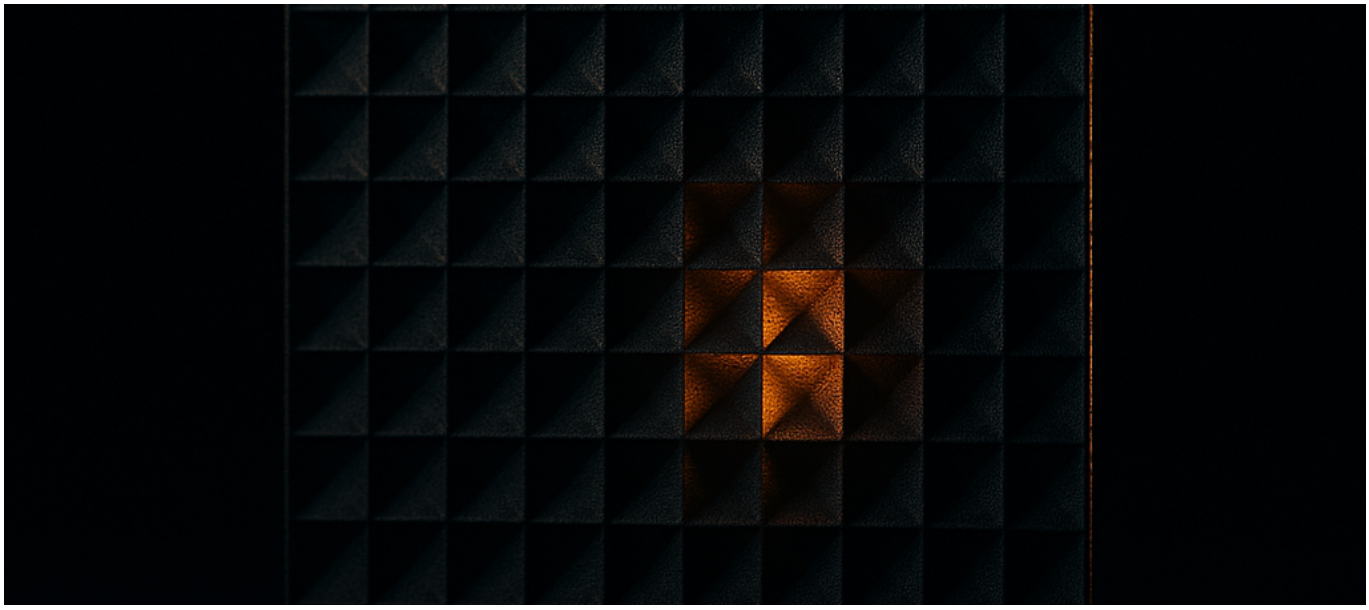

Tencent Open-Sources Hy4 Preview — 770B MoE, 49B Active, 1M-Token Context, Apache 2.0

Tencent shipped a 770-billion-parameter model under a real Apache 2.0 license — no MAU caps, no bespoke community terms. And it published zero benchmark numbers to go with it.

The News

On August 28, 2026, Tencent released Hy4 preview, an open-source model targeting software engineering, research, and financial analysis workloads. The headline specs:

- **770 billion total parameters**, approximately **49 billion active per token**
- **Context window exceeding 1,000,000 tokens** — among the largest of any open-weight model available

- **Apache License 2.0**, with zero MAU thresholds
- **$0.834 / $2.501** per million input/output tokens on Tencent Cloud TokenHub

The official GitHub repository was created on August 27, one day before the public announcement. Weights ship as `tencent/Hy4-preview` on Hugging Face, with mirrors on ModelScope, GitCode and CNB — a distribution pattern that says Tencent expects volume from inside and outside the Great Firewall simultaneously.

Day-one product integration covers WorkBuddy (both Chinese and international editions), CodeBuddy, Yuanbao and ima, with third-party API routing available through OpenRouter.

The license is the part that should stop you. Most Chinese flagship open-weight releases — and plenty of Western ones — arrive wrapped in a custom "community license" that revokes itself above some user threshold. AI Tools Review confirms Hy4 carries no such conditions. Apache 2.0 means you can fork it, fine-tune it, sell access to it, and never call Tencent's legal department.

Why It Matters

The cost floor for long-context workloads just moved, and it moved because of architecture, not subsidy.

At 49B active parameters, Hy4's per-token inference compute sits closer to a mid-size dense model than to anything you'd expect from a 770B system. AIWeekly's analysis makes this point directly: the sparse design compresses the cost floor despite the enormous total capacity. Tencent's $0.834/$2.501 pricing is the visible consequence, but the more important number for CTOs is the one you compute yourself — what does 49B active cost on your own hardware, amortised over your own utilisation?

Who wins: teams with long-context problems and compliance constraints. Legal document review, codebase-wide refactoring, multi-quarter financial analysis — workloads where you need a million tokens of context and cannot send them to a hosted API you don't control. Hy4 is the first model at this scale in that category with a license that survives contact with an enterprise procurement team.

Who loses: vendors whose moat is a permissive-sounding license that isn't. If you're shipping under a restricted community license and charging enterprise rates, Apache 2.0 at 770B is a direct competitive problem. The comparison your prospects will make is not benchmark-to-benchmark. It is lawyer-to-lawyer.

The second-order effect is geopolitical and unavoidable: the ceiling on unrestricted open weights is now set by a Chinese lab. That is a statement about license terms and parameter counts, not about capability. But it changes the shape of the conversation for anyone building on open weights as a strategic hedge.

