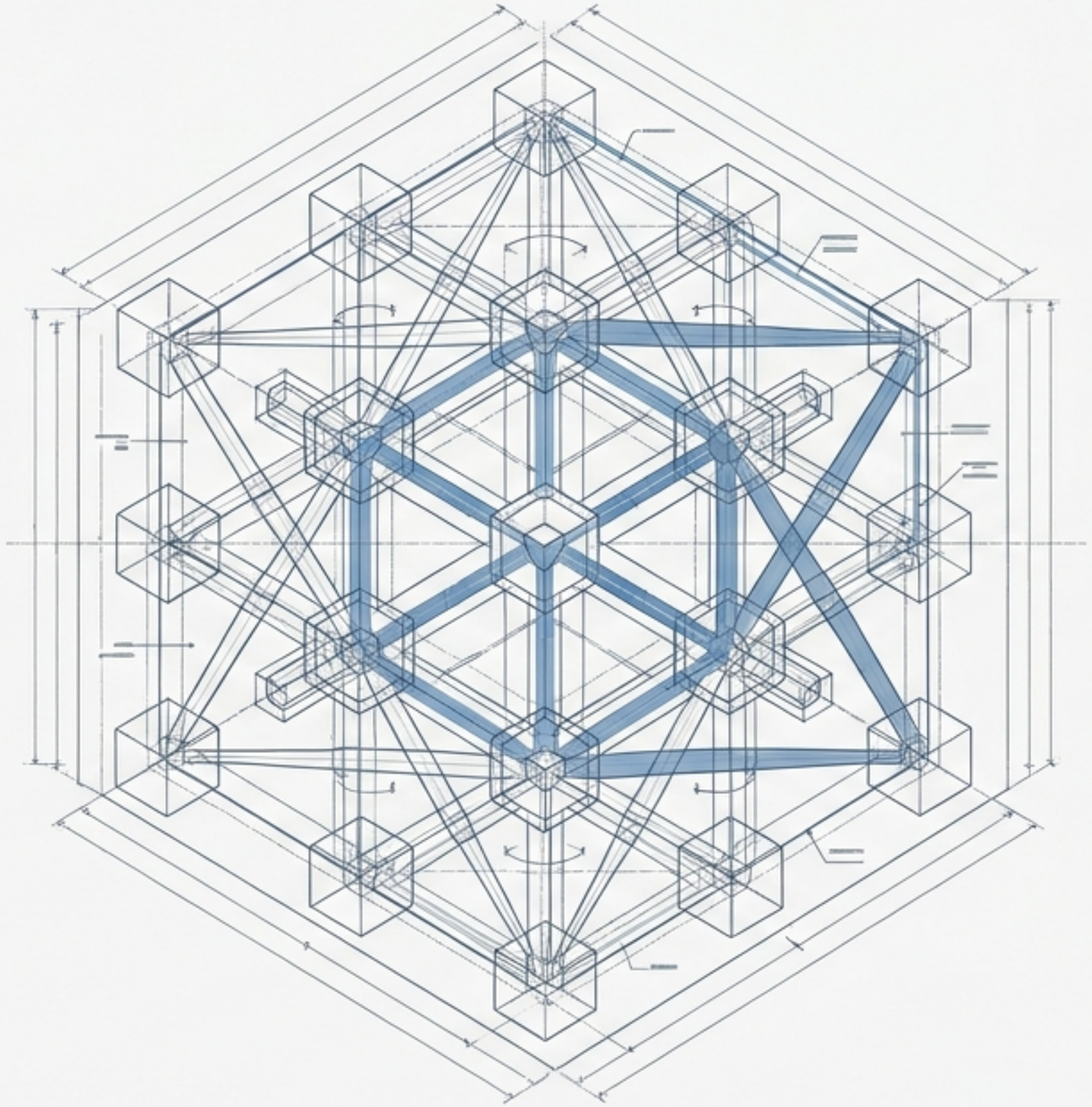


Tencent Hy4 preview 徹底解剖：7700億パラメータの実像と影

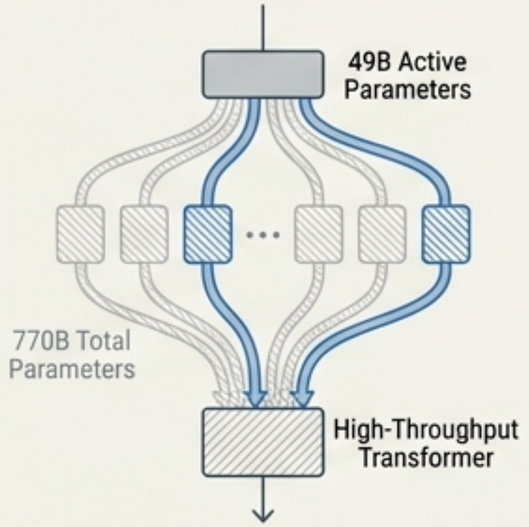
独立した技術検証に基づくアーキテクチャ、デプロイメントコスト、およびデータ透明性の分析ドキュメント。

巨大な野心と残された空白：Hy4 previewの全体像

革新的

アーキテクチャ

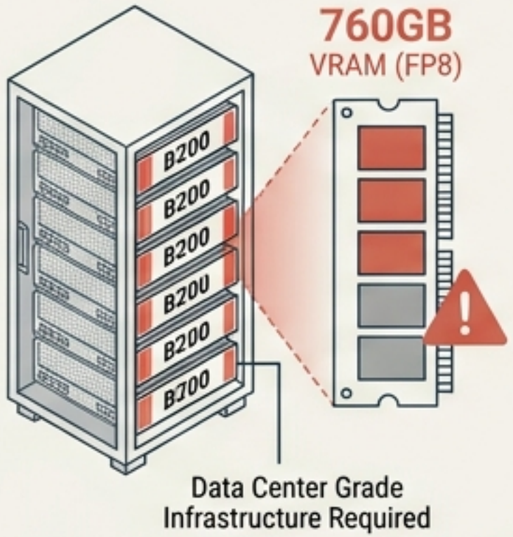
実態は770Bの密なモデルではなく、1トークンあたり49Bのみを活性化する長文脈・高スループット特化型のTransformer系MoE。



要大規模インフラ

デプロイ要件

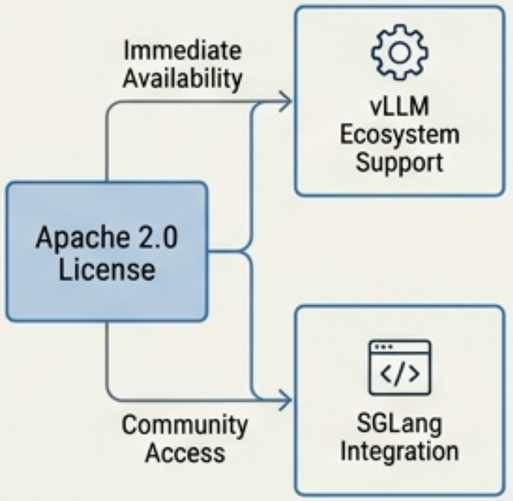
FP8でも約760GB（8x B200相当）のVRAMを要求。「ローカルLLM」の枠を超えデータセンター級のモデル。



高いアクセス性

ライセンスと流通

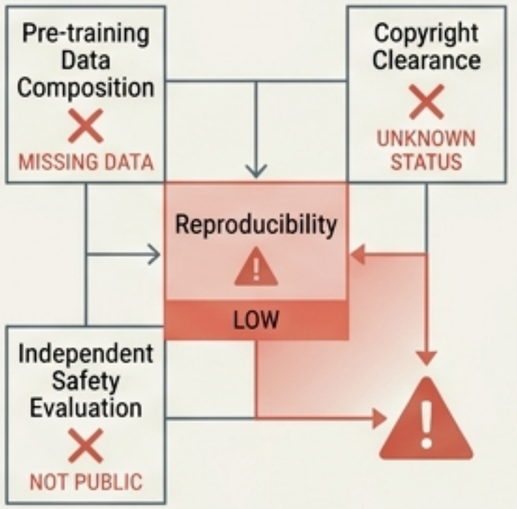
Apache 2.0でウェイトを開放し、vLLMやSGLangのエコシステムに初日から対応。



