ChatGPT

Tencent「Hy4 preview」徹底分析：7700億パラメータの実像、性能、再現性、コスト、リスク

エグゼクティブサマリー

Tencentが2026年8月28日に公開したモデルの正式名称は **「Hy4 preview」** であり、公式仕様では **バックボーン7700億（770B）パラメータ、1トークンあたり490億（49B）を活性化するMixture-of-Experts（MoE）モデル、100万トークン級コンテキスト** とされている。さらに推測的デコード用の **MTP（Multi-Token Prediction）層が別に100億パラメータ** あり、Hugging Face上のチェックポイントが約780Bと表示されることと整合するため、「770B」は実質的にバックボーンのみを指す数字である。 [1]
技術的には通常のDense Transformerを巨大化したものではなく、77個のMoE層、256 routed experts、Gated DeepSeek Sparse Attention、IndexCache、4本の残差ストリームを持つiHCなどを組み合わせた、**長文脈・低アクティブ比率・高スループットを狙ったTransformer系MoE** である。 [2]
一方、ウェイト、Apache 2.0ライセンス、推論・ファインチューニング手順はかなり開かれているものの、**事前学習トークン数、データセット一覧、データカットオフ、データ権利処理、学習計算量、詳細なpost-training手法、安全性評価は未公開** であり、「完全に再現可能なオープンソース基盤モデル」というより、現時点では「非常に大規模なApache-2.0オープンウェイト＋推論/FTコード公開モデル」と評価するのが正確である。 [3]
また、TencentはGPQA Diamond 92.3など強い数値を公表しているが、**MMLUとHellaSwagのHy4公式スコアは2026年8月29日時点で確認できない**ため、GPT-4、Llama、Mistralとの一般ベンチマーク上の直接比較はまだ成立しない。 [4]

公式ソースと公開範囲

まず重要なのは、今回リリースされたものが最終版「Hy4」ではなく **preview版** である点である。Tencent自身は2026年8月28日の発表で、Hy4 previewを公開・オープンソース化したと発表し、770B総パラメータ、49Bアクティブパラメータ、100万トークン超のコンテキストを主要仕様として掲げている。 [5]

種別	公式ページ	確認できる内容
Tencent公式発表	Tencent News – Hy4 preview	リリース日、770B/49B、1M+ context、製品展開、価格、社内評価。 [6]
Tencent研究ページ	Tencent Hy Research	「770B、49B active、1M context」、スケーリング方針。 [7]
GitHub	Tencent-Hunyuan/Hy4-preview	README、モデル構造、推論例、finetune、ライセンス、配布先。 [8]
Hugging Face	tencent/Hy4-preview	モデルウェイト、構造、ベンチマーク、推論方法。 [9]
Hugging Face FP8	tencent/Hy4-preview-FP8	FP8ウェイトと推論用モデル。 [10]
ModelScope	Tencent-Hunyuan/Hy4-preview	中国圏向け公式配布ミラー。 [11]

種別	公式ページ	確認できる内容
ライセンス	GitHub LICENSE	Apache License 2.0。 12
ファインチューニング	finetune/README_CN.md	DeepSpeed、LLaMA-Factory、ms-swift、LoRA/full tuning。 13
vLLM	vLLM Hy4 recipe	FP8、Tensor Parallel、MTPなどのサービング方法。 14
SGLang	SGLang Hy4 recipe	GPU別メモリ要件、BF16/MXFP8、分散構成。 15

Tencent公式発表の核心部分は、英語原文で **“770B total parameters and 49B active parameters”** と明記されている。 6 Hugging Faceのモデルカードも同じ770B/49Bを明記する一方、後述するMTP層を別枠としている。 9

ライセンス文書は明快で、冒頭に **“Tencent Hy4 preview is licensed under the Apache-2.0.”** とある。Apache 2.0は、著作権表示・ライセンス文書・変更通知などの条件を守れば、複製、改変、派生物、再配布、サブライセンスを含む非常に広い権利を付与し、特許ライセンス条項も含む。Tencent独自の「非商用限定」条項や一定規模以上の企業に対する追加ライセンス条項は、このHy4 LICENSEには確認できない。したがって、**モデルウェイトについては商用利用可能**と解するのが妥当である。 12

ただし「Apache 2.0で公開」＝「学習過程までオープン」という意味ではない。公開GitHubにはREADME、推論手順、`finetune/`、図表、LICENSE等はあるが、770Bモデルをゼロから再学習するための完全なデータパイプライン、事前学習データ、クラスタ設定、学習ログは提供されていない。したがって、**推論と下流ファインチューニングの再現性は比較的高いが、事前学習の再現性は低い**。 16