Technical Depth

The architecture, per the MindStudio breakdown and Tencent's spec sheet, is a 78-layer transformer with a specific asymmetry:

- **Layer 1:** dense feed-forward network
- **Layers 2–78:** MoE layers, 77 of them
- **Each MoE layer:** 256 routed experts + 1 shared expert
- **Routing:** top-8 routed experts per token, plus the shared expert always active

That gives roughly 6–7% effective sparsity. The dense first layer is a deliberate choice — it gives every token a common representation pass before routing kicks in, which tends to stabilise training in very sparse regimes. The persistent shared expert serves a related purpose: it absorbs the general-purpose computation that would otherwise force the router to waste capacity duplicating basics across all 256 experts.

The speculative decoding module

Hy4 ships with a separate multi-token prediction module — roughly 10B total parameters, 0.7B active — used exclusively for speculative decoding. This is not part of the main forward pass. It is a draft model bolted on to reduce latency, and at 0.7B active it is cheap enough to run alongside the main model without meaningfully changing your memory budget.

Tencent also shipped an FP8 quantized variant. Between FP8 weights and the MTP draft

head, Tencent is signalling that they expect people to actually self-host this thing, not just admire it on Hugging Face.

What we don't have

There are no published numeric benchmarks. No GSM8K, no MATH, no HumanEval, no MBPP, no MMLU tables. Coverage across Reuters, AIWeekly and Memujo is uniformly qualitative and vendor-framed. Analysts place Hy4 "alongside" the largest open-weight systems from DeepSeek and Zhipu AI and describe it as Tencent's largest and strongest open-weight model to date — a comparison of scale and a statement about Tencent's internal ranking, not a measured capability claim.

This is becoming a pattern rather than an accident. I noted the same gap when Alibaba launched Qwen3.8-Max at 2.4 trillion parameters with zero official benchmarks three weeks earlier. Parameter counts have become the marketing surface; evaluation has become optional.

The contrast is instructive: when LG AI Research shipped K-EXAONE 2.0 at 750B, they led with long-context benchmark comparisons. Similar parameter class, entirely different disclosure posture.

The Contrarian Take

My take: the license is the release. Everything else is table stakes.

Most coverage is leading with 770B and the million-token context. Both are impressive and both are, at this point in 2026, unremarkable as differentiators. DeepSeek, Zhipu AI, Alibaba Qwen and Moonshot Kimi are all shipping in this race — the field is crowded, and Kimi's K2 series already holds a reputation as the agentic/coding specialist. Another large MoE with a long context window is a Tuesday.

Apache 2.0 at this scale is not a Tuesday. The restricted community license has been the standard defensive move: publish weights, capture mindshare, retain the option to monetise anyone who gets big. Tencent gave that option up. My reading is that this is a deliberate trade — legal frictionlessness in exchange for becoming the default substrate

that other people's products get built on, with Tencent Cloud positioned as the path of least resistance for inference.

Second: **the absence of benchmarks should lower your confidence, not just your patience.** A lab that has strong numbers publishes strong numbers. The qualitative framing across every outlet covering this release — "aimed at software engineering, research and financial analysis tasks," "alongside the largest open-weight systems" — is what you write when you cannot yet write "beats X on Y." The word "preview" in the model name is doing real work here.

That is not an accusation of weakness. It is an observation that we have zero third-party evidence either way, and the burden of proof sits with the vendor.

Third: the $0.834/$2.501 pricing is Tencent's own hosted API, and Tencent Cloud has every incentive to price aggressively at launch. Do not treat it as the true economic cost of running Hy4. Treat it as the number you have to beat if you self-host — and a 770B weight footprint is a memory problem before it is a compute problem.

Practical Implications

If you are evaluating this seriously, run your own benchmarks. There is no alternative, because none exist publicly.

Concrete steps, in order of cost:

- **Cheapest test:** route through OpenRouter or Tencent Cloud TokenHub and run your actual production prompts. At $0.834 per million input tokens, a meaningful eval on your own data costs less than an afternoon of engineering time.
- **Test the context window specifically.** A stated 1M-token window and a usable 1M-token window are different things, and no published long-context results exist for Hy4. If long context is why you're interested, verify that first.
- **Before planning self-hosting, do the memory arithmetic.** All 770B parameters must be resident even though only 49B activate per token. The FP8 variant exists precisely because of this. Your bottleneck is memory capacity, not FLOPs.
- **Have legal read the license.** Not because it's suspect — Apache 2.0 is Apache 2.0 —

but because if you have previously rejected Chinese open-weight models on license grounds, that specific objection no longer applies and your policy may need updating.

For teams already running a restricted-license open-weight model in production: Hy4 is worth a comparison run purely on licensing risk reduction, independent of capability. Trading a small capability delta for the elimination of a MAU cliff is a good trade for most products.

Forward Look

I expect independent benchmarks within 30 days of the preview tag being dropped, and I expect them to place Hy4 competitive-but-not-leading against DeepSeek and Kimi K2 on coding tasks. That is a prediction, not a report. The reasoning: 49B active parameters is a real constraint, and Kimi's K2 series has already established itself as the agentic/coding specialist in this field.

I also expect the Apache 2.0 decision to be copied. Once one lab at this scale demonstrates that giving up license leverage buys distribution, the restricted-community-license posture becomes harder to defend internally at every competitor. My prediction: at least one other major Chinese lab ships a 500B+ model under a genuinely permissive license within six months.

Worth watching: whether Tencent maintains Apache 2.0 for the non-preview release. "Preview" gives them a clean exit. Nothing in the current announcements commits them to the same terms at general availability.

The pricing floor is the piece I'd watch most closely. If $0.834/$2.501 holds and Hy4 turns out to be genuinely competent at long-context code work, it puts pressure on every hosted API in that segment — not through better capability, but through a sparse architecture that makes 770B parameters cost roughly what 49B costs to serve.

The license, not the parameter count, is what makes this release matter — and there are still no published benchmarks to tell us whether the model deserves the attention its terms have earned.

WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting