Tencent Hy4 Preview: 7700億パラメータMoEと実用AIの新基準

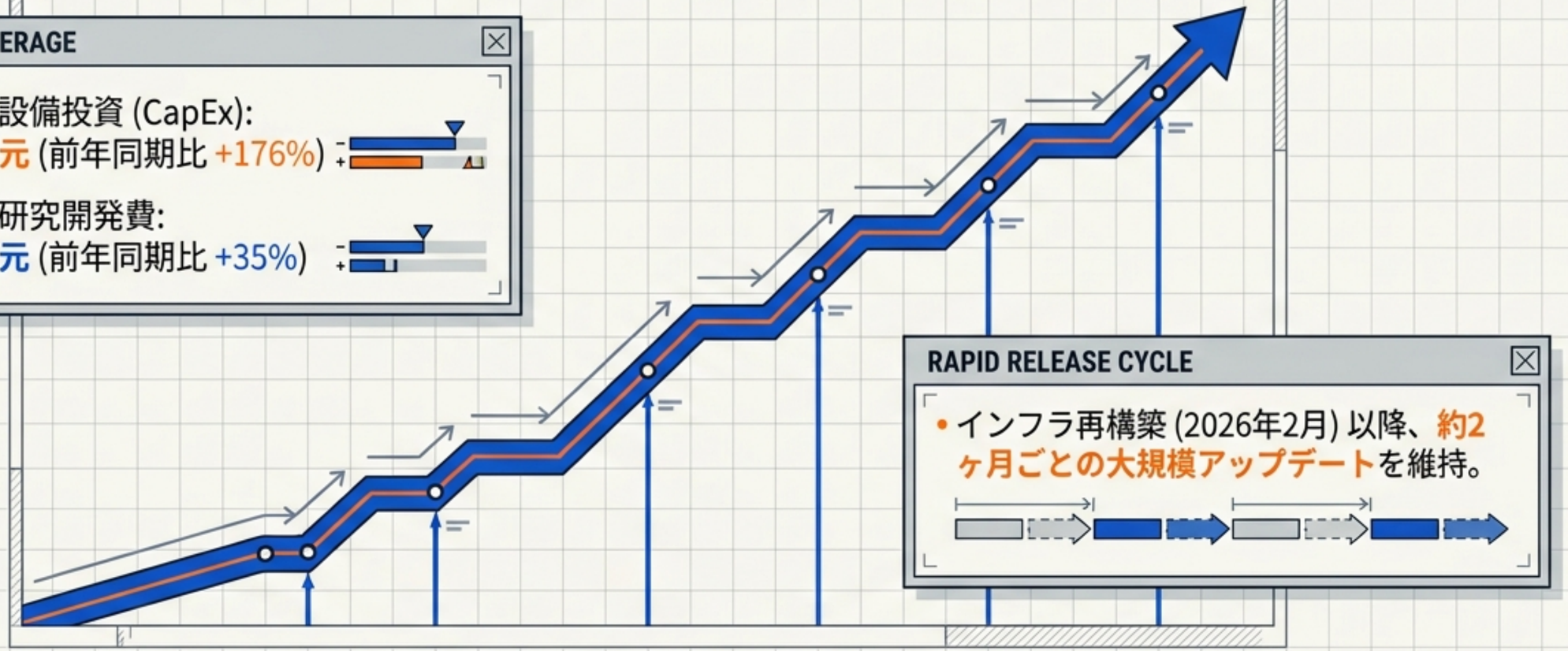
The New Standard for Practical AI: Structural Innovations & Strategic Market Impact

CONFIDENTIAL SYSTEM BRIEFING / / AUGUST 2026

戦略的コンテキスト：圧倒的な資本投下と高密度なイテレーション

FINANCIAL LEVERAGE

- 2026年Q2 設備投資 (CapEx):
  528億人民元 (前年同期比 +176%)
- 2026年Q2 研究開発費:
  273億人民元 (前年同期比 +35%)