技術仕様と学習方法

Hy4 previewの本質は「770Bすべてを毎トークン計算する巨大Denseモデル」ではない。公式モデルカードによると、バックボーンは78層で、第1層だけがDense FFN、残る77層がMoEである。各MoE層には **256個のrouted expertと1個のshared expert** があり、各トークンではrouted expertのうち **Top-8** とshared expertが利用される。これにより保存すべきウェイト総量は770Bと巨大な一方、トークンごとのアクティブパラメータは49Bに抑えられている。 17

仕様	Hy4 preview
バックボーン総パラメータ	**770B**
バックボーンactive parameters / token	**49B**
MTP追加パラメータ	**10B total / 0.7B active**
層数	**78**
Dense FFN層	1
MoE層	77
Hidden size	**6,144**
Attention heads	**64**
Query compression	2,048

仕様	Hy4 preview
KV compression	512
Indexer heads	32
Indexer head dimension	128
Indexer Top-k	2,048
Routed experts / MoE layer	**256**
Shared experts / MoE layer	**1**
Activated routed experts / token	**8**
MoE intermediate size	2,048
Dense FFN intermediate size	18,432
Residual streams	**4**
最大コンテキスト	**約1M tokens**
Vocabulary size	**120,832**

すべてTencent/Hugging Face公式仕様に基づく。 17

「7700億」の数字は何を数えているのか

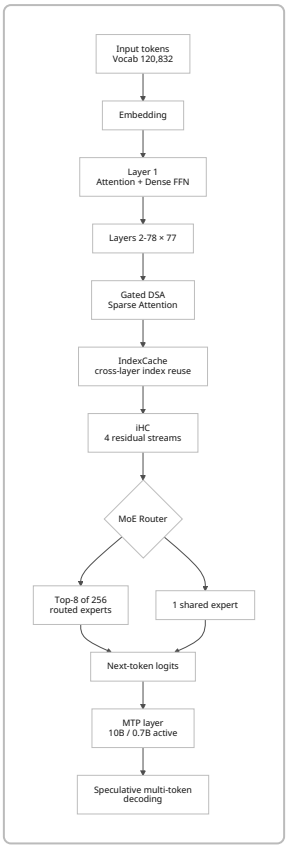
ここには少し紛らわしい点がある。公式アーキテクチャ表の **770Bはbackboneのみ** であり、モデルカードはさらにnative MTP layerについて **10B total / 0.7B active** と記載する。一方、Hugging Face側のチェックポイントメタデータはモデルサイズを約 **780B** と表示している。したがって、

770B backbone + 10B MTP ≈ 780B checkpoint

と理解するのが、公開資料を最も矛盾なく説明する。MTPまで実際に使用する経路では、単純な概念上のactive weightは49B + 0.7B ≈ **49.7B** になるが、Tencentのマーケティング上の「49B active」はバックボーン値である。 9

770Bの大半がMoE expertに存在することも構造から明らかである。77層 × 256 routed experts = **19,712個のlayer-specific routed-expert instances** を持つ一方、1トークンが各層で利用するrouted expertは8個にすぎない。この「巨大な知識容量を保持するが、1トークンで全部を計算しない」という設計が、770Bという総容量と49Bという計算量の差を生み出している。 18

アーキテクチャ概略は次のようになる。



Attention部分も通常のfull self-attentionではない。Hy4は **Gated DeepSeek Sparse Attention（Gated DSA）** を採用し、さらに **IndexCache** により隣接層間で類似するTop-kインデックスを再利用する。IndexCache論文では、DSAのindexer自体が長文脈時に大きな計算コストとなることを指摘し、cross-layer index reuseによってindexer計算の75%を削減し、30B実験では最大1.82倍のprefill高速化、1.48倍のdecode高速化を報告している。Hy4がこの方式を採用したのは、100万トークン級contextを現実的な推論コストへ近づけるためと解釈できる。 2

さらにHy4では **iHC（identity Hyper-Connections）による4本のresidual streams** が使われる。これは従来の「一本のresidual streamを各Transformer blockへ順次通す」構造より複雑であり、Tencentは大規模モデルでの情報伝播・最適化を狙った構成としている。 18

トークナイザーについて確定できるのはvocabulary sizeが120,832であること、チャットテンプレートが `reasoning_content`、thinking/non-thinking、tool call用の特殊形式を扱うことまでである。公式ファインチューニング例は `AutoTokenizer.from_pretrained(..., trust_remote_code=True)` を使用する。SGLang資料にもthinking/tool-call用の構造トークンが記載されている。**BPE、SentencePiece等のアルゴリズム分類については一次資料上で明示確認できず、未確認**である。 19

