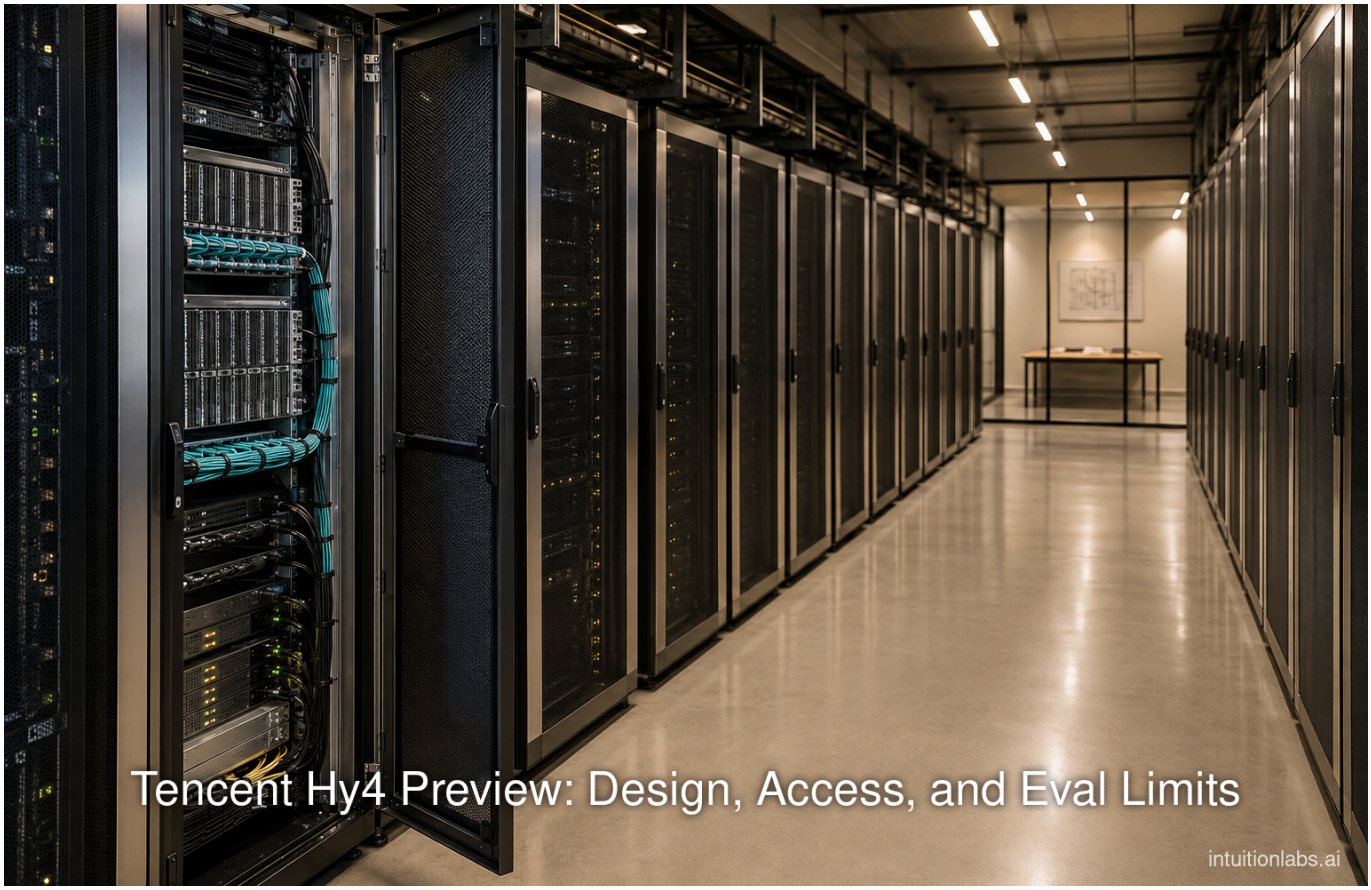
Tencent Hy4 Preview: Design, Access, and Eval Limits

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · tencent hy4 · hunyuan · hy4 preview · open source llm · mixture of experts · chinese ai models · vllm deployment · ai model licensing



Executive Summary

Tencent released **Hy4 preview**, an open-weight large language model developed by the Tencent Hy Team, on **August 28, 2026** (www.tencent.com). The model is a **Mixture-of-Experts (MoE)** design comprising **770 billion total parameters**, of which only **49 billion are activated per token** (huggingface.co) (www.reuters.com). It is released under the **Apache License 2.0**, with weights published as both a standard and an FP8-quantized checkpoint on Hugging Face, ModelScope, GitCode, and CNB (huggingface.co).