RAPID RELEASE CYCLE

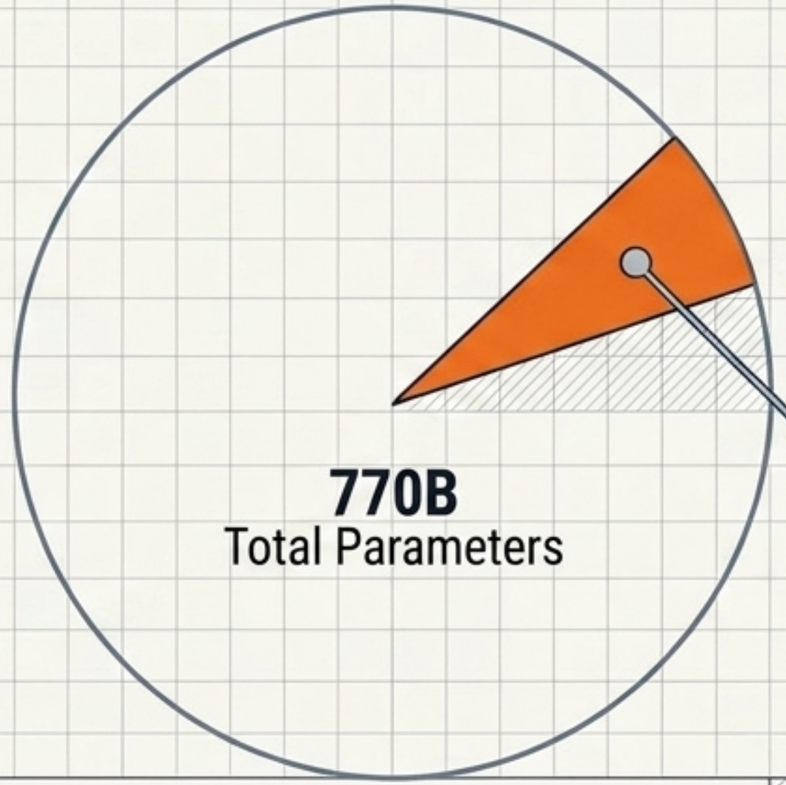
- インフラ再構築 (2026年2月) 以降、**約2ヶ月ごとの大規模アップデート**を維持。

単純なパラメータ規模の拡大から、「実稼働環境での推論効率」と「コンテキスト長の拡張性」へ。
Tencentの巨大な資本投下が、高密度なリリースサイクルとアーキテクチャの刷新を駆動している。

アーキテクチャのパラドックス：圧倒的スケールと極限の計算効率

THE ENGINE SPECS

- **総パラメータ数: 7,700億** (Backbone Capacity)
- **活性化パラメータ: 490億 / トークン** (Active Payload)