学習データと学習手法

ここはHy4公開資料の最も大きな空白である。TencentはHy4について、前世代より **model size、context length、training-data volumeを拡大し、pre-trainingとpost-trainingを強化した** と説明している。また、software engineering、gaming、finance、security等のTencent社内専門家と共同で専門データを作成したとしている。 5

しかし、2026年8月29日時点の公開資料からは以下を確認できない。

事前学習トークン総数、Web/書籍/コード/論文/会話データの比率、言語別比率、日本語データ比率、データカットオフ、重複除去率、著作権フィルタ、個人情報フィルタ、合成データ比率は未確認である。したがって、たとえばMetaがLlama 3について15T超のtraining tokensを明示しているレベルのデータ透明性には達していない。 20

同様に、Hy4固有のpost-trainingがSFT、RL、DPO、GRPO、rejection sampling、teacher distillationをどの割合で使用したかも公開資料から特定できない。Tencentが「pre-training」と「substantially larger post-training」を明示していることまでは確認できるが、**「Hy4はRLで学習された」「他モデルから蒸留した」と断定できる一次資料は今回確認できなかった**。特に蒸留は「未確認」と扱うべきである。 7

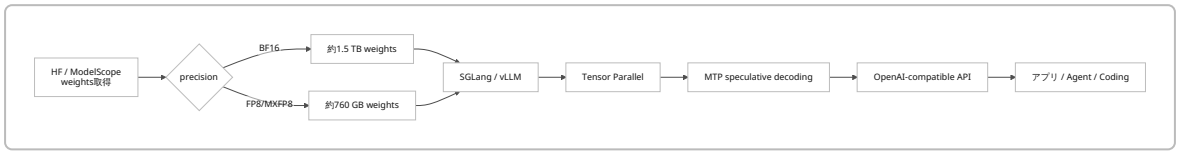
興味深いのはTencentが、Hy4自体が学習方式、データ戦略、評価、低レベルoperator最適化について提案・実験・反復を行い、開発プロセスを補助したと説明している点である。ただしこれは第三者が検証した「recursive self-improvement」の実証ではなく、現段階では **Tencent自身による開発プロセス上の主張** として扱うべきである。 6

実装・再現性・導入

Hy4の「オープンさ」は用途別に評価すると明確になる。**ウェイト入手：高、推論再現性：高、ファインチューニング：中〜高、事前学習再現性：低** である。GitHubには推論例とfinetuneフォルダがあり、Hugging Face/ModelScope等からウェイトを入手できるが、770Bをゼロから作り直せるtraining stack/data recipeは公開されていない。 16

公式に中心となる推論ランタイムは **vLLMとSGLang** である。TencentのREADMEはFP8モデルをTensor Parallel 8でサーブする例を示し、vLLM側にはHy4用recipeが用意され、MTPを用いたspeculative decodingも利用できる。Tencentは独自推論最適化によってbaseline比 **31.8%のend-to-end throughput改善** を達成したとしている。 21

導入の概略は以下の通りである。



SGLangのHy4 deployment recipeは、ウェイトだけで **BF16約1.5TB、MXFP8約760GB** と見積もっている。つまり、一般的な「GPU 1〜2枚でローカルLLM」というカテゴリーではない。 22

GPU構成例	Precision	公式recipe上の構成	例示context
H200 141GB	BF16	**16 GPU = 2×8 nodes、TP16**	131K
B200 192GB	MXFP8	**8 GPU、TP8**	262K
B200 192GB	BF16	**16 GPU**	262K
B300 288GB	MXFP8	**4 GPU、TP4**	262K
B300 288GB	BF16	**8 GPU、TP8**	262K

SGLang公式recipeに基づく。MXFP8はSM100世代以降を前提とするため、H200ではBF16構成が提示される。 15

ここで重要なのは、**「モデルが1M contextをサポート」することと「1M contextを安価にサーブできる」ことは別問題**という点である。公開SGLang recipeの代表構成は131K〜262K程度であり、1Mを実運用するにはさらに大きなKV cache、低い同時実行数、分散・offload等が必要になる可能性が高い。したがって「1M contextの実運用ハードウェア構成・実測性能」は現時点では未確認とするのが安全である。 22