重大な欠落

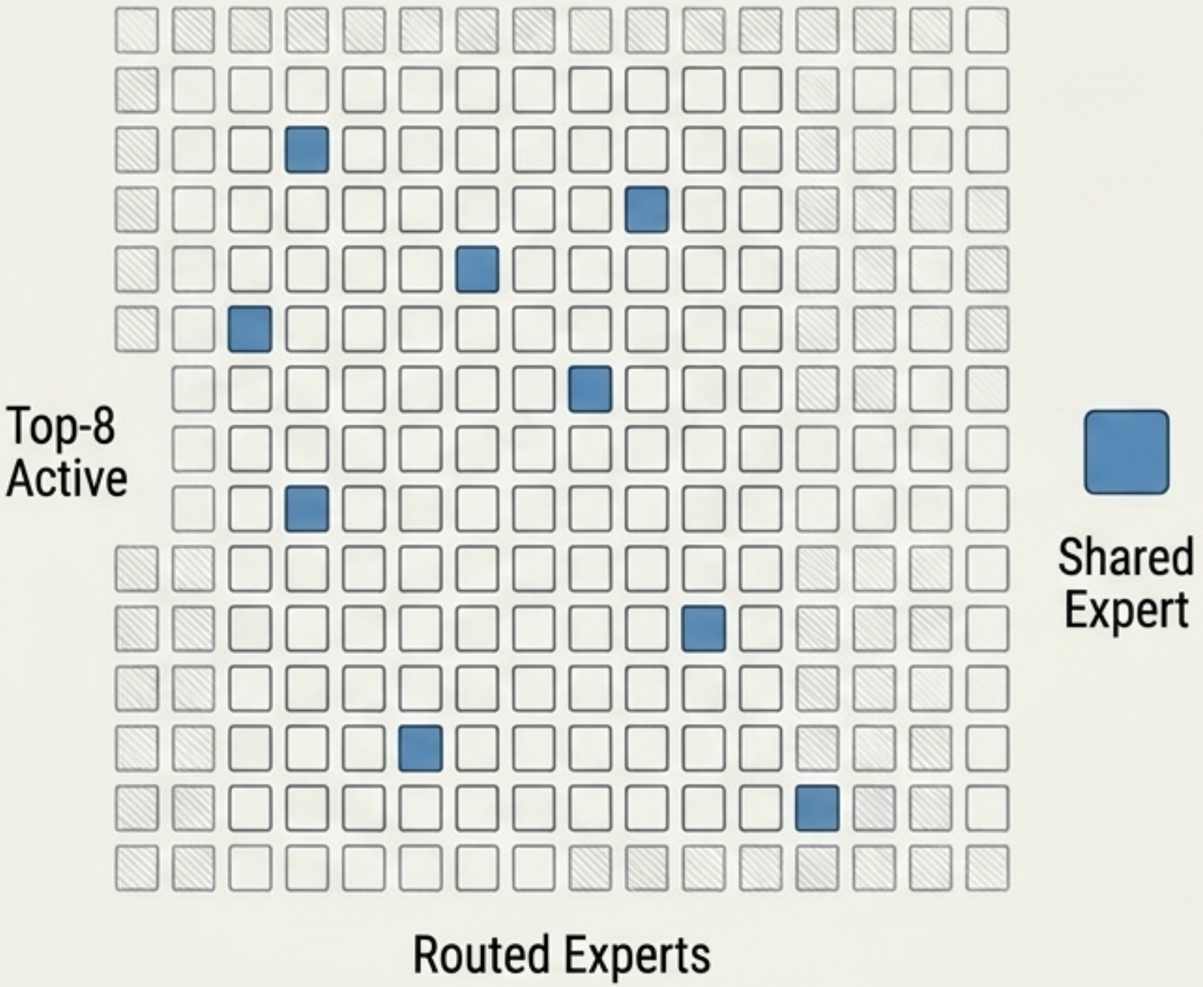
透明性と安全性

事前学習データの内訳、著作権処理、独立した安全性評価の公開がなく、再現性は極めて低い。



「7700億」の錯覚：真の計算コストは490億パラメータ

770Bの総容量は知識の貯蔵庫にすぎず、1トークンの推論に使用されるのは全体の約6%に留まる。

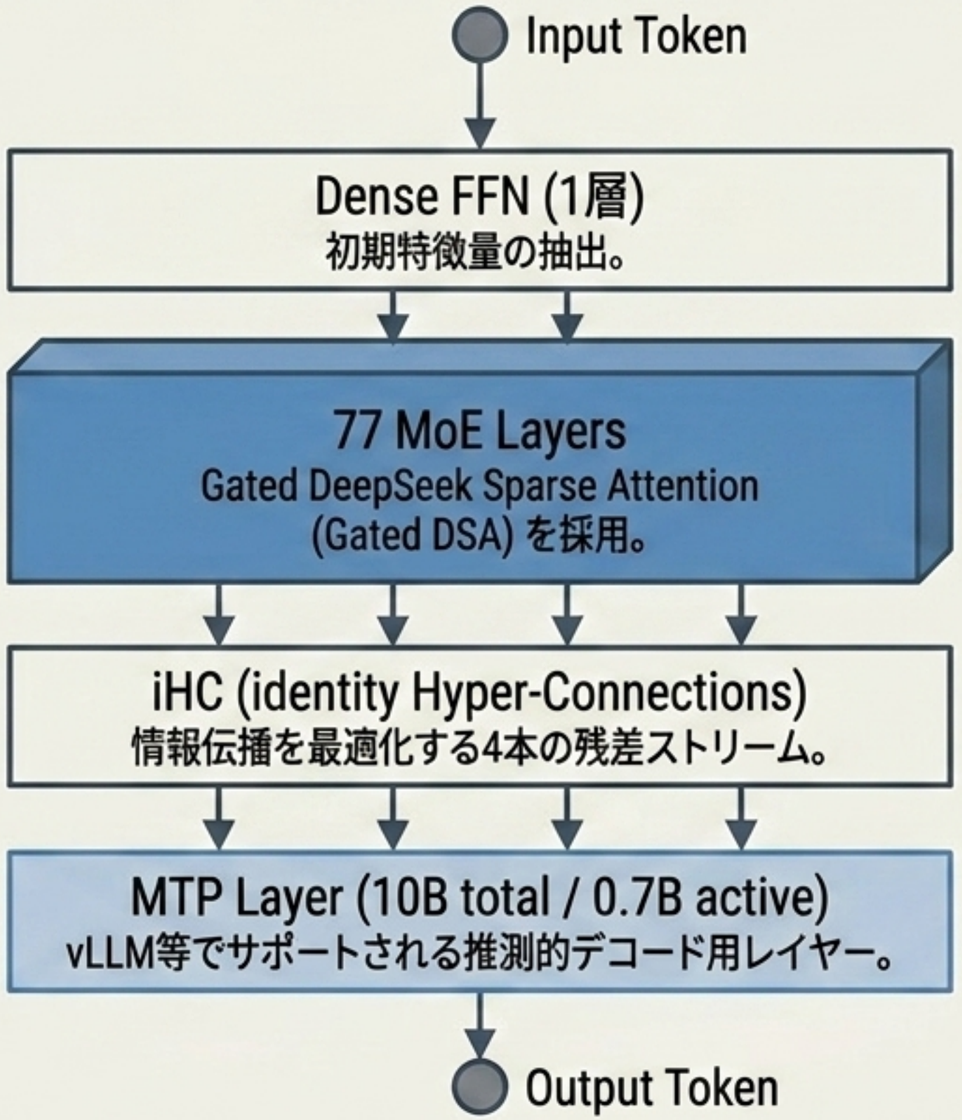