Architecturally, Hy4 preview uses a mixture-of-experts backbone, detailed in full in the next section, that routes each token through a small fraction of its total parameters. It supports a context window of roughly **1 million tokens** and includes a built-in multi-token prediction layer for speculative decoding, drawing on a custom sparse-attention mechanism (openrouter.ai). This represents roughly a **2.6-fold** parameter increase and a **4-fold** context-length increase over Hy4's immediate predecessor, Hy3, which shipped on July 6, 2026 with 295 billion total and 21 billion active parameters (www.scmp.com) (recipes.vllm.ai).

Tencent's blind comparative evaluation was conducted internally; the detailed results and methodological limitations appear in "Evaluation Methodology and Limitations."

Access is available through self-hosted inference with vLLM or SGLang, Tencent Cloud TokenHub API, and third-party routing via OpenRouter, at a published price of **$0.834** per million input tokens and **$2.501** per million output tokens (www.tencent.com) (openrouter.ai). Hy4 preview sits within a fast-moving field of large Chinese open-weight releases in mid-2026, alongside DeepSeek-V4, Alibaba's Qwen3.8-Max, Zhipu's GLM-5.3, and Moonshot's Kimi K3, several of which are substantially larger in raw parameter count but carry more restrictive licenses (github.com). This report documents the preview's architecture, licensing, deployment paths, and the specific limitations of its published evaluation methodology, treating it as a checkpoint on the way to a future final release rather than as Tencent's finished flagship.

Introduction and Background

Tencent's Hy Team has released a preview build of its next flagship large language model, referred to across official channels as **Hy4 preview**, as detailed in the summary above. Tencent describes the release as a Mixture-of-Experts flagship model, an architecture in which a router selects a small subset of specialized sub-networks ("experts") to process each token rather than passing every token through the model's entire parameter set. This design is now standard among large open-weight releases from Chinese labs, including DeepSeek, Alibaba, Zhipu, and Moonshot, because it allows a very large total parameter count while keeping the compute cost of each inference call closer to that of a much smaller dense model.

The "preview" label is deliberate and consistent with Tencent's prior release pattern: the company's Hugging Face model card states plainly that Hy4 is "an early version" with "real headroom left in both pre-training and post-training," language it also used for the prior Hy3 preview checkpoint (huggingface.co). Tencent has confirmed that "the next batch of models in the Hy4 series is expected to roll out soon," meaning the preview evaluated in this report is explicitly a precursor to further Hy4-family releases, not the finished product (www.tencent.com).

This matters for how the model's evaluation results should be read. This report documents what Tencent, independent press, and the developer community have verifiably published about Hy4 preview's design, licensing, benchmark evidence, and access paths as of **September 5, 2026**, and it is explicit throughout about which claims are self-reported by Tencent versus independently corroborated.

Product and Platform Architecture

Hy4 preview's technical specification, as published on its Hugging Face model card, describes a 78-layer transformer backbone in which the first layer is a conventional dense feed-forward network and the remaining 77 layers are Mixture-of-Experts (MoE) layers, each containing 256 routed experts plus a single shared expert (huggingface.co). For every token processed, the model activates the top 8 routed experts together with the shared expert, which is how a 770-billion-parameter model keeps its active compute footprint down to 49 billion parameters per token (huggingface.co).