ファインチューニングの現実的な重さ

Tencentは **DeepSpeed/Hugging Face Trainer、LLaMA-Factory、ms-swift** の3系統のファインチューニング方法を説明している。LoRA、full tuning、DeepSpeed ZeRO-2/3、CPU offload、FSDP、gradient checkpointing、BF16等を利用できる。LLaMA-FactoryではSFT、full/LoRAの両方が示され、LoRAではrank 64、alpha 128などのサンプル設定が掲載されている。 13

ただし必要ハードウェアは驚くほど大きい。Tencentがテストした最低ラインは、**LoRAでも8ノード・64 GPU、各GPU 96GB以上、各ノードCPU RAM 2TB以上**。full fine-tuningは **16ノード・128 GPU、各GPU 96GB以上、各ノードCPU RAM 2TB以上** とされる。実際の必要量はsequence lengthとbatch sizeで変わる。LoRAは更新対象パラメータを減らすが、770Bの基礎ウェイトそのものを保持・計算する必要があるため、小規模環境に急に降りてくるわけではない。 13

量子化についてはTencentがAngelSlimを案内しており、低bit量子化や推論最適化の経路が存在する。**分散推論**はTP8/TP16等が明示され、**speculative decoding** は10Bのnative MTP layerで正式にサポートされる。一方、「Hy4を小型モデルへ蒸留する公式pipeline」は今回の一次資料では確認できず、**蒸留サポートは未確認**である。 23

実装面のセキュリティ上は、公式finetune例が `trust_remote_code=True` を使用している点にも注意が必要である。企業環境ではモデルリポジトリのrevisionを固定し、remote codeをレビューしてから実行する方が望ましい。これはHy4固有の悪意を示すものではなく、Hugging Faceのremote-code実行一般に伴うsupply-chainリスクである。 13

性能評価と比較

Tencent/Hugging Faceが公開しているHy4 previewの代表的な性能は以下である。古典的な一般知識ベンチマークより、reasoning・coding・software engineering系に重点が置かれている。 4

公開ベンチマーク	Hy4 preview
GPQA Diamond	**92.3**
DeepSWE	**64.3**
SWE-bench Pro	**65.7**
SkillsBench v1.1	**62.9**
SWE-bench Multilingual Resolved	**82.9**

数値はTencent公式Hugging Face model cardに掲載されたもの。 4

Tencent自身のblind evaluationでは、**163人の専門家・203件のengineering tasks** を用い、4点満点でHy4が **2.99、GLM-5.3が2.92、Kimi K3が2.94** と報告されている。差は小さく、またTencentによる内部評価なので、独立第三者ベンチマークとは明確に区別する必要がある。 6

TencentがGitHubで公開している公式ベンチマーク図も確認できる。[24]

Tencent Hy4 preview公式ベンチマーク図

MMLU・HellaSwagとGPT-4/Llama/Mistral

ここは非常に重要な結論である。**2026年8月29日時点で、Hy4 previewについてTencent公式モデルカード・GitHubからMMLUおよびHellaSwagの数値を確認できなかった。** よって、これらのベンチマークで「Hy4がGPT-4/Llama/Mistralを上回った」とする定量的主張は現時点では裏付けられない。[25]

比較可能な一次資料を並べると次のようになる。

モデル	MMLU	HellaSwag	評価条件・注意
Hy4 preview	**未確認**	**未確認**	Tencent公式資料に数値を確認できず。[4]
GPT-4	**86.4%**	**95.3%**	OpenAI Technical Report。MMLU 5-shot、HellaSwag 10-shot。[26]
Llama 3.1 405B Instruct	**約 87.3%**	—	MetaはMMLUを5-shot/no-CoTで評価し、405BがGPT-4を上回ると報告。HellaSwagは同じpost-trained比較表で直接比較できる値をここでは採用しない。[27]
Mistral 7B Base	**60.1%**	**81.3%**	Mistral技術報告。MMLU 5-shot、HellaSwag 0-shot。[28]
Llama 2 13B Base	**55.6%**	**80.7%**	Mistralが同一pipelineで再評価した値。[28]

この表には**プロンプト数、base/instruct、CoTの有無が異なる**という根本的な制約がある。GPT-4 Technical Report自身もMMLUを5-shot、HellaSwagを10-shotで測定しており、Mistral 7BはHellaSwagを0-shotで測定しているため、数字だけを横一列にして厳密なモデル順位と解釈してはいけない。[29]