バックボーン層: 770B (77 MoE layers)	770B
MTP層（推測的デコード用）:	10B
Hugging Face チェックポイント総量:	**約780B**

Per-Token Activation

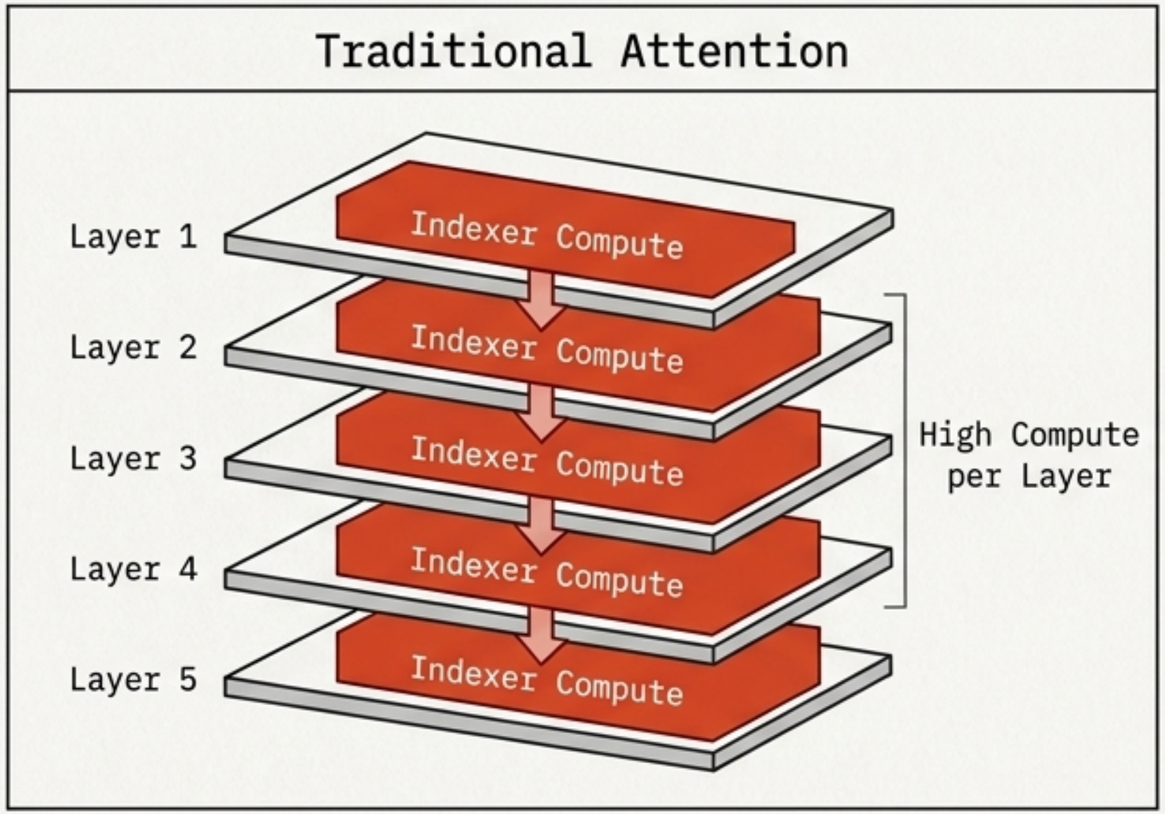
256 Routed Expertsの内、各層で利用されるのは上位8個と1個のShared Expertのみ。
実質のActive Parameterは「49B」。