Two architectural features distinguish Hy4 preview from a conventional dense transformer or a simpler MoE design. First, its attention module uses **Gated DeepSeek Sparse Attention (Gated DSA)** with an "IndexCache" mechanism for reusing sparse attention indices across layers, an efficiency technique intended to keep long-

context inference tractable (huggingface.co). Second, its residual pathway uses "identity Hyper-Connections" (iHC), described by Tencent as a technique to expand inter-layer information flow beyond what a standard residual connection provides (huggingface.co). The model also includes one native **multi-token prediction (MTP)** layer of 10 billion total and 0.7 billion active parameters, built in specifically to support speculative decoding at inference time (huggingface.co).

Table 1 below summarizes the published specification sheet, drawn primarily from Tencent's Hugging Face model card (huggingface.co), alongside a few figures that require light interpretation or independent cross-checking.

Specification	Value	Source note
Total parameters	770 billion backbone; 10 billion MTP layer	The model card says its specification table lists backbone parameters only and separately identifies a native 10B MTP layer; the repository metadata's 780B figure should therefore not be presented as a discrepancy (huggingface.co)
Active parameters per token	49 billion	Confirmed independently by Reuters (www.reuters.com)
Backbone layers	78 (1 dense + 77 MoE)	Per specification sheet cited above
Routed experts / activated per token	256 routed + 1 shared / top-8 + shared	Per specification sheet cited above
Context length	Approximately 1,048,576 tokens ("1M" on the model card)	OpenRouter's listing states the precise figure of 1,048,576 tokens (openrouter.ai)
Hidden size / attention heads	6,144 / 64	Per specification sheet cited above
Vocabulary size	120,832 tokens	Per specification sheet cited above
MTP layer	10B total / 0.7B active	Per specification sheet cited above
License	Apache License 2.0	(github.com)

The model card labels the 770B specification table as backbone-only and separately identifies a 10B MTP layer, so the 780B Safetensors metadata should not be presented as a discrepancy. Recommended inference settings on the model card default to a temperature of 0.9 and top-p of 1.0, with the reasoning mode defaulting to "high" (deep chain-of-thought) unless a developer explicitly disables it (recipes.vllm.ai).

Use Cases and Functional Capabilities

Tencent positions Hy4 preview around four workload categories: software engineering, office and document work, game development, and scientific research (huggingface.co). For software engineering specifically, the model card emphasizes long-horizon planning, debugging, and self-verification of multi-step coding tasks rather than single-shot code generation alone. For office and productivity use, Tencent's press release describes covering the "full workflow from information processing through to the creation of documents, spreadsheets, and presentations," positioning the model as an engine behind end-to-end document work rather than a narrow drafting assistant (www.tencent.com).

Tencent also highlights a game-development use case in which "Hy4 preview can generate a playable prototype from a single natural-language request," illustrating an agentic, multi-file code-generation capability rather than isolated snippet completion (www.tencent.com). These use cases are consistent with the model's built-in tooling: the deployment recipe published for vLLM references dedicated tool-call and reasoning parsers (`hy_v4`) tuned for agentic, tool-using workflows rather than plain chat completion.

Distribution runs through two channels that differ in who bears the compute cost. Inside Tencent's own product ecosystem, Hy4 preview is reachable through the coding assistant **CodeBuddy**, the productivity assistant **WorkBuddy**, and the consumer assistants **Yuanbao** and **ima**, and Tencent made it free to use on WorkBuddy and CodeBuddy for two weeks after launch while separately extending free access to the older Hy3 model through September 30, 2026 (www.tencent.com). Outside that ecosystem, developers can self-host the open weights or call Tencent's own API. One methodological wrinkle worth flagging for enterprise buyers: Tencent states Hy4 preview was involved in "the automated optimization of training methods, data strategies, evaluation frameworks" used to build it, meaning parts of its own development and evaluation pipeline were themselves touched by the model under review, a detail relevant to how independently its self-reported scores should be read (www.tencent.com).

