

Forum:[Hack Forums](#)

Thread: [LFM 2.5: Fastest AI Vision Model Runs On Your Laptop And It Feels Unreal](#)

JoseMar - March 23, 2026, 9:12 am

LFM 2.5: Fastest AI Vision Model
Runs On Your Laptop And It Feels Unreal

Here again, found a new AI too
and I think this one's worth your time
Alright, so I just scrolling on TikTok, half paying attention, and then boom, I run into this thing called LFM 2.5 from Liquid AI.

And I had to double check because... it didn't sound right.

A vision model, running locally, on a normal laptop? No cloud? No "send your data somewhere"?
Just... runs?

Yeah, that.

Wait, This Runs Where Exactly?
So apparently, this thing runs entirely on your device. Laptop, phone, whatever. No server dependency.

Like... fully offline if you want.

I don't know, maybe I'm late to the party, but this feels like one of those shifts people ignore at first and then suddenly everything is built around it.

We've been so used to sending everything to the cloud that it just feels "normal." But when you think about it, it's kinda messed up, especially for anything sensitive.

This? Doesn't leave your machine.

Privacy people are probably smiling right now.

Also, it uses WebGPU and ONNX, which, I mean, sounds technical, but the point is... it's fast enough without needing a monster setup.

That's the weird part.

The Architecture Sounds Like Overkill, But Works
They mixed convolution stuff with something called grouped query attention.

I'm not gonna pretend I fully "feel" that intuitively, but the result is what matters.

It can understand images, text, both together, and do stuff like captioning, answering questions about images, even analyzing documents.

Basically, it's not just "look at image and say cat."

It's more like, "look at this messy document and explain what's going on."

Which... on a laptop... still sounds wrong to me.

But here we are.

The Part That Actually Surprised Me
Less than 1GB RAM.? No, seriously.

I've seen browser tabs eat more than that doing nothing.

And this thing handles a 32K context window. That's not small. That's like, real work, long documents, actual use cases.

Also supports images up to 512x512 natively, and bigger ones using tiling, which is basically chopping them up smartly so your system doesn't die.

It's kinda clever, honestly.

Not flashy, just efficient.

And maybe that's the point.

Real Time Stuff, On Device...
This is where I paused.

Real-time video analysis, object detection, text recognition... locally.

No lag from sending data somewhere.

No dependency on internet speed.

If this actually works as advertised, that's a big deal for a lot of use cases people don't even think about yet.

Like, imagine working somewhere with bad connection, or dealing with sensitive data you can't upload.

This just... works. I guess.

Still feels weird saying that.

So... Is This The Direction Now?
I don't know if this kills cloud AI, probably not, at least not fully.

But it definitely chips away at it.

And honestly, it makes you question why we're sending everything to servers in the first place.

Convenience? Sure.

But if models like this keep improving... then what?

Local, private, fast enough.

That combo wasn't really "a thing" before. Now it kinda is.

DramaKing - March 23, 2026, 9:50 am

Multimodal model with 1.2B parameters that isn't total crap? Sounds like a fantasy.

BiNBo - March 23, 2026, 10:12 am

Have to side with [DramaKing](#) here on this one. I can see it running with that small of a parameter count on pretty lightweight hardware, and with some purpose training and tweaking I could see it even being a good model for something like a smart camera project. However for general use? I wouldn't trust it to handle anything too heavy.

DramaKing - March 23, 2026, 10:19 am

Heck, I get model optimization and models getting smaller, but Whisper is bigger than this, and it only does TTS.

JoseMar - March 23, 2026, 10:43 am

(Mar 23, 2026 10:12 AM)BiNBo Wrote: [\[CITE\]](#) Have to side with [DramaKing](#) here on this one. I can see it running with that small of a parameter count on pretty lightweight hardware, and with some purpose training and tweaking I could see it even being a good model for something like a smart camera project. However for general use? I wouldn't trust it to handle anything too heavy.

(Mar 23, 2026 10:19 AM)DramaKing Wrote: [\[CITE\]](#) Heck, I get model optimization and models getting smaller, but Whisper is bigger than this, and it only does TTS.

Not long ago, a 1.2B multimodal model would've sounded like a joke. Text model have 1.2B
Vision model 1.6B

But I think we are missing the main point here, it's not about replacing large models.

If it runs locally, in real time, under 1GB RAM, and handles vision + text decently... that's already a win.

Check it out:

<https://www.liquid.ai/blog/introducing-lfm2-5-the-next-generation-of-on-device-ai#language-models>

7 - March 23, 2026, 10:47 am

well maybe its not that quick on 1gb of ram but maybe if we unlock it, it can be very useful for documents and K*C

JoseMar - March 23, 2026, 11:04 am

(Mar 23, 2026 10:47 AM)7 Wrote: [\[CITE\]](#) well maybe its not that quick on 1gb of ram but maybe if we unlock it, it can be very useful for documents and K*C

Check this video

[Video: <https://www.youtube.com/watch?v=GOyp0E4rGK4>]

MaîtreCornichon - June 9, 2026, 2:33 pm

Sounds like the future of AI is finally here; local processing could change everything for privacy.

Djangoo - June 26, 2026, 7:47 pm

Extremely useful information as always. Djangoo is excited to experiment with this today on his laptop.?