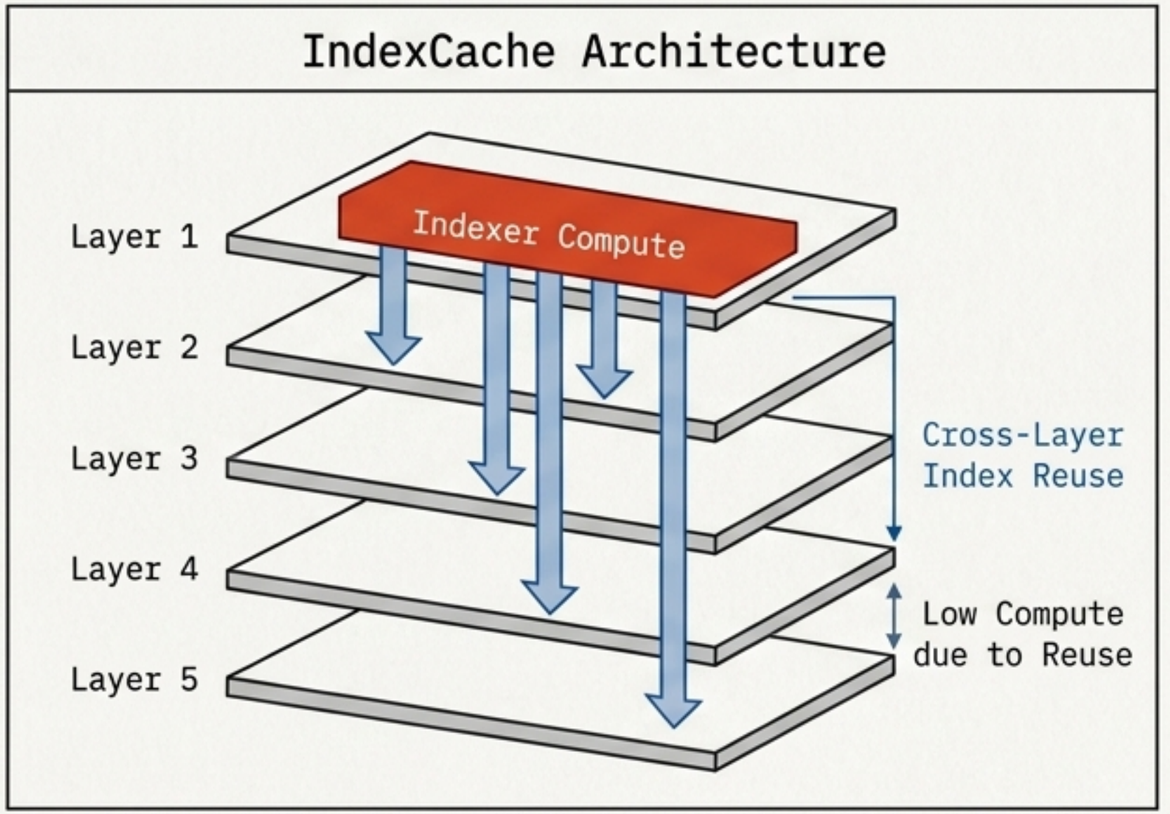
トークン処理の内部構造：長文脈と高スループットを支えるエンジン



100万トークン文脈の経済的成立：IndexCacheによる計算削減

1Mコンテキストは単にサポートするだけでは計算コストが高すぎて実運用できない。





Gated DSAにおけるIndexer自体の計算コストを圧縮。Indexer計算量を最大大75%削減し、巨大な文脈長の実運用に向けた推論コストを最適化。

表面的な「オープン」の実態：高い利用権と極端に低い再現性

Accessible（下流タスク）

- ✓ 商用利用・再配布（Apache 2.0ライセンス）
- ✓ 推論エコシステム（vLLM，SGLang完全対応）
- ✓ ファインチューニング（DeepSpeed，LLaMA-Factory対応）

