

Forum: [Hack Forums](#)

Thread: [Google Gemma 4 Release Feels Bigger Than It Looks At First Glance](#)

JoseMar - April 3, 2026, 3:36 pm

Google Gemma 4 Release
Feels Bigger Than It Looks At First Glance

Alright so I just finished reading about Gemma 4 and... I'm not gonna lie, I didn't expect to actually stop and reread parts of it.

Like yeah, "new model release", we see that every week now. Most of the time it's just slightly better benchmarks, slightly cheaper inference, same story...?

This Feels Like Google Is Playing A Longer Game

So first thing that hit me, this isn't just one model. It's a whole family, from tiny on-device stuff all the way up to 31B.

And yeah, we've seen that before, but the weird part is how usable the smaller ones actually sound.

E2B and E4B running multimodal, text, image, even audio... on local devices?

Like... are we at the point where your phone can legit run something that used to require a GPU setup?

Because if that's real, not just marketing, that's kind of insane.

Also 128K context on small models, 256K on bigger ones. I remember when 8K felt like luxury. Now it's like... background noise.

The MoE Thing Is Actually Kinda Smart

The 26B model with only ~4B active parameters...

That's one of those things that sounds boring until you think about it for a second.

It basically runs like a small model but thinks like a bigger one, at least in theory.

And if that holds up in real usage, not just benchmarks, then yeah... this could be the sweet spot everyone's been trying to hit.

Fast, cheap, but still capable.

Feels like the direction everything is heading anyway.

Benchmarks Look Strong But That's Not Even The Point

Yeah the numbers are good, obviously.

MMLU, coding, reasoning, all solid improvements.

But honestly? I kinda stopped caring about benchmarks the moment everything started scoring “high”.

What matters more is the combo here.

Multimodal, long context, reasoning mode, function calling, system prompts... all baked in.

It's like they're not trying to win one category, they're trying to make something that just works everywhere.

Which is... actually more useful?

The On-Device Part Is What Keeps Sticking In My Head

I keep coming back to this.

If smaller models can handle real tasks locally, coding, reasoning, even audio... what happens to cloud dependency?

Like are we slowly moving back to local-first AI?

Because that would flip a lot of things.

Privacy, cost, speed... even how people build apps.

Feels like one of those shifts that doesn't look big at first but then suddenly everything changes around it.

Also... Why Is Nobody Freaking Out About This?

Maybe I'm just in the wrong circles, but this didn't blow up as much as I expected.

And I don't really get why.

Open weights, Apache 2.0, strong performance, multimodal, scalable from phone to server...

That checks a lot of boxes.

Maybe people are just burned out from constant releases.

Or maybe everyone's waiting to see real-world performance before caring.

Fair enough, I guess.

Still... something about this feels like it's gonna matter more in a few months than it does right now.

Curious what you guys think, hype or actually a big deal?

[Video: <https://www.youtube.com/watch?v=5aqF1HVpjdC>]

There's also a lot of talk about dynamic model switching. Using smaller, lighter models when you can and bigger, more expensive models when the smaller model hallucinates, or the prompt requires complex reasoning. Could be big.

MaîtreCornichon - *June 7, 2026, 10:09 am*

The potential for local processing on devices is a game-changer; it could really redefine accessibility in AI.