Qwen 3.7 Open Weights Watch: The June Window Is Closing

May 20, 2026 · Updated: June 19, 2026

Download this guide as PDF

> **Quick Answer:** Open weights aren't out — and by Alibaba's own cadence, they're overdue. The 3.5→3.6 open-to-open lag was 51-59 days, projecting June 6-14 for 3.7. We're now 5-13 days past, and the QwenLM/Qwen3.7 GitHub repo doesn't exist yet. Running Qwen 3.6 today? Stay put. InsiderLLM's firsthand RTX 3090 + 3060 bench publishes within 24 hours of the drop. Closed-tier tail: 3.7-Max May 20 (56.6 AAI, #5 overall and top Chinese model), Qwen-VLA May 29, 3.7-Plus June 1 — all paid endpoints. HF API confirms zero Qwen3.7-* repos as of June 19.

> **Status — June 19, 2026: ⏳ NOT YET RELEASED — and overdue against precedent.** Closed-tier shipments since May 20: **Qwen 3.7-Max** (May 20, AAI v4.0 score 56.6, #5 overall and top Chinese model), **Qwen-VLA** (May 29, robotics), **Qwen 3.7-Plus** (June 1, multimodal agent). All three are paid endpoints with no public weights. InsiderLLM's HF API monitor confirms zero `Qwen3.7-*` repos under the official `Qwen` org as of this morning, and the `QwenLM/Qwen3.7` GitHub repo does not exist yet either.
>
> **Where the cadence puts us:** the inter-generation open-to-open lag (3.5→3.6 = 51-59 days) projected a **June 6-14 drop**. We're now **5-13 days past** that window. Realistic landing zone is **late June through mid-July**, with the June probability declining each day. **Alibaba has confirmed no official date** — this is a precedent extrapolation, not an announcement. Full analysis in the section below.
>
> **InsiderLLM's commitment:** firsthand RTX 3090 + RTX 3060 benchmarks within 24 hours of release. (When the drop happens, this status block flips to the launch-piece cross-link.)

For broader strategic context on Qwen's two-track split (closed frontier vs open mid-tier), see Is Qwen Going Closed? Open Weights vs Frontier (2026).

Qwen 3.7 Max preview landed on Alibaba's API on May 19, 2026, and scored **56.6 on Artificial Analysis's Intelligence Index v4.0** (the headline in this article's title and slug rounds to 57). That's a 4.8-point jump over Qwen 3.6 Max Preview's **51.8**. AA places the new model at #5 of 218 ranked entries on its public leaderboard, the highest-ranked Chinese model and sitting in the global top 5 alongside Claude Opus 4.7 and GPT-5.5. On Arena AI's text leaderboard the model sits at 1,489 Elo, ranked #14 overall.

What InsiderLLM readers actually want to know is different. When do the open 27B and 35B variants drop, will they work with the same llama.cpp / MTP / DFlash pipeline that runs Qwen 3.6 today, and is it worth holding off on a hardware purchase to wait?

This article walks through what's verified about the AAI score, what's been announced for the open weights, and what InsiderLLM still doesn't know. When the open weights ship, the firsthand bench on RTX 3090 plus RTX 3060 will land within 24 hours of release — and the Status block above flips to point at it.

The scoring news

Artificial Analysis published its evaluation of Qwen3.7 Max on May 19, 2026. The headline numbers (source):

- **Intelligence Index v4.0 score: 56.6**
- **Rank: #5 of 218** ranked models on AA, **highest-ranked Chinese model** (sits in the global top 5 alongside Claude Opus 4.7 and GPT-5.5)
- Context window: 1 million tokens
- Modality: text in, text out (no image input)
- Mode: reasoning model only ("This page shows the reasoning version")
- Output tokens generated during eval: 97 million, against a 26 million median for the evaluated set

The 97M output figure is worth a pause. AA describes the model as "very verbose in comparison to the average." Reasoning models trade tokens for accuracy, and Qwen 3.7's verbosity sits on the high end even within that category. That has cost implications for anyone running the Max API at scale, and it has timing implications for anyone planning to run an open variant locally.

The Intelligence Index aggregates ten evaluations: GDPval-AA, τ^2-Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, and CritPt. The 56.6 score reflects performance across that whole suite, not any single eval.

For comparison, Qwen 3.6 Max Preview sits at **51.8 AAI** under the same methodology. That's a **+4.8 absolute jump**, or +9.3% relative. Both Max releases are reasoning-classified hosted models with verbose-output profiles.

On Arena AI's text leaderboard, Qwen 3.7 Max Preview shows 1,489 Elo at rank #14 overall (verified May 20, 2026). Arena Elo is human pairwise preference, a different signal from AA's

aggregate eval score. Both numbers are based on the hosted API model. Neither has any direct relationship to open-weight inference on consumer hardware.

That last point is the whole reason this article exists.