Availability, Deployment, and Market Adoption

Hy4 preview's weights are published in two forms, a full-precision **Hy4-preview** checkpoint and an **Hy4-preview-FP8** quantized checkpoint, both distributed through Hugging Face, ModelScope, GitCode, and CNB (huggingface.co). For self-hosted inference, Tencent's official vLLM deployment recipe describes Hy4 preview as a "scaled-up MoE language model (770B total / 49B active) with a 10B MTP layer" and requires **vLLM 0.29.0 or later**; the reference launch command uses 8-way tensor parallelism with the FP8 checkpoint and speculative decoding enabled through the built-in MTP layer (recipes.vllm.ai). Tencent recommends the same temperature=0.9, top_p=1.0 defaults for vLLM deployments as for its hosted API (recipes.vllm.ai). An alternative deployment path uses SGLang via Tencent's official prebuilt multi-architecture Docker image, `lmsysorg/sglang:hy4-preview`, supporting both x86 and Arm hosts with NEXTN speculative decoding (github.com). For teams that need a smaller footprint, Tencent points to its own **AngelSlim** toolkit for quantization, low-bit compression, and speculative sampling rather than publishing bespoke Hy4 quantization instructions (github.com).

The vLLM recipe documents a Hy4-preview configuration on 8x B300 GPUs with MTP (recipes.vllm.ai). Separately, South China Morning Post reported that Tencent compressed the full 1.5-terabyte Hy4 preview checkpoint into a 214-gigabyte version to lower the hardware bar for local deployment, though the outlet did not specify the quantization method used to achieve that reduction (www.scmp.com).

For hosted access, Hy4 preview can be reached "via API through Tencent Cloud TokenHub and OpenRouter," Tencent's press release states, without requiring self-hosting (www.tencent.com). OpenRouter's own listing independently confirms pricing of $0.834 per million input tokens, $2.501 per million output tokens, and $0.042 per million tokens for cache-hit reads, matching Tencent's stated figures (openrouter.ai). Third-party routing documentation from LiteLLM shows TokenHub exposes both an OpenAI-compatible endpoint and a separate Anthropic-Messages-compatible endpoint, meaning teams already integrated against either API shape can point their existing client code at Hy4 preview with minimal changes (docs.litellm.ai).

On naming lineage, the "Hy" branding is a contraction of Tencent's longer-running **Hunyuan** model family: the vLLM Recipes registry for Tencent's GitHub organization lists Hunyuan-A13B-Instruct, HunyuanOCR, Hy3-preview, Hy3, and Hy4-preview as sequential entries, with Hy3-preview added in April 2026 and the full Hy3 release following on July 6, 2026 before Hy4 preview arrived in late August (recipes.vllm.ai) (www.tencent.com). Tencent describes Hy4 preview as a next-generation large language model with 770B total parameters, 49B active parameters, and a context window exceeding 1M tokens (www.tencent.com).

Hy4 in the Open-Weight Landscape: Comparative Positioning

Hy4 preview is one of several large, open-weight Chinese models released in mid-2026, and its relative position depends heavily on which axis is measured: raw parameter count, active compute per token, license permissiveness, or task benchmark scores. **DeepSeek-V4-Pro**, released April 24, 2026, is considerably larger in total parameters at 1.6 trillion, though it activates a comparable 49 billion parameters per token and, like Hy4 preview, supports roughly a 1-million-token context window (fe-static.deepseek.com) (fe-static.deepseek.com). DeepSeek licenses its open-source weights and code under the **MIT License**, one of the most permissive terms among this peer group (fe-static.deepseek.com).

Alibaba's Qwen3.8-Max is a 2.4-trillion-parameter Mixture-of-Experts flagship model (github.com). **Moonshot's Kimi K3** is larger still, at 2.8 trillion total and 104 billion active parameters (huggingface.co). Its license requires a separate agreement before commercial use only when the licensee or its affiliates operate a defined "Model as a Service" business and their aggregate revenue exceeds $20 million over a consecutive 12-month period; the license excludes certain embedded end-user products from that definition and exempts internal use and access through Moonshot AI's official products or certified inference partners (github.com). Products meeting the separate 100-million-monthly-active-user or $20-million-monthly-revenue thresholds must display "Kimi K3" in their user interface; readers should review the license for their use case (github.com). **Zhipu's GLM-5.3** was made available through Zhipu's own coding subscription service on August 14, 2026, with open-weight publication following roughly two weeks later (mlq.ai), a staggered release pattern distinct from Tencent's simultaneous open-weight launch.