さらにGPT-4の報告は2023年、Mistral 7Bは2023年、Llama 3.1は2024年であり、Hy4は2026年のモデルである。MMLU/HellaSwagはもはやfrontier model間の差を測るには飽和・data-contamination問題が大きく、TencentがGPQA Diamond、SWE-bench系、DeepSWEへ比重を移していること自体には合理性がある。ただし、それでも**一般的な横断比較のためMMLU/HellaSwagを公開してほしいという再現性上の要求は残る。** GPT-4とLlama 3の技術報告もbenchmark contaminationを明示的に検討している。[30]

モデルカード自身もpreview版の制約として、**複雑な課題で必要以上に長くreasoningし、過度にverificationする傾向**を認めている。これは推論品質だけでなく、output-token課金とlatencyにも直接影響し得る実運用上の弱点である。[31]

セキュリティ・倫理・ライセンス

Hy4の公開方針で最も強い点はライセンスであり、最も弱い点はモデル安全性・データ透明性である。

ライセンス面では商用利用可能性は高い。 Apache 2.0は複製・改変・派生物作成・配布・サブライセンスを認め、特許条項も含む。再配布時にはライセンスの添付、変更ファイルの通知、著作権・特許・商標・NOTICE表示の保持等が必要で、Tencentの商標権が自動的に与えられるわけではない。またソフトウェアは

無保証で提供される。Hy4固有LICENSEには、Llama型の利用規模制限や禁止用途リストは確認できない。[12]

したがって、たとえば自社SaaSへの組み込み、社内オンプレ運用、追加学習モデルの提供などはApache 2.0上は原則可能である。ただし、**モデルのライセンスが商用利用を許すことと、学習データ・生成物について第三者の著作権、個人情報、営業秘密、各国AI規制への対応が不要になることは別問題**である。特にTencentはtraining corpusの具体的出所や権利構成を開示していないため、規制産業で導入する場合は別途法務デューデリジェンスが必要となる。[32]

データ透明性は低い。 Tencentはsoftware、gaming、finance、security等の専門家がデータ作成に関与したことは明らかにしたが、公開Web、コードリポジトリ、書籍、ニュース、ライセンス済みデータ、ユーザーデータ、合成データ等の内訳を公開していない。個人情報除去、copyright filtering、benchmark decontaminationについてもHy4固有の詳細は確認できない。[5]

バイアス・有害出力の定量評価も不足している。 今回確認したHy4公式model card/repositoryには、ToxiGen、BBQ、RealToxicityPrompts、jailbreak attack success rate、false-refusal rate、CBRN/cyber uplift、multilingual safety等に相当する包括的なsafety reportを確認できなかった。対照的にMetaのLlama 3技術報告はadversarial prompts、false refusals、cyber/chemical-biological uplift、red teamingなどを独立した安全性章で扱っているため、Hy4には今後同程度の公開が望まれる。[33]

Hy4の**高いcoding能力＋tool calling＋100万トークン文脈**は有益である一方、悪用可能性も高める。たとえば脆弱コード生成、malware支援、長いRAG文書中に仕込まれたprompt injection、agent toolの誤実行などである。ウェイトがApache 2.0で完全に手元へ置けるため、Tencent Cloud側の安全フィルタだけで全利用を制御することもできない。これは「Hy4に既知の重大脆弱性がある」という意味ではなく、**強力なopen-weight agentic modelに一般的に付随するリスク**である。[34]

特にSGLang側のHy4 tool-use実装には、tool-choiceの一部が厳密強制されないケースなど実装上の注意事項が記載されている。高権限のエージェントに接続するなら、LLMのtool-call出力を直接実行せず、schema validation、allowlist、least privilege、sandbox、人間承認を別レイヤーに置くべきである。[35]

総じて安全性を表にすると、現時点の公開状態は次のように評価できる。

項目	公開状況	評価
モデルウェイト	公開	高
ライセンス	Apache 2.0明示	高
Architecture	詳細あり	高
Training token数	**未確認**	低
Training corpus内訳	**未確認**	低
データ権利・provenance	**未確認**	低
Bias/toxicity評価	**未確認**	低
Red-team/jailbreak結果	**未確認**	低
Cyber/CBRN uplift	**未確認**	低
Fine-tuning recipe	公開	高

項目	公開状況	評価
推論recipe	公開	高

これは「安全でない」という判定ではなく、**第三者が安全性を判断するための公開証拠がまだ少ない**という評価である。[16]