The open-weights situation

Qwen 3.7 ships in two announced tiers:

1. **Closed frontier (Max, Plus, VLA).** Hosted via Alibaba's API. No weights public. Three closed shipments since May 20: **3.7-Max** (May 20, reasoning, AAI 56.6), **Qwen-VLA** (May 29, vision-language-action for robotics), and **3.7-Plus** (June 1, multimodal agent). Following the pattern of the Qwen Max line in 3.5 and 3.6, weights are unlikely to be released for these tiers.
2. **27B and 35B open variants.** Announced as forthcoming. No public release date. **Direct HF API check today (June 19, 2026): zero `Qwen3.7-*` repos under the official `Qwen` org**, and zero matching `*-GGUF` quants under `unsloth`, `bartowski`, or `lmstudio-community`.

Why the distinction matters: the 56.6 AAI score is for the Max model. Local AI readers care about the 27B and 35B. There is no reliable way to predict open-weights quality from Max scoring. Alibaba has not claimed parity. Precedent from 3.6: Max Preview scored 51.8 AAI; the open 27B (reasoning) scored 46 AAI on the same methodology — a ~6-point gap. If 3.7 follows the same pattern, the open 27B lands around 51 AAI. That's essentially tied with 3.6 Max Preview (51.8), well below 3.7 Max (56.6), and still in the strong open-weights tier.

The architecture for the 27B and 35B is unannounced. The following is informed speculation, flagged as speculation:

- The 27B will probably stay dense, inheriting Qwen 3.6's hybrid Gated DeltaNet plus Gated Attention layout.
- The 35B is likely to remain a Mixture-of-Experts model, plausibly carrying forward the 35B-A3B configuration (3 billion active parameters per token). This is a guess, not a confirmed spec.
- MTP layer availability depends on whether Alibaba ships MTP-trained variants the way they did for parts of the Qwen 3.6 family. If they don't, the community would need to train MTP layers from scratch.
- DFlash and EAGLE3 compatibility depends on draft model training. None exists yet for 3.7.
- Expected VRAM range for the 27B at Q4_K_M: roughly 15-16 GiB, similar to 3.6. For 35B-A3B at Q4: roughly 18-22 GiB. Both numbers are extrapolations from 3.6 and should not be quoted as confirmed.

GGUF release timing typically lags weight drop by 24 to 72 hours. Bartowski, Unsloth, and RDson have shipped Qwen quants within that window on previous releases. None of that timing is guaranteed for 3.7.

When Will the Open Weights Drop? (And Why It's Overdue)

The widely-repeated framing on Qwen 3.7 is "Max launched, open weights follow weeks later." The verified dates show the opposite: in both Qwen 3.5 and 3.6, **open weights shipped before Max** — by 29 days and 5 days respectively. Qwen 3.7 inverted that pattern on May 20 by launching Max first. The "Max → open lag" lens points the arrow backwards.

Generation	First open release	Max release	Ordering
Qwen 3.5	Feb 24, 2026 (27B, 35B-A3B, 122B-A10B)	March 25, 2026 (LM Arena only — no GA launch)	**Open first by 29 days**
Qwen 3.6	April 16, 2026 (35B-A3B)	April 20, 2026 (Max-Preview)	**Open first by 5 days**
Qwen 3.7	**— (not yet)**	May 20, 2026	**Max first — pattern inverted**

Open-weight dates: QwenLM/Qwen3.5 and QwenLM/Qwen3.6 GitHub README News sections — Qwen's own release announcements. Max dates: qwen.ai/blog and Apsara Summit coverage.

The cadence the data actually supports is **inter-generation, open-to-open**. From Qwen 3.5's first open release (Feb 16, 2026) to 3.6's first open release (April 16, 2026) was 59 days. Same-architecture comparison (Qwen 3.5-35B-A3B → Qwen 3.6-35B-A3B): 51 days. **Range: 51-59 days.**

Applying that range to 3.6's first open (April 16, 2026) projects Qwen 3.7's first open weights at **June 6 to June 14, 2026**. **Today is June 19. The drop is 5-13 days past that window — overdue by the precedent.**

Four honest hedges balance the call:

1. **3.7 broke the 3.5/3.6 ordering pattern** (Max first). The inter-gen cadence is being extrapolated across a broken pattern, not a fitted curve — useful as a marker, not a guarantee.
2. **The `QwenLM/Qwen3.7` GitHub repo doesn't exist yet** (a direct fetch returns HTTP 404). In both prior cycles that repo existed at or before the open-weights drop with its News section pre-populated. Its current absence is a soft "not days-away" signal.
3. **Alibaba changed cycle shape two generations running.** They skipped Qwen 3.5-Max's GA launch entirely (LM Arena debut only) and then inverted 3.7 to Max-first. "The cadence will

hold" is a weaker assumption against two consecutive cycle-shape changes than a single-precedent extrapolation would be.

4. **Alibaba has confirmed no official date.** The projection above is a precedent extrapolation, not an announcement. Worth restating because the rest of the analysis can read like commitment if this line is missed.