Table 2 below places these five models side by side on the dimensions with clear, dated sourcing.

Model	Total / active parameters	Context window	Open-weight license	Release / preview date
Hy4 preview (Tencent)	770B / 49B	Approx. 1M tokens	Apache License 2.0	Aug 28, 2026 (www.tencent.com)
DeepSeek-V4-Pro	1.6T / 49B	Approx. 1M tokens	MIT License	Apr 24, 2026 (fe-static.deepseek.com)
Qwen3.8-Max (Alibaba)	2.4T total parameters (github.com)	Not independently confirmed in this research	Not assessed in this comparison	Not independently confirmed in this research
Kimi K3 (Moonshot)	2.8T / 104B	Approx. 1M tokens	Custom; a separate agreement is required for commercial use only if the licensee or affiliates operate a defined Model as a	Reported ahead of Qwen3.8-Max, mid-July 2026 (www.marktechpost.com)

Model	Total / active parameters	Context window	Open-weight license	Release / preview date
			Service business and exceed $20M aggregate revenue in 12 consecutive months; internal use and official/certified-partner access are exempt (github.com)	
GLM-5.3 (Zhipu)	Not independently confirmed in this research	Not independently confirmed in this research	Custom license for the flagship checkpoint (differs from GLM-5.2's MIT terms per secondary reporting)	Aug 14, 2026 via coding service; open weights approx. 2 weeks later (mlq.ai)

This comparison does not support a smallest-model conclusion: Tencent's model card says Hy4 preview has 770B total parameters, while Z.ai's GLM-5.3 listing displays a 753B model size (huggingface.co) (huggingface.co). Published parameter figures should be compared cautiously because sources may use different parameter-accounting conventions. The table distinguishes Qwen3.8-Max from the open Qwen3.8-2.4T-A95B model that its official card identifies as the basis for the hosted version; GLM-5.3's precise parameter count could not be independently confirmed on a primary source during this research.

Data Analysis and Evidence

The clearest, most reproducible number in Hy4 preview's public record is its price. Tencent Cloud's TokenHub API and OpenRouter both list identical figures: $0.834 per million input tokens, $2.501 per million output tokens, and $0.042 per million cache-read tokens (www.tencent.com) (openrouter.ai), and Forbes independently reported the input price rounds to "about 83 cents per million input tokens" (www.forbes.com).

Benchmark evidence is more layered, and this report separates vendor-reported figures from independent scrutiny of them. On Tencent's own published leaderboard results, Hy4 preview scores 92.3 on GPQA Diamond, a graduate-level science question set, and 62.9 on SkillsBench v1.1 (huggingface.co). Korean outlet The Elec, reporting on Tencent's own 12-benchmark internal testing, listed generation-over-generation gains against the prior Hy3 model: Terminal-Bench 2.1 rose from 70.8 to 85.4, DeepSWE from 28.0 to 64.3, SWE Atlas Refactoring from 32.9 to 53.3, Toolathlon-Verified from 56.2 to 74.1, APEX-Agents from 24.4 to 37.1, and OneMillionBench-plus-tools from 51.6 to 65.4 (www.thelec.net). Against peer models, the same reporting found Hy4 preview's 82.9 on SWE-bench Multilingual ahead of DeepSeek V4 Pro (77.3), GLM 5.3 (81.3), and Kimi K3 (80.8), but trailing Claude Opus 5 (89.5 and 85.8 across two configurations) (www.thelec.net). Head-to-head against Kimi K3 specifically, Hy4 preview trailed on Terminal-Bench 2.1 (85.4 versus 88.3) and DeepSWE (64.3 versus 67.5, a 3.2-point gap), while leading on SWE-bench Pro (65.7 versus 63.3) (www.thelec.net). South China Morning Post separately reported that on the DeepSWE benchmark, Hy4 preview's 64.3 surpassed Alibaba's Qwen-3.8 Max (56.6) and DeepSeek-V4 Pro (62.7) (www.scmp.com).

Table 3 summarizes the human-preference evaluation Tencent ran to compare Hy4 preview against two named competitors.