Black Box

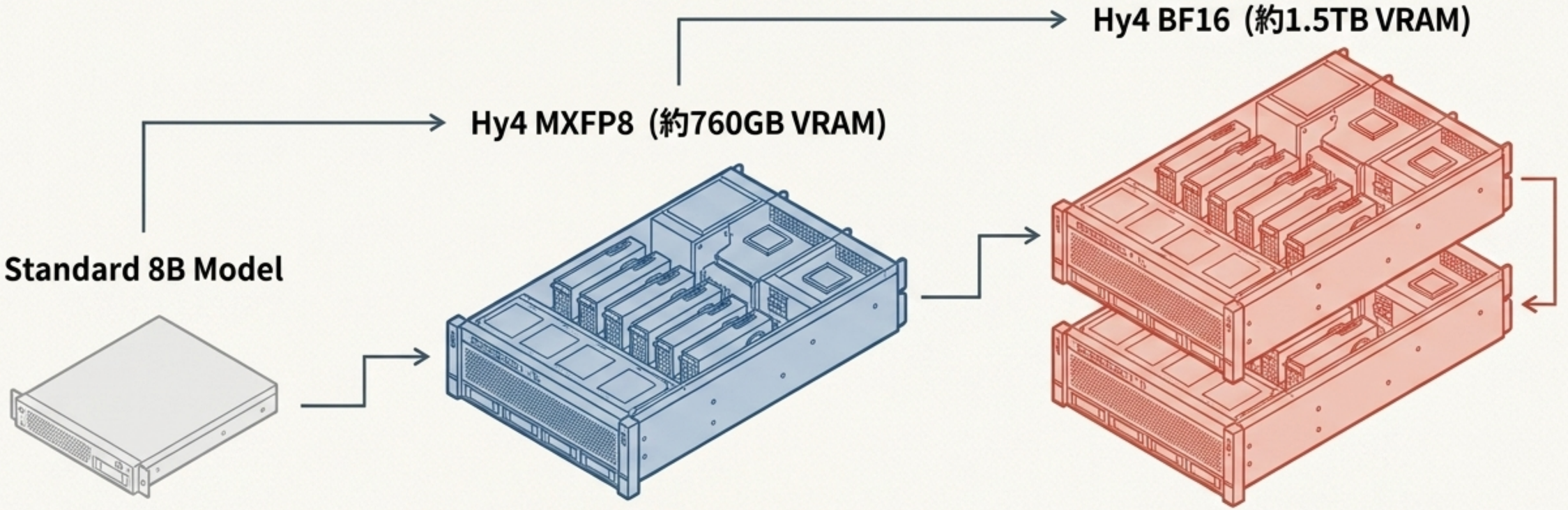
（上流プロセス）

- ✗ 事前学習データセットとデータパイプライン（未公開）
- ✗ 学習クラスタ設定とログ（未公開）

結論：これは「完全に再現可能なオープンソース」ではなく、「オープンウェイト+推論コード公開モデル」である。

運用に必要な物理インフラ：ローカルLLMの枠組みを超えるVRAM要件

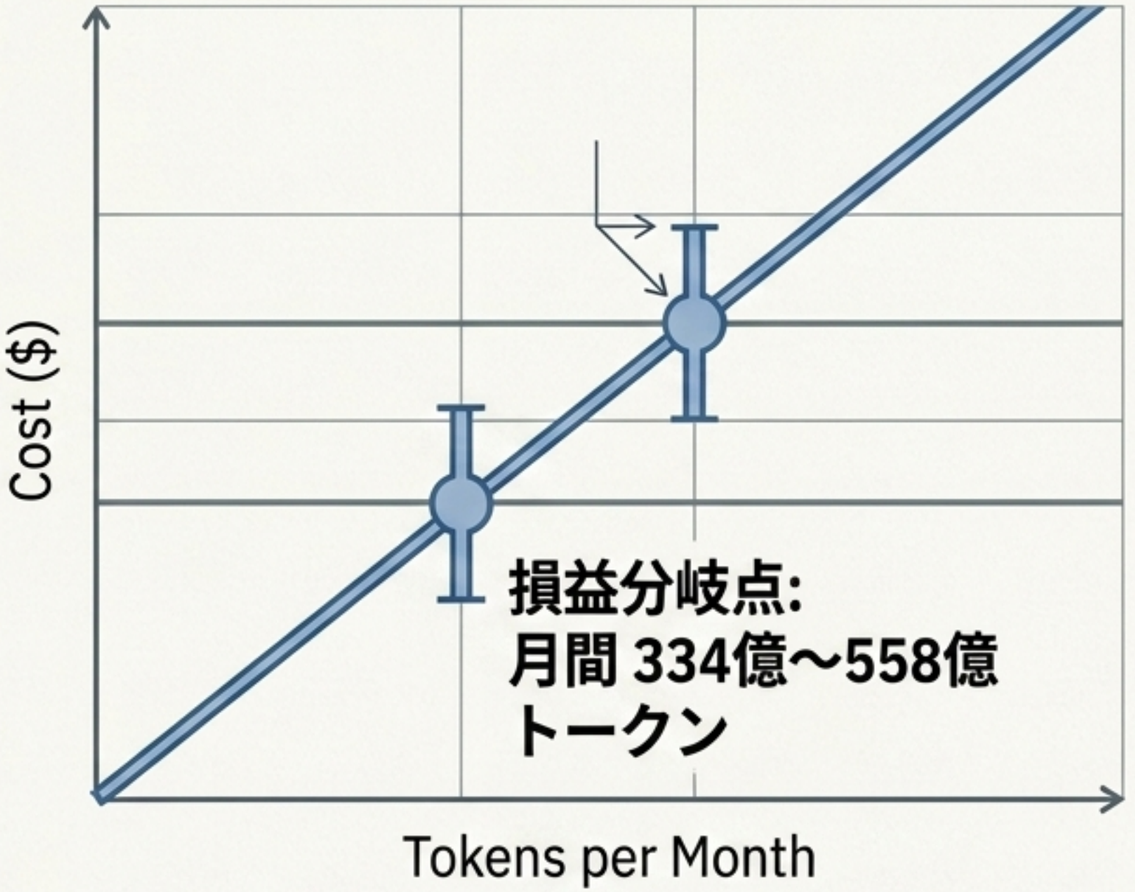
SGLang公式レシピに基づく最低要件。単一のGPUでは到底動作しない。



ファインチューニングの現実：最小構成のLoRAでさえ、8ノード・64基のGPUを要求。フルチューニングには128基のGPUが必要。

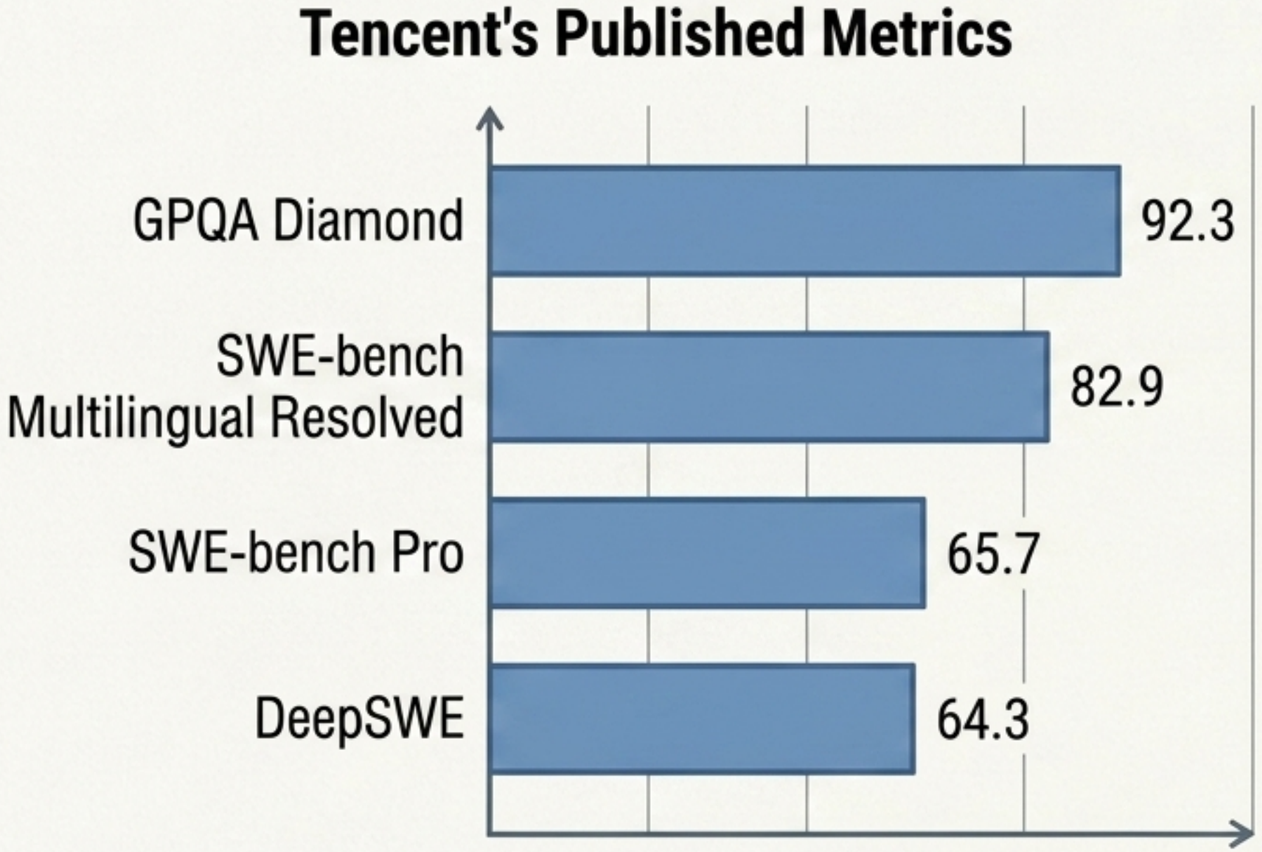
展開コストの分岐点：API利用とセルフホスト環境の損益分岐

Tencent API公式	Self-Hosted Cloud (月額固定費)
ブレンド単価: 約$1.25 / 1M tokens (Based on Input $0.834, Output $2.501)	**8x B200 (FP8): 約$41,756 / 月** **16x H200 (BF16): 約$69,730 / 月**



セルフホストの固定費をAPI利用で相殺するには莫大な処理量が必要。初期はAPI利用が経済的に合理的。