エコノミクス・エコシステム・初期反応

TencentはHy4を単にHugging Faceへ置くだけでなく、自社の **WorkBuddy/CodeBuddy、Yuanbao、ima** に投入し、Tencent Cloud TokenHubおよびOpenRouter経由でもAPI提供している。ローンチ時にはWorkBuddy/CodeBuddyで2週間無料提供すると発表した。この初日からの「ウェイト＋自社アプリ＋クラウドAPI＋第三者API」の多面的流通は、研究用モデルだけでなく商用プラットフォームとして普及させる戦略を示している。[36]

APIのTencent公式価格は、**input $0.834 / 1M tokens、output $2.501 / 1M tokens、cache hit $0.042 / 1M tokens**。770B総パラメータのfrontier-class modelとしては、自前GPUを常時保有するよりかなり低い参入コストである。[6]

クラウド運用コストの概算

SGLangが示すMXFP8構成の一例は8×B200である。Nebiusの公開価格ではB200がオンデマンド **$7.15/GPU-hour** なので、単純計算すると、

8 × $7.15 = $57.20/時、
$57.20 × 730時間 ≈ $41,756/月

となる。これはGPU rentalのみの概算であり、CPU、RAM、NVMe、ネットワーク、ストレージ、egress、冗長化、監視、エンジニア人件費等を含まない。[37]

BF16をH200で動かすSGLang構成では16 GPUが必要である。AWS Capacity Blocksのp5e.48xlarge（8×H200）は公開価格例で **$47.76/時/8GPU node** なので、2ノードなら、

$95.52/時 ≈ $69,730/月

となる。この場合もcapacity-block価格を単純に24時間×730時間へ換算した比較用概算であり、実際の予約可能性、地域、契約、ストレージ等は別である。[38]

APIとの単純な費用比較を行うため、ワークロードを **input:output = 3:1** と仮定すると、100万総トークンあたりの平均API費は、

0.75 × $0.834 + 0.25 × $2.501 ≈ **$1.25075**

となる。この仮定では、8×B200の$41,756/月に等しいAPI料金へ達するのは約 **334億総tokens/月**、16×H200の$69,730/月では約 **558億tokens/月** である。これはあくまで「支出額の等価点」であり、セルフホスト環境が実際にそのトークン量を処理できることを意味しない。またcache hitが多ければTencent API側はさらに安くなる。[39]

運用形態	概算直接費	メリット	デメリット
Tencent API	$0.834/M input、$2.501/M output	初期投資なし、運用容易	データ外部送信、利用量依存
8×B200 FP8クラウド	約**$41.8k/月**	完全制御、専有環境	高い固定費、運用要員
16×H200 BF16クラウド	約**$69.7k/月**	BF16、既存Hopper資産を利用可	2ノード分散、さらに高コスト
オンプレ	**設備価格は未確認**	データ主権、長期高稼働で有利な可能性	B200/H200サーバ、NVLink/IB、電力・冷却が必要

オンプレの正確なCapExについては、B200/H200 HGX/DGXサーバの企業向け価格が構成・地域・保守契約によって大きく変わるため、今回確認した一次情報だけから信頼できる金額を断定できない。**未確認**としてベンダー見積もりを取得すべきである。一方、必要HBM量だけでもFP8約760GB、BF16約1.5TBなので、小規模ワークステーション市場ではなくデータセンター級投資になることは明白である。 22

競争環境への意味

Hy4の戦略的な重要性は「770Bだから最大」という単純な話ではない。**770Bの知識容量を49B active/tokenまで疎化し、1M context、sparse attention、MTP、FP8、vLLM/SGLangをセットで出す**ことで、モデル品質とserving economicsを同時に競争軸にしている点が重要である。 40

Tencent自身がGLM-5.3とKimi K3を社内blind evaluationの比較対象としていることからも、中国のopen-weight frontier競争における直接的な競争相手がZhipu/GLM、Moonshot/Kimi、さらにDeepSeek、Alibaba/Qwen等になることは自然な推論である。ただし、Hy4の独立評価がほとんど揃っていないリリース翌日の段階で「最高性能」と断定するのは早い。 6

特に **IndexCache採用** は興味深い。Sparse attentionを導入しても、長文脈では「どの過去tokenを見るか」を選ぶindexerそのものが重くなる。IndexCacheは層間でindexを再利用することでそこを圧縮する技術であり、Hy4は単にパラメータ数を増やすだけでなく、「巨大MoEをどう安くサーブするか」をarchitectureの中核に置いている。 41

