Qwen 3.8 & Kimi K3: Open in Name, Closed in Practice — Run This Instead (2026)

July 19, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Two trillion-parameter models got open-weight announcements in ten days — Qwen 3.8 at 2.4T and Kimi K3 at 2.8T — and you can't run either one. The benchmark-topping tier is the closed, paid Max API; the open weights are unshipped, undated, or datacenter-only. But the tier you actually run didn't move: Qwen's own history shows when runnable open weights realistically follow a flagship, which labs ship consumer-sized weights day one, and what tops the independent quality charts on a single 24GB card right now. The headlines aren't aimed at your GPU. You don't need them to be.

📚 **Related:** The Open Frontier Left Your Hardware Behind · Qwen 3.6 Local Guide · Best Uncensored Local LLMs · Running 70B Models Locally

Two trillion-parameter “open” models got announced in ten days. You can’t run either one.

Kimi K3 landed on July 16 — 2.8 trillion parameters, open weights promised for July 27. Today, July 19, Alibaba announced Qwen 3.8: 2.4 trillion parameters, open weights “soon.” If you’re on a 3090 or a 4070 or a Mac and you read those headlines wondering when you get to download something, here’s the honest answer up front: not from either of these, not in any form your GPU can load. The runnable question — the only one that matters if you’re the one buying the electricity — has a completely different answer, and it’s a good one. Let me get the news out of the way first, because the gap between what was announced and what shipped is the whole story.

Open in name, closed in practice

Here’s what Alibaba actually posted today, from the official @Alibaba_Qwen account: “Qwen3.8 is launching and going open-weight soon!” With, it says, 2.4 trillion parameters, “second only to Fable 5.”

Read that carefully, because two different things are wearing one name.

The thing that’s **live right now** is Qwen3.8-Max-Preview. It’s a closed, paid API tier — routable through Alibaba’s Token Plan and Qoder, same subscription bucket as qwen3.7-max and

glm-5.2. No model card, no per-token price sheet, no benchmark table, no weights. You can pay to call it. You cannot download it.

The thing that would actually matter to you — open weights — does not exist yet. Not on Hugging Face's Qwen org, not anywhere. The newest open repos there are Qwen3-ASR and an AgentWorld MoE from a few weeks back. There is no Qwen3.8 of any size, and critically, no small variant was even announced. The tweet promises a 2.4T model "soon" with no date. A 2.4-trillion-parameter model is a datacenter object no matter how hard you quantize it — the same shape as Kimi K3's 2.8T, the same shape as Inkling's 975B whose smallest 1-bit quant is 270GB. "Open" in the license sense. Unrunnable in the practical sense.

And "second only to Fable 5"? That's Alibaba's own claim — the tweet literally says "we believe." It isn't a third-party result. Artificial Analysis, the independent index, doesn't list Qwen 3.8 at all. Its current top looks like this: Claude Fable 5 at 60, GPT-5.6 at 59, Kimi K3 at 57, Claude Opus 4.8 at 56. A genuine number-two would sit around 59, above GPT-5.6. Maybe Qwen 3.8 lands there when someone measures it. But we've seen this exact move one generation ago: Qwen 3.7 Max was billed as China's number one, and when Artificial Analysis actually scored it, it came in at 56.6 — fifth. Still strong. Just not what the announcement said.

I want to be precise about the criticism here, because it's easy to overshoot. Qwen builds good models. 3.7 Max at fifth in the world is a real achievement. The problem isn't the models. The problem is a self-reported ranking presented as fact before anyone independent has checked it, attached to weights nobody can download. Claim and delivery, both running ahead of reality.

What Qwen's own history says about your 3.8 weights

Now the useful part, the one you can't get from the horse-race coverage. If you want to know when — or whether — you'll get runnable 3.8 weights, the best predictor is Qwen's own track record. I pulled the actual Hugging Face commit dates for the last two generations.

Qwen 3.5 launched February 16, 2026, with the big open 397B-A17B on day one. The runnable weights followed fast: the 35B-A3B MoE and the 27B dense both hit Hugging Face on February 24, and the 9B — the one that fits an 8GB card — on February 27. Call it eight to eleven days from flagship to something you can actually load.

Qwen 3.6 was even tighter. The 35B-A3B open MoE landed April 15–16, the 27B dense April 21–22, and the closed Max-Preview showed up around April 20 — meaning the open weights arrived alongside the paid tier, the MoE actually a few days ahead of it. No wait at all.

So for two straight generations, Qwen was one of the good actors: runnable open weights, consumer-sized, within a week or two of the headline, sometimes before it. That's the pattern that makes 3.8's silence loud.

Because then came **Qwen 3.7**, and the pattern broke. Qwen3.7-Max shipped in May as a proprietary, API-only model. No open weights ever followed — not a 27B, not a 9B, nothing. I checked again this week: there is still no official open Qwen 3.7 of any size. The open line went 3.6, skipped 3.7 entirely, and now arrives at 3.8 as a 2.4T model with no small variant announced.

So here's the honest forecast, and it's a fork, not a date. If 3.8 follows the 3.5/3.6 cadence, a runnable 27B or 35B-A3B-class open variant appears within roughly one to two weeks and this whole worry evaporates. If it follows the 3.7 precedent — Max tier only, runnable open tier quietly skipped — you get a 2.4T datacenter drop labeled "open" and nothing for your GPU, same as Kimi K3. Alibaba has never pre-committed a date for the small weights, so anyone telling you a specific day is guessing. Watch the Qwen Hugging Face org, not the tweets. Weights or a named small variant is the only signal that counts.