公式ベンチマークの主張：推論とソフトウェア・エンジニアリングへの特化



内部ブラインド評価 (Internal Blind Evaluation)

- 163人の専門家によるブラインドテストを実施。
- Hy4 (2.99) が GLM-5.3 (2.92) や Kimi K3 (2.94)と同等以上の評価を獲得したと主張。

※第三者機関による独立評価ではない点に留意。

比較不能な一般指標：MMLUとHellaSwagデータの意図的な欠落

公開時点で、業界標準である公式スコアが意図的または事後的に欠落している。

Model	MMLU	HellaSwag
GPT-4	86.4%	95.3%
Llama 3.1 405B	87.3%	N/A
Mistral 7B	60.1%	81.3%
Hy4 preview	[UNVERIFIED]	[UNVERIFIED]

基礎的な指標とデータ汚染（Contamination）の検証報告がない限り、GPT-4やLlama 3.1を凌駕したという定量的な断定は成立しない。

隠蔽された学習プロセス：事前学習データとコンプライアンスの空白

Redacted Matrix

学習トークン総数 (Meta Llama 3の15T明示との落差)	データセットの出所 (Web/コード/書籍の構成比)	著作権処理 (フィルタリングとオプトアウトの有無)
日本語データ (学習量と品質)	蒸留手法 (他モデルからのDistillation適用の有無)	

