# Qwen 3.7's open weights are overdue — by the math, not vibes

June 22, 2026 · by Mark Bartlett

Download this post as PDF

Two things landed this week, and one of them is a non-event that's actually the story: Qwen 3.7's open weights still haven't shipped, and by Qwen's own release math, that's now late — not "any day now" late, measurably-past-the-window late. Plus a new guide on running GLM-5.2 locally, which is a frontier open model and a genuine hardware gut-check.

---

## Qwen 3.7 Open Weights: Overdue by the Numbers

Everyone waiting on the 27B and 35B open weights — the ones that actually run on a 3090 — keeps getting told "two to six weeks, who knows." So we did the math instead of guessing.

Qwen's last two generations shipped open weights on a tight inter-generation cadence: 3.5-open on Feb 24, 3.6-open on Apr 16. That's a ~51–59 day rhythm. Apply it to 3.6-open and you get a **June 6–14 window** for 3.7. Today is June 22. We're past it. There's a hard corroborating signal too: the `QwenLM/Qwen3.7` GitHub repo still 404s, and in past cycles that repo existed at or before the drop.

Now the honest hedge, because this is a projection and not an announcement: Alibaba hasn't confirmed any date, and 3.7 already broke pattern by going Max-first (closed API in May) before any open release. So treat this as cadence math — I'd put it around 55–65% late June, 35–45% slipping into July. But "overdue" is a fact, not a vibe.

We've got a live monitor watching for the moment the weights drop, and our firsthand RTX 3090 benchmark publishes within 24 hours of it.

📖 **Our Qwen 3.7 watch page, with the full cadence breakdown, is here.**

---

## GLM-5.2, and What It Actually Takes to Run It

Z.ai's GLM-5.2 is one of the better open models out right now — roughly 750B parameters as a Mixture-of-Experts, MIT licensed, frontier-class on the leaderboards. So naturally the question is "can I run it at home," and the honest answer is: yes, if you bring real hardware.

This is not a 24GB-card model. Even the aggressive 2-bit quant lands around 240GB on disk, which puts it in 256GB-Mac or big-RAM-server territory, not “load it on your 3090.” I wrote up the full picture — quant ladder, memory math, where the realistic entry point actually sits, and where the API makes more sense than buying hardware.

One straight-up disclosure: this is a guidance piece, not a firsthand bench. Our test rig has 64GB of RAM and physically can’t hold this model, so I’m not going to pretend I clocked tokens-per-second on it. What I can do is the honest math on what it takes.

📖 **The full GLM-5.2 local guide is here.**

---

### Quick Hits

- **Run Qwen 3.6 today.** While we wait on 3.7, the 3.6 line — 27B dense and 35B-A3B MoE — is still the open backbone to actually deploy. Don’t hold your project hostage to a release window. 📖 **Qwen 3.6 guide.**
- **llama.cpp is rolling through the b9750s.** Incremental stuff — sampler efficiency, broader multimodal input handling — nothing that forces an update if you’re on a recent build.
- **The “is Qwen going closed” question is live.** The Max-first, open-weights-later pattern on 3.7 is exactly the shift worth watching. We laid out the case here.

---

## I’ll Be at AI Engineer World’s Fair

Quick personal note before I sign off. I’ll be at the AI Engineer World’s Fair at Moscone in San Francisco, June 29 – July 2. If you’re going to be there, find me and say hi — I’d genuinely love to talk local AI in person rather than over a screen for once.

And if you’re not going but have something on your mind — a weird rig, a benchmark, a disagreement with something I wrote — reply to this, or hit me at hello@insiderllm.com. I read everything.

That’s the week. Next edition lands next week.

— Mark, InsiderLLM

---

Source: https://insiderllm.com/blog/newsletter-2026-06-22/