Comparison	Hy4 preview average score	Competitor average score	Win / tie / loss
vs. **GLM 5.3**	2.99 / 4.00	2.92 / 4.00	46.8% / 12.8% / 40.4% (huggingface.co)

Comparison	Hy4 preview average score	Competitor average score	Win / tie / loss
vs. **Kimi K3**	2.99 / 4.00	2.94 / 4.00	51.2% / 7.9% / 40.9% (www.tencent.com)

This evaluation involved 163 internal Tencent experts rating outputs across 203 engineering tasks in a blind, side-by-side format (technode.global). The win/tie/loss splits in Table 3 show margins well short of a decisive result: against GLM 5.3, Hy4 preview lost 40.4% of comparisons outright, and against Kimi K3 it lost 40.9%. TechNode's independent coverage of the same figures explicitly cautioned that "those results were produced through an internal evaluation and have not been independently verified" (technode.global).

Beyond model-specific figures, Hy4 preview's release fits a broader open-weight adoption pattern. Hugging Face's own platform report found Chinese-origin models "accounted for the plurality or 41% of downloads" on its platform over the preceding year, with the Qwen family alone spawning more than 113,000 derivative models, more than Google's and Meta's derivative counts combined (huggingface.co) (huggingface.co). The same report found industry's share of overall model development on the platform fell from around 70% before 2022 to roughly 37% in 2025, as independent and unaffiliated developers grew their share correspondingly (huggingface.co). A March 2026 research paper from the U.S.-China Economic and Security Review Commission characterized the strategic backdrop plainly: "China has opted to go all in on an open-source approach to AI. Most Chinese labs publish model source code and weights" (www.uscc.gov), a strategy the same paper frames as constrained by United States export controls on advanced semiconductors but backed by sustained state support (www.uscc.gov). Hy4 preview's simultaneous open-weight release alongside a commercial API is consistent with that broader pattern rather than an outlier within it.

Evaluation Methodology and Limitations

Tencent's own model card is unusually candid about Hy4 preview's shortcomings for a vendor-published document. It names two specific behavioral issues: the model "spending longer than necessary reasoning through complex tasks," and "a tendency to over-verify" its own completed work, both framed as known limitations rather than hypothetical edge cases (huggingface.co). Reuters independently confirmed the same disclosure, reporting that "the early-release model can sometimes take longer than necessary to work through complex questions and may over-verify its own answers" (www.reuters.com).

The evaluation methodology behind Hy4 preview's headline comparative claim, its win/tie/loss performance against GLM 5.3 and Kimi K3, has three specific limits worth stating plainly. First, it is self-reported: Tencent's own press release describes "In a blind evaluation conducted internally by Tencent involving 163 experts and 203 engineering tasks," meaning the raters, the task set, and the scoring rubric were all controlled by the company whose model was being evaluated (www.tencent.com). Second, the margins involved are narrow rather than decisive, as Table 3 shows. Third, the comparative result described here is Tencent's internal evaluation.

Community reproductions that do exist are explicit about their own limits. A GitHub issue on Tencent's official Hy4-preview repository, documenting an independent developer's comparison, describes its own test as "a six-category engineering smoke benchmark, not an independent leaderboard," and states directly that "the sample is too small for broad claims about general capability" (github.com) (github.com). In that specific 36-call test across six task categories, Hy4 preview scored 95 out of 100 against a comparison model's 100, with its single failure being a strict JSON formatting error rather than a reasoning failure, a granular, low-sample-size data point that illustrates the gap

between anecdotal developer testing and a validated benchmark. Separately, community commentary aggregated by Medium noted that Hy4 preview shows clear weak points on benchmarks such as "SWE-Marathon, ProgramBench, HorizonMath, and BrokenArXiv," concluding plainly that "the model is not a universal dominator" across task categories, a useful corrective to headline comparisons that emphasize only the categories where Hy4 preview leads (medium.com) (medium.com).

Taken together, these caveats do not invalidate Hy4 preview's published results, but they do mean the model's benchmark standing should be read as a snapshot of a specific, self-selected internal test suite rather than as a settled, third-party-audited ranking. Readers citing this preview's benchmark numbers should attribute them explicitly to Tencent's own testing, note the "preview" status of the checkpoint, and avoid extrapolating these figures to whatever final, non-preview Hy4 model Tencent eventually ships.