コミュニティ採用については二極化すると考えられる。Apache 2.0、Hugging Face、ModelScope、vLLM、SGLang、LLaMA-Factory、ms-swiftという標準的なエコシステムへ最初から対応しているため、**API利用・クラウド利用・大規模研究機関での採用障壁は比較的低い**。一方、FP8でも約760GBというウェイト量とLoRAで64×96GB GPUという推奨環境は、個人・小規模研究室のself-host adoptionを強く制約する。したがって、ダウンロード数が増えても、実利用の多くがTencent Cloud/OpenRouter等のhosted inferenceへ流れる可能性が高い、というのが現時点での合理的な予測である。 42

公開デモ・サンプルと初期反応

公式に試せる経路としてTencentは **WorkBuddy/CodeBuddy、Yuanbao、ima、Tencent Cloud TokenHub** を挙げ、OpenRouterからもAPIアクセスできる。したがって、ウェイトを760GB以上ダウンロードしなくても公開サービス経由で評価可能である。 36

一方、**再現可能な「同一prompt→公式出力全文」の体系的sample-output集は、今回確認したGitHub/model cardでは乏しい**。vLLM/SGLang資料にはAPI呼び出しやprompt例があるものの、SGLang側には出力例が未更新の箇所もある。したがって「Tencent公式サンプル出力」を性能証拠として長文引用するより、公

開API上でtemperature、reasoning effort、tool use設定を固定して第三者が再評価する方が有用である。[43]

独立メディアではReutersがリリース直後に、Hy4が770B total / 49B activeのMoEとして公開されたことを報じており、少なくとも「Tencentがfrontier-size open model競争を本格化した」という事実は主要メディアでも確認されている。[44]

初期コミュニティ反応はまだ速報段階である。LocalLLaMAでは、あるユーザーが **“Around glm 5.3 level size and capability”** と評する一方、別の反応ではDeepSWE等でGLM側が強いとの指摘もあり、ベンチマーク解釈に慎重な声が出ている。また「今後半年の競争が面白くなる」という趣旨の反応も見られる。これは専門家査読ではなく、**リリース直後のコミュニティ感想にすぎない**ため、性能評価の根拠ではなく初期センチメントとして扱うべきである。[45]

総じて、2026年8月29日という公開翌日の時点では、**Tencent内部評価と公式benchmarkはあるが、独立した大規模再評価、査読論文、長期運用報告はまだ不足**している。最も興味深い検証ポイントは「770B/49BというMoE構成が、同世代モデルに対して品質だけでなく実トークン/秒・$/tokenでも優位になるか」である。[46]

不明点と追加検証事項

現時点で、以下は特に「未確認」または追加検証が必要である。

- **事前学習データ総量：未確認。** 何兆tokensで学習したのか、Tencentは具体値を公開していない。[5]
- **データセットの出所と権利処理：未確認。** Web、GitHub、書籍、論文、Tencent内部データ、licensed data、synthetic dataの割合、opt-out/copyright処理を確認する必要がある。[6]
- **日本語学習量・日本語品質：未確認。** vocabulary sizeは公開されているが、日本語token比率、日本語MMLU/JMMLU、JHumanEval等の公式評価は追加確認が必要である。[18]
- **Tokenizer方式：未確認。** 120,832 vocabularyとchat/template tokensは判明しているが、BPE/SentencePiece等の詳細仕様をconfig/tokenizer filesレベルで監査する余地がある。[19]
- **post-training recipe：未確認。** SFT/RL/DPO/GRPO/rejection sampling等の組み合わせ、reward model、RL compute、alignment dataset量は公開説明が不足している。[47]
- **蒸留の有無：未確認。** Hy4の学習にteacher model distillationを用いた証拠、およびHy4から小型モデルへ蒸留する公式pipelineは今回確認できなかった。
- **770Bの完全なmatrix-by-matrix parameter accounting：未確認。** 770B backbone＋10B MTPというトップレベル内訳は明確だが、attention、expert、embedding、router、indexer等の正確な個別parameter表は公開資料にない。[9]
- **MMLU/HellaSwag：未確認。** Tencentによる公式値または同一evaluation harnessでの第三者値が必要で、現時点ではGPT-4/Llama/Mistralとの同指標直接比較はできない。[4]
- **1M contextの実運用性能：未確認。** 1M対応はモデル仕様として公表されているが、公開SGLang構成は主に131K〜262Kであり、1MでのGPU数、TTFT、throughput、同時実行数、精度劣化を測る必要がある。[46]
- **量子化品質：追加検証必要。** FP8は公式提供されるが、4-bit等の強い量子化でGPQA/SWE-bench/long-context/tool callingがどの程度維持されるかは独立測定が必要である。[48]
- **安全性：大幅な追加検証が必要。** Bias、toxicity、jailbreak、prompt injection、privacy memorization、cyber uplift、CBRN、多言語安全性について、Tencentによる定量的system card相当資料が必要である。[49]
- **benchmark contamination：未確認。** training corpusが非公開である以上、GPQA/SWE-bench等のtest contaminationを第三者が完全には監査できない。GPT-4やLlama 3が公開したようなdecontamination methodologyの提示が望まれる。[30]