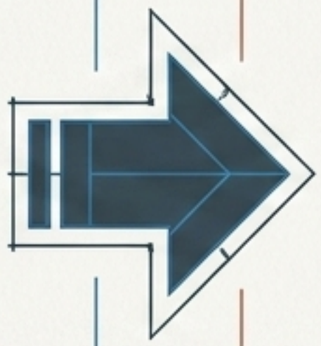
Enterprise Implication

データの権利処理が不明なため、規制産業でのオンプレミス導入には極めて高い法務デューデリジェンス・リスクが伴う。

運用上のセキュリティ脅威：
強力な機能と不足する安全性評価の交差点

Capability Risks（強力な機能）

- 高度なコーディング能力
  + ツール呼び出し
  = マルウェア生成支援のリスク
- 100万トークンの長文脈
  = プロンプトインジェクションの検知困難化
- Apache 2.0のローカル運用
  = クラウド側のセーフティフィルタによる強制遮断が不可

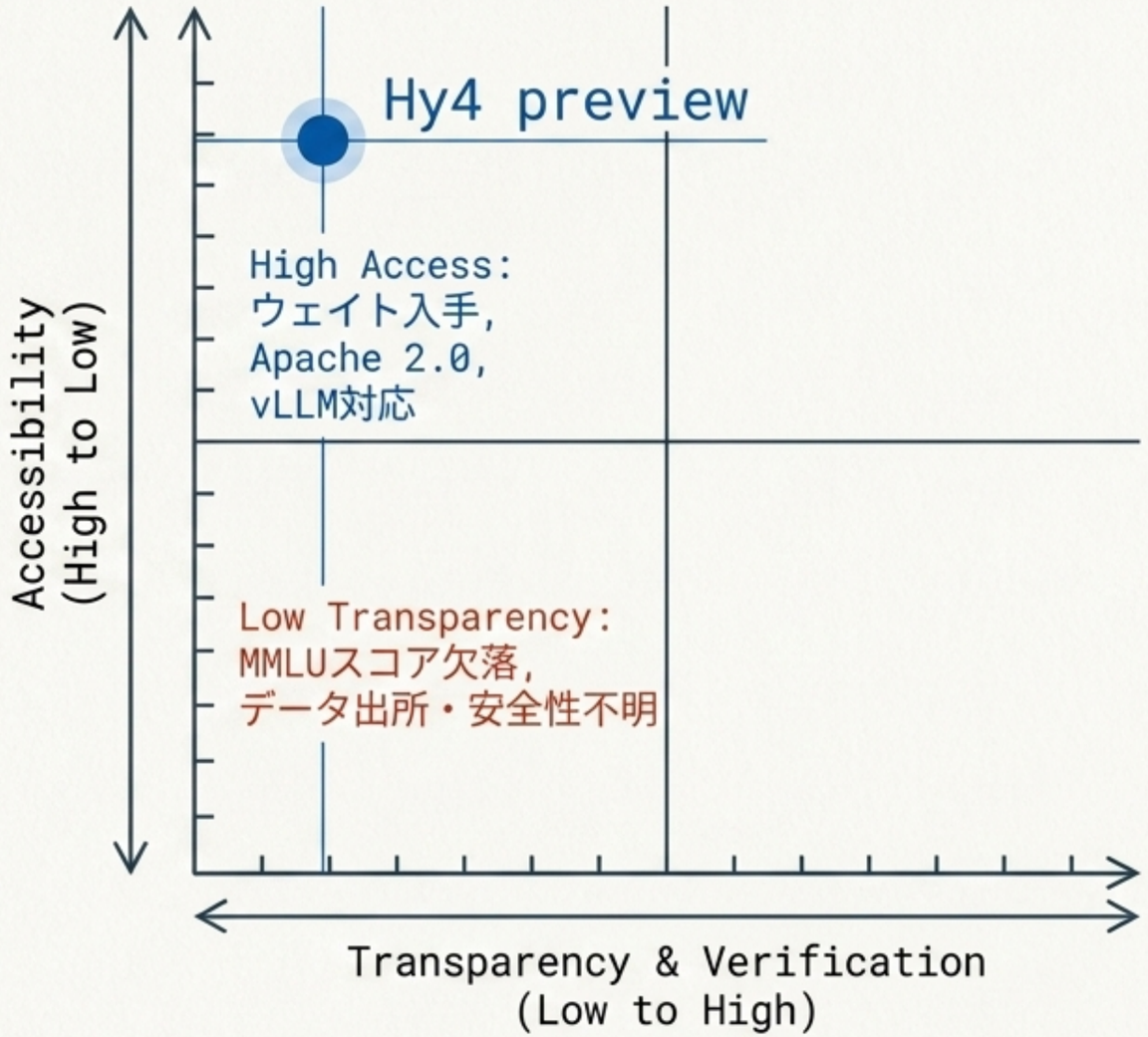


Missing Guardrails（不足する防壁）

- ToxiGen, BBQ等のバイアス評価
- Jailbreak耐性レポート
- サイバー/CBRNのレッドチーム検証

System Card相当の安全性評価が一切公開されていない。

総合評価：Hy4の実力は「770B」という数字にあらず



Final Verdict

Hy4の真の価値は「7700億」というマーケティング上の規模ではない。

この野心的な49B-active MoEモデルが、競合に対して**「実測スループット」「トークン単価」「実際の安全性」**で**優位に立てるかどうかが、今後の独立評価における最大の焦点となる。**