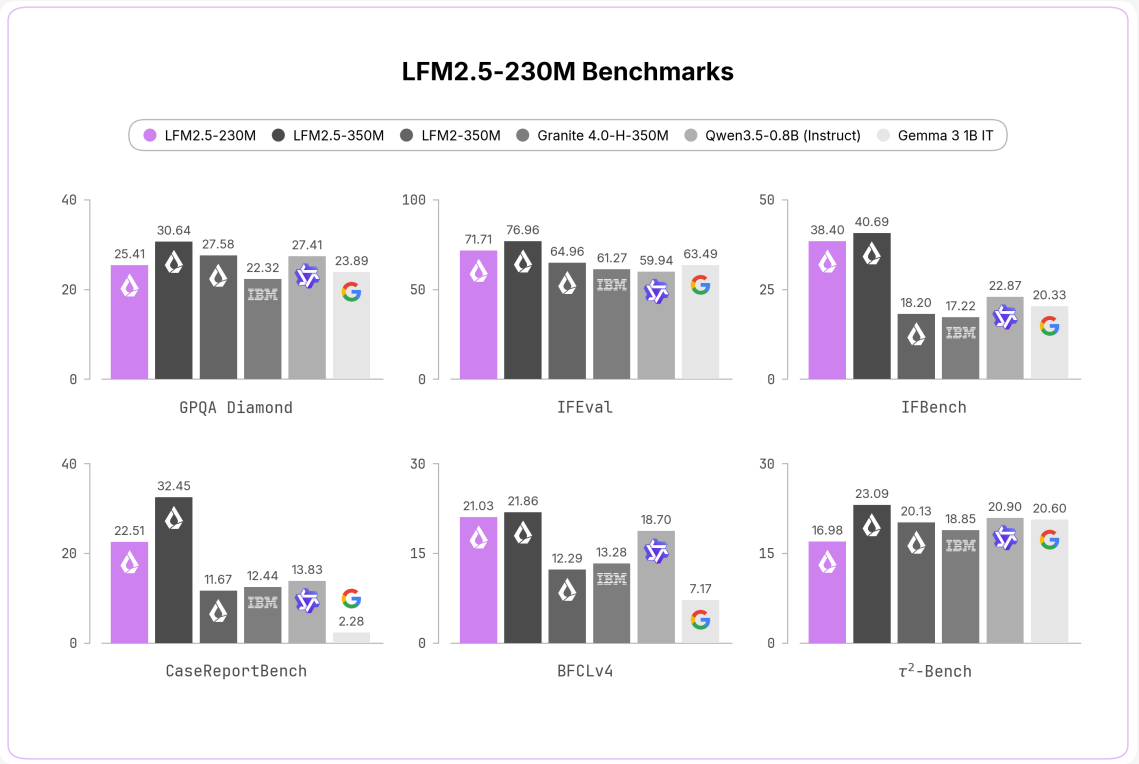
AUTHORS
Liquid AI

PUBLISHED
June 25, 2026

Training & Fine-tuning
Benchmarks
Fast Inference Everywhere

Today, we're releasing **LFM2.5-230M**, our smallest model yet. It's a fast, lightweight foundation for developers to fine-tune and deploy in agentic workflows. Built on the LFM2 architecture, it delivers exceptionally fast inference and runs everywhere, from cloud GPUs to low-cost CPUs (213 tok/s decode speed on Galaxy S25 Ultra, 42 tok/s on a Raspberry Pi 5).

The base (LFM2.5-230M-Base) and post-trained (LFM2.5-230M) models are available today on Hugging Face. Check out our docs on how to run and fine-tune them locally.



Training & Fine-tuning

The model was pre-trained for 10T tokens, including a 32K context

The recipe consists of three stages: **(1) supervised fine-tuning with distillation from LFM2.5-350M, (2) direct preference optimization, and (3) multi-domain reinforcement learning**. The final checkpoint balances strong out-of-the-box capabilities with adaptability to downstream specialization, while remaining competitive with larger models.

As an early look at ongoing work, we deployed LFM2.5-230M on a Unitree G1 humanoid robot, running entirely on-device on its onboard NVIDIA Jetson Orin. Here the model acts as a skill-selection layer: it takes a single natural-language instruction and decomposes it into a sequence of tool calls that invoke pre-trained low-level skills provided by NVIDIA's SONIC framework. After a quick fine-tune for this task, the model turns a free-form command such as

> *"Hold still for 2 seconds, then walk forward at 1 meter per second for 3 meters, hold a forward one-leg kneel for 5 seconds, and walk backward at 0.5 meters per second for 3 meters"*

into a structured, multi-step plan, chaining skills like timed walking at a target velocity and a one-legged kneel. While the behaviors are deliberately simple at this stage, we think it's a compelling signal: a 230M-parameter model can be quickly fine-tuned and deployed on-device to serve as the natural-language control interface for a humanoid.