Did they say why?

No. I looked — the announcement thread, the Qwen blog, statements from the team — and Alibaba has given no reason for going Max-first or for the open-weight delay. The tweet just says "soon." I'm not going to invent a motive, and you should be skeptical of anyone who does. We know the pattern (Max-first, twice now). We don't know why, and there's a real difference between reporting the first and guessing the second. For the longer view on where Qwen's open posture may be heading, I dug into that separately in Is Qwen Going Closed?.

There is one piece of documented external context worth knowing — and only as context. On July 7, Reuters reported that China's Ministry of Commerce held talks with Alibaba, ByteDance, and Z.ai about restricting overseas access to the country's most advanced AI models, talks that per the reporting explicitly covered open-weight releases and even unreleased future models. Alibaba was a named participant. That's real, recent, and relevant to open-weight timing across every Chinese lab. It is not a stated cause of anything here: the talks are at the consultation stage, no rules exist, no timeline was given, and Alibaba has not connected them to Qwen's release plans. File it as background, not as the reason for 3.8. The honest position is that we know the pattern and we don't know the motive.

Who actually ships weights you can run

The trillion-parameter labs aren't the whole field. Some labs still ship consumer-sized open weights the day they announce, and it's worth naming them, because it's a real difference in behavior, not vibes.

Google's Gemma 4 is the clean example. It launched April 2, 2026, Apache 2.0, with open weights on Hugging Face the same day — and in sizes that fit one card. The 31B dense runs in about 17–18GB at Q4 on a 24GB GPU. The 26B-A4B MoE (roughly 4B active per token) is lighter still and comfortable on 16GB. No trillion-parameter flex, no "coming soon." Announce and ship, runnable, day one.

The contrast with the other big open releases is instructive. Meta's **Llama 4** was open at launch in April — but Scout (109B total) and Maverick (~400B) are datacenter MoE; neither runs on a single 24GB card in any honest sense. Mistral's newest, **Small 4**, is Apache 2.0 and open, but it's a 119B MoE that only "loads" on 24GB via heavy RAM offload at single-digit tokens per second. The genuinely runnable Mistral is still last generation's Small 3.2 24B dense. So "open at announcement" and "runnable on your hardware" have quietly split into two different things, and only a couple of labs — Google most cleanly, and Qwen itself back in the 3.5/3.6 days — deliver both at once.

That's the real fault line in mid-2026 open AI. Not China versus America. Not Qwen versus Kimi. It's whether a lab ships weights sized for the machine you own, or weights sized for the machine they own.

Run this today

While the trillion-parameter headlines play out, here's what actually fits your GPU right now. None of it is new this week, which is exactly the point — the runnable tier didn't move, so you don't have to wait for anything.

The best open model you can run on a single **24GB** card is still Qwen 3.6-27B dense. On the independent Artificial Analysis Intelligence Index it scores 37 — number one among open models in its class — with Gemma 4 31B behind it at 29. And it's genuinely fast: on my own RTX 3090 the 27B runs around 38 tok/s at Q4 on llama.cpp, and about 2.5x that with DFlash speculative decoding on batch coding and math. If you'd rather have the MoE, our 35B-A3B guide clocks Qwen 3.6-35B-A3B at 101 tok/s at UD-Q4_K_XL (22.4GB) on a 3090, because only 3B of its parameters fire per token. Either one is the "run this instead of waiting" answer.

Stepping down the tiers:

- **16GB:** Gemma 4 26B-A4B (the MoE, ~4B active) is the natural fit, or Qwen 3.6-35B-A3B at UD-Q3_K_M (16.6GB) if you want the Qwen family.
- **12GB:** the smaller Qwen 3.5 dense models — the 9B especially — or Gemma 4's compact variants. Quality per gigabyte here is the best it's ever been.
- **8GB:** Qwen 3.5-9B at Q4 lands around 5–6GB and leaves room for context. For the full tier-by-tier breakdown, including the uncensored/abliterated options, the best uncensored local LLMs guide and the VRAM requirements guide go deeper than I can here.

If you're weighing whether a bigger model is worth the squeeze, the running 70B models locally guide has the honest math on where quantization stops paying off.

The bottom line

Two trillion-parameter models went "open" in ten days and neither one is for you. Kimi K3's weights are promised for July 27 and will need a server rack to load. Qwen 3.8's weights are promised "soon," undated, at 2.4T, with no small variant on the roadmap and a top-of-the-charts claim that no independent index has confirmed. That's not a scandal. It's just a category of release that stopped being aimed at consumer hardware.

The good news is the tier that is aimed at you never slowed down. Qwen 3.6 still tops the independent open-model charts on a 24GB card and runs at 38 tok/s on a used 3090. Gemma 4 ships runnable and Apache-licensed the day it's announced. The headlines will keep chasing the trillion-parameter number, and you can keep ignoring them, because the model that fits your GPU is already downloaded, already fast, and already better than anything you could have run a year ago.

Watch the Qwen Hugging Face org for a small 3.8 variant. If it shows up in a week or two, great — history says it might. If it doesn't, you didn't lose anything. You were never the audience for the 2.4T drop.