Honest probability split: **~55-65% the drop lands in late June, 35-45% slips to early or mid July**, with the June probability declining each day past the window. InsiderLLM's HF monitor polls every 10 minutes for new `Qwen3.7-*` repos under the official Qwen org and the three quantizer accounts (`unsloth`, `bartowski`, `lmstudio-community`). First sight triggers a Telegram alert and starts the firsthand-bench publication clock.

What this means for local AI buyers

Four scenarios with clear guidance.

Already running Qwen 3.6 27B or 35B on a 24GB card. Stay put. The stack works, MTP is mainline-bound through PR #22673, and the DFlash plus MTP comparison on the same hardware documents 1.5x to 2.5x speedup options today. Wait for weights, wait for benches, then re-evaluate.

Planning a hardware purchase soon (RTX 4090, RTX 5090, Mac Studio). No reason to delay. 24GB runs the 27B comfortably. 32GB and up handles the 35B with room to spare. The hardware decision is independent of which Qwen generation a buyer targets, because consumer-GPU VRAM ceilings move slowly compared to model releases.

Running Qwen 3.5 or older locally. The 3.6 family is the more practical upgrade. It's available now, has 60 tok/s MTP on RTX 3090, and full coverage of the 27B dense vs 35B MoE choice is already published. Skipping 3.6 to wait for 3.7 means leaving real performance on the table for an unknown release window.

API users evaluating Qwen 3.7 Max. That is a different question than local AI. The hosted Max preview competes with Claude Opus, GPT-5.5, and DeepSeek R1 on reasoning quality and per-token cost. InsiderLLM's domain is local hardware. Cloud comparisons are out of scope here.

What's still missing

Five concrete unknowns, none of which the launch announcement addressed:

- **Local inference performance.** Zero benches are possible until weights ship. Anyone publishing numbers today is either repackaging API results or guessing.
- **MTP layer availability.** If Alibaba doesn't ship MTP-trained variants, mainline llama.cpp speculative decoding via PR #22673 won't work day-zero on 3.7. The community would have to train MTP layers post-release.
- **GGUF release timing.** Estimated at 24-72 hours after weights drop based on prior cadence. Not guaranteed.
- **DFlash / EAGLE3 compatibility.** Depends on draft model training that hasn't happened yet. The speculative decoding primer covers why draft availability matters.
- **MoE variant configuration.** Whether the 35B is A3B-style, a different MoE shape, or replaced entirely is unknown.

Anyone claiming firm answers on these points today is speculating. InsiderLLM's commitment is to flag the speculation as such until evidence ships.

How InsiderLLM is preparing

When the open weights drop, the editorial plan is:

- **Day 0.** Initial article with download links, model card analysis, build requirements, and setup notes for llama.cpp.
- **Day 1-2.** Firsthand bench on RTX 3090 plus RTX 3060 12GB using am17an's gist harness, the same nine-prompt setup as the existing 3.6 benches.
- **Day 3-5.** MTP compatibility report (does PR #22673 understand the 3.7 layers?) and DFlash compatibility report (does an EAGLE3 draft exist for 3.7?).
- **Week 1.** Head-to-head Qwen 3.6 versus Qwen 3.7 on identical hardware, identical harness.

Reference points readers can use today, all firsthand:

- Qwen 3.6 complete guide
- Wicked Fast Qwen 3.6 27B with MTP on RTX 3090, 60 tok/s on Miu
- Best way to run Qwen 3.6 35B MoE locally
- DFlash vs MTP head-to-head on RTX 3090
- Speculative decoding explained

Where we are today (June 19, 2026)

The Cloud Summit verdict landed in the “likely case” path. May 20 brought the 3.7-Max preview announcement and the 56.6 AAI headline; the open 27B and 35B variants were named but unscheduled. Four weeks later:

- **3.7-Max** continues as a hosted reasoning preview.
- **Qwen-VLA** shipped May 29 — robotics-focused, closed, no local angle.
- **3.7-Plus** shipped June 1 — multimodal agent, closed, no local angle.
- **No 3.7-tier open weights** have appeared. As detailed in the cadence section above, the inter-generation pattern projected a **June 6-14 drop**; we are now **5-13 days past** that window, and Alibaba has confirmed no official date.

Three live signals worth tracking:

- **Cadence projection:** the verified 3.5→3.6 inter-generation lag (51-59 days) points to mid-June. We’re past it. Realistic landing zone is now late June through mid-July, with the June probability declining each day.
- **Closed-tier cadence is strong.** Three closed-only releases in the four weeks after the Max launch is a serious commercial push. Whether the open tier is delayed against the 3.5/3.6 precedent (cadence-honest reading) versus widening structurally behind closed (a stronger claim) is a question this article doesn’t try to resolve — the open weights vs closed frontier piece walks through both reads.
- **InsiderLLM’s HF API monitor polls every 10 minutes** for new `Qwen3.7-*` repos under the official Qwen org and matching `-GGUF` quants under `unsloth`, `bartowski`, and `lmstudio-community`. First sight triggers a Telegram alert and kicks off the bench-and-publish pipeline.

This article updates as the situation moves. The Status block at the top flips to the launch-piece cross-link the moment the open weights ship.