Implications and Future Directions

Hy4 preview's Apache 2.0 license, 49-billion active-parameter footprint, and self-hosting recipes for vLLM and SGLang give organizations an open-weight option to evaluate. Kimi K3's license should be reviewed for the intended use: its separate-agreement condition applies only when a licensee or its affiliates operate the license-defined Model as a Service business and exceed the specified aggregate-revenue threshold, with stated exclusions and exemptions (github.com). At the same time, the narrow, self-reported margins in Tencent's own blind evaluation, together with the known tendency to over-verify that Tencent discloses, argue for organizations to run their own task-specific validation before committing production workloads to this specific checkpoint rather than relying on vendor-published rankings alone (huggingface.co).

For organizations evaluating Hy4 preview, Tencent states that the model is available as an open-source model and can be accessed through Tencent products (Tencent).

Looking ahead, the single most important open question is what Tencent's confirmed "next batch of models in the Hy4 series" will contain and how its benchmark standing compares with this preview's self-reported figures (www.tencent.com). Tencent states that the next batch of models in the Hy4 series is expected to roll out soon, but it does not provide a release date. Independent evaluation platforms testing the preview, rather than relying solely on Tencent's internal comparisons, would materially strengthen the evidence base for any final release.

Conclusion

Hy4 preview is a verifiably real, openly licensed release: a 770-billion-parameter Mixture-of-Experts model, 49 billion parameters active per token, released under Apache 2.0, deployable through vLLM or SGLang, and available via Tencent's own API at a published price of $0.834 per million input tokens. Its architecture, including Gated DeepSeek Sparse Attention, identity Hyper-Connections, and a native multi-token-prediction layer, reflects design choices shared across several 2026-era Chinese open-weight releases rather than a unique approach. Its published benchmark evidence and the limits of Tencent's internal evaluation are detailed in "Evaluation Methodology and Limitations."

Tencent has been explicit that this is a preview checkpoint, not a finished flagship, and has confirmed that further Hy4-series releases are planned. Readers, developers, and other analysts citing this release should therefore anchor any claim to this specific checkpoint and its stated evaluation date, distinguish Tencent's own figures from

independently reproduced ones, and revisit the comparison once a final Hy4 release and any independent benchmark verification become available.

Frequently Asked Questions

What is Tencent Hy4?

Hy4 preview is an open-weight large language model released by Tencent's Hy Team on August 28, 2026, using a Mixture-of-Experts architecture with 770 billion total and 49 billion active parameters (www.reuters.com).

Is Hy4 open weights, and under what license?

Yes. Both a standard and an FP8-quantized checkpoint are published on Hugging Face, ModelScope, GitCode, and CNB under the Apache License 2.0 (huggingface.co).

How can Hy4 preview be downloaded and run?

Weights are downloadable from Hugging Face, ModelScope, GitCode, or CNB; self-hosted inference is supported through vLLM (version 0.29.0 or later, with an official deployment recipe) or through SGLang using Tencent's prebuilt `lmsysorg/sglang:hy4-preview` Docker image (recipes.vllm.ai) (github.com). The official recipe documents a Hy4-preview configuration on 8x B300 GPUs with MTP (recipes.vllm.ai).

How does Hy4 preview compare to earlier Hunyuan models?

It follows Hy3, which released July 6, 2026 with 295 billion total and 21 billion active parameters and a 256,000-token context window; Hy4 preview roughly doubles the active parameter count, more than doubles total parameters, and quadruples the context window to approximately 1 million tokens (recipes.vllm.ai) (www.scmp.com).

Can Hy4 preview's benchmark results be generalized to a future final Hy4 release?

No. Tencent's own model card calls this an early version with acknowledged headroom remaining in both pre-training and post-training, and has confirmed further Hy4-series models are planned; TechNode's independent reporting separately notes the comparative benchmark figures "have not been independently verified" (technode.global). This report's findings apply specifically to the checkpoint dated August 28, 2026 and should be revisited once a final release is documented.