Benchmarks

We evaluated LFM2.5-230M across ten benchmarks covering both core capabilities and applied tasks. Despite its size, it **competes with and often beats models more than twice as large**, spanning knowledge (GPQA Diamond, MMLU-Pro), instruction following (IFEval, IFBench, Multi-IF), data extraction (CaseReportBench), and tool use (BFCLv3, BFCLv4, τ^2-Bench Telecom and Retail).

Model	GPQA Diamond	MMLU-Pro	IFEval	IFBench
LFM2.5-230M	25.41	20.25	71.71	38.40
LFM2.5-350M	30.64	20.01	76.96	40.69
LFM2-350M	27.58	19.29	64.96	18.20
Granite 4.0-H-350M	22.32	13.14	61.27	17.22
Granite 4.0-350M	25.91	12.84	53.48	15.98
Qwen3.5-0.8B (Instruct)	27.41	37.42	59.94	22.87
Gemma 3 1B IT	23.89	14.04	63.49	20.33

Model	CaseReportBench	BFCLv3	BFCLv4	τ^2-Bench Telecom
LFM2.5-230M	22.51	43.26	21.03	5.26
LFM2.5-350M	32.45	44.11	21.86	18.86
LFM2-350M	11.67	22.95	12.29	10.82
Granite 4.0-H-350M	12.44	43.07	13.28	13.74
Granite 4.0-350M	0.84	39.58	13.73	2.92
Qwen3.5-0.8B (Instruct)	13.83	35.08	18.70	12.57
Gemma 3 1B IT	2.28	16.61	7.17	9.36

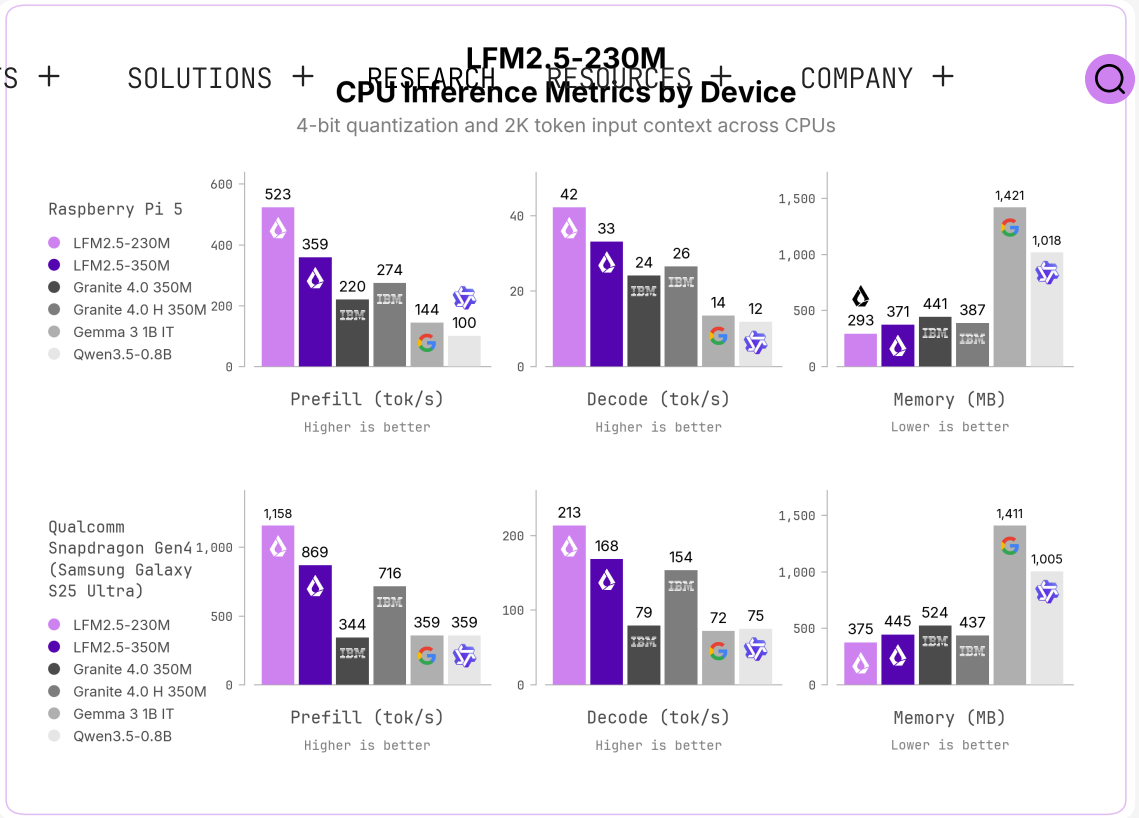
This makes LFM2.5-230M an ideal solution to power large-scale data extraction pipelines or lightweight on-device agentic workloads. However, given its compact size, we do not recommend it for reasoning-heavy workloads such as advanced math, code generation, or creative writing.