- **オンプレCapEx：未確認。** 8×B200または16×H200級の構成が現実的な基準になるが、日本国内のサーバ本体、InfiniBand/NVLink、保守、電力・冷却を含むTCOはベンダー見積もりで検証すべきである。 22
- **独立評価：今後の最重要事項。** 公開からまだ1日程度であり、同一prompt、同一sampling条件、同一hardwareでGLM/Kimi/DeepSeek/Qwen/Llama/Mistralや主要closed modelsを比較した第三者測定が、Hy4の実際の位置づけを決める。現段階ではTencentの「2.99/4」内部blind評価を独立benchmarkと同等には扱えない。 6

現時点で最も堅い結論は、**Hy4 previewは「770B全部を毎回計算する巨大モデル」ではなく、770BのMoE容量を49B active/tokenへ疎化し、Gated DSA＋IndexCache＋MTPによって1M-token級long-contextを現実的にサーブしようとするモデル**だということである。技術設計とApache 2.0公開は非常に積極的だが、training data、post-training、安全性、一般ベンチマークの透明性はモデル規模に比してまだ不十分である。したがってHy4の本当のインパクトを測る指標は「770B」という見出しの数字そのものではなく、今後の独立評価で **49B-activeあたりの品質、実測throughput、1M-contextでの安定性、$/token、fine-tuning可能性、安全性** がどこまで実証されるかにある。 50

---

1 5 6 20 32 36 39 46 Tencent Releases and Open-Sources Tencent Hy4 preview - Tencent
https://www.tencent.com/tencent-releases-and-open-sources-tencent-hy4-preview/

2 4 9 17 18 25 31 33 34 40 43 49 50 tencent/Hy4-preview · Hugging Face
https://huggingface.co/tencent/Hy4-preview

3 8 16 21 23 GitHub - Tencent-Hunyuan/Hy4-preview · GitHub
https://github.com/Tencent-Hunyuan/Hy4-preview/tree/main

7 47 Introducing Hy4 preview - Tencent Hy
https://hy.tencent.ai/research/hy4-preview?utm_source=chatgpt.com

10 48 tencent/Hy4-preview-FP8
https://huggingface.co/tencent/Hy4-preview-FP8?utm_source=chatgpt.com

11 Model Details
https://modelscope.cn/models/Tencent-Hunyuan/Hy4-preview?utm_source=chatgpt.com

12 LICENSE · tencent/Hy4-preview-FP8 at main
https://huggingface.co/tencent/Hy4-preview-FP8/blob/main/LICENSE

13 19 42 Hy4-preview/finetune/README_CN.md at main · Tencent-Hunyuan/Hy4-preview · GitHub
https://github.com/Tencent-Hunyuan/Hy4-preview/blob/main/finetune/README_CN.md

14 tencent/Hy4-preview | vLLM Recipes
https://recipes.vllm.ai/tencent/Hy4-preview

15 22 35 37 38 Hy4 preview - SGLang Documentation
https://lmsysorg.mintlify.app/cookbook/autoregressive/Tencent/Hy4-Preview

24 Hy4-preview/assets at main · Tencent-Hunyuan/Hy4-preview · GitHub
https://github.com/Tencent-Hunyuan/Hy4-preview/tree/main/assets

26 29 30 GPT-4 Technical Report
https://arxiv.org/html/2303.08774v6

27 [2407.21783] The Llama 3 Herd of Models
https://ar5iv.labs.arxiv.org/html/2407.21783

28 Mistral 7B
https://arxiv.org/html/2310.06825?utm_source=chatgpt.com

41 IndexCache: Accelerating Sparse Attention via Cross-Layer Index Reuse
https://arxiv.org/abs/2603.12201?utm_source=chatgpt.com

44 China's Tencent releases new open-source AI model for ...
https://www.reuters.com/world/asia-pacific/chinas-tencent-releases-new-open-source-ai-model-coding-research-tasks-2026-08-28/?utm_source=chatgpt.com

45 Tencent/Hy4-preview 770B-A49B weight dropped
https://www.reddit.com/r/LocalLLaMA/comments/1w0igxk/tencenthy4preview_770ba49b_weight_dropped/?utm_source=chatgpt.com