Fast Inference Everywhere

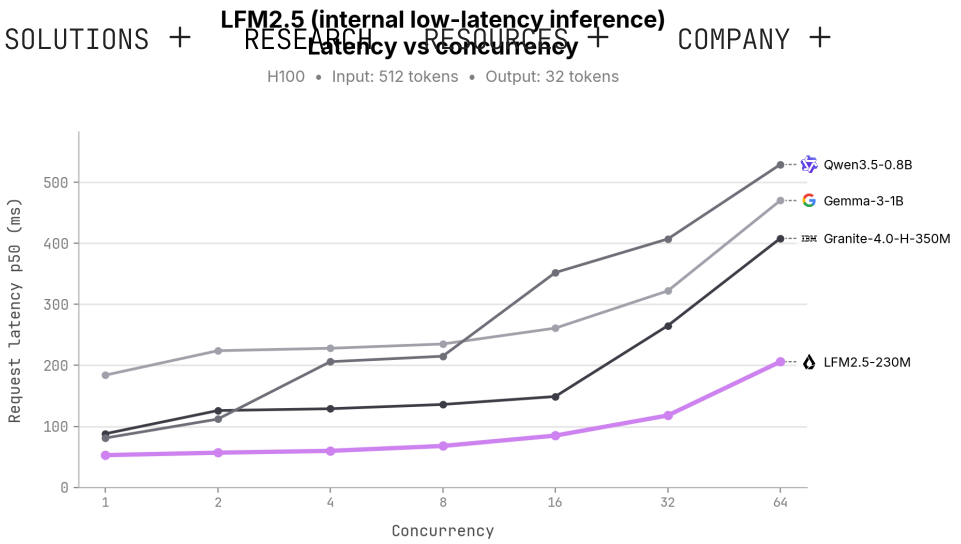
LFM2.5-230M ships with day-one support across the inference ecosystem.

- **llama.cpp** — GGUF checkpoints for efficient edge inference
- **MLX** — Optimized inference for Apple Silicon
- **vLLM** — GPU-accelerated serving for production throughput
- **SGLang** — GPU-accelerated serving for production throughput
- **ONNX** — Cross-platform inference across diverse accelerators

CPU inference. Thanks to the efficient LFM2 architecture, LFM2.5-230M is considerably faster than similar-sized models, including SSM hybrids and Gated Delta Networks. On both a Raspberry Pi 5 and a Qualcomm Snapdragon Gen4 (Samsung Galaxy S25 Ultra), it delivers the highest prefill and decode throughput in its class while keeping the smallest memory footprint. We tune the flash-attention flag per device to maximize prefill on each platform: enabled (-fa 1) on the Raspberry Pi 5 and disabled (-fa 0) on the Snapdragon Gen4.



GPU inference. For production-grade enterprise deployments, we have also developed an internal GPU inference stack that delivers extremely low-latency serving. We benchmark it against other small models running on SGLang, and across all concurrency levels, LFM2.5 models achieve considerably lower end-to-end latency.



Get Started

Start building today with LFM2.5-230M and LFM2.5-230M-Base, available on Hugging Face.

With LFM2.5, we're delivering on our vision of AI that runs anywhere. These models are:

- **Open-weight** — Download, fine-tune, and deploy without restrictions

- **A complete family** — From base models for customization to specialized audio and vision variants, one architecture covers diverse use cases

The edge AI future is here. We can't wait to see what you build.

🤗 **Download on Hugging Face** ↗

Read the Docs ↗

Citation

Please cite this article as:

> Liquid AI, "LFM2.5-230M: Built to Run Anywhere", *Liquid AI Blog*, Jun 2026.

Or use the BibTeX citation:

```
@article{liquidAI2026230M,
    author = {Liquid AI},
    title = {LFM2.5-230M:
    Built to Run Anywhere},
    journal = {Liquid AI Blog},
```

Get Liquid Apollo for free

Vibe check LFMs directly on your phone.

Download on the App Store

GET IT ON Google Play

SOLUTIONS

- Enterprise Solutions
- Start Up Solutions
- Automotive
- Consumer Electronics
- Ecommerce
- Financial Services
- Healthcare & Life Sciences
- Industrial & Robotics
- Case Studies

TECH & PRODUCTS

- Liquid Foundation Models
- LEAP
- Liquid Apollo
- Research
- Pricing & Licensing

DEVELOPER RESOURCES

- Demos
- Documentation
- Developer Community
- Hackathons

COMPANY

- About
- Blog
- Press
- Careers ↗
- Contact Us
- Shoutouts

LEGAL

- LFM License
- Privacy Policy
- Privacy Choices
- Terms & Conditions

314 Main St, Cambridge, MA 02142