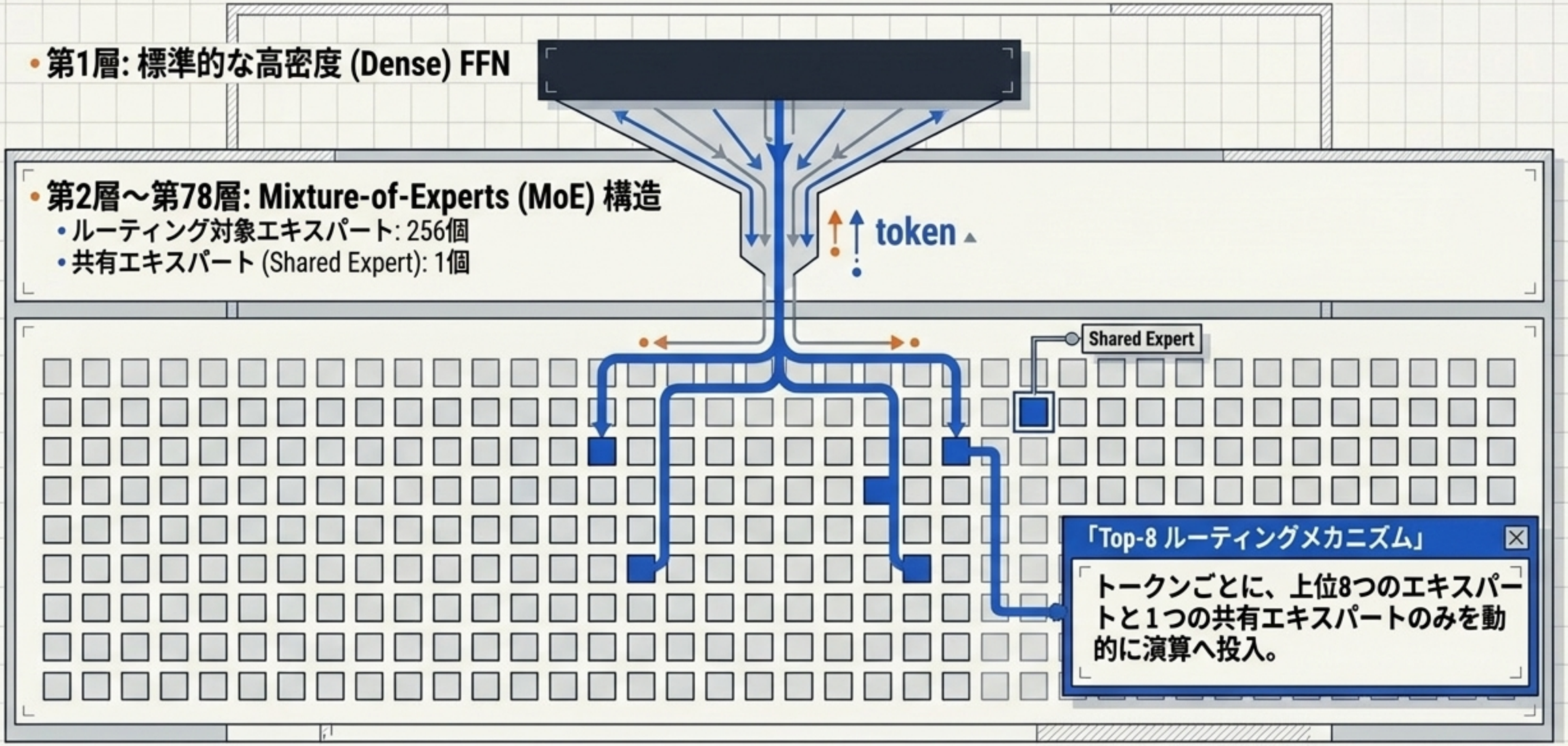
CONTEXT CAPACITY

- **最大コンテキスト長: 100万トークン**

「6.4% Activation Dial」：膨大なパラメータ容量による高度な知識表現を維持しながら、推論時の計算負荷を総容量のわずか6.4%に抑え込む疎結合（Sparse）設計。

MoEレイヤーの解剖学：スパース性（疎結合）の視覚化

- 第1層: 標準的な高密度 (Dense) FFN



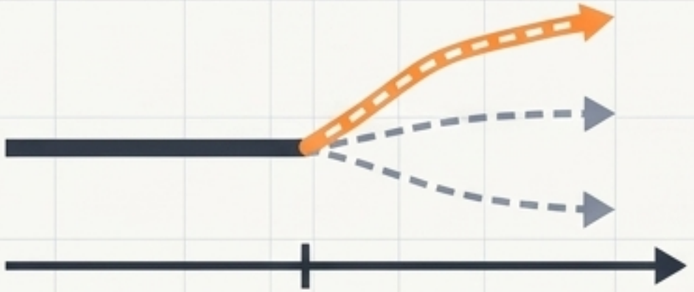
スループットとアテンションの革新

Gated DSA & IndexCache

- Gated DeepSeek Sparse Attention
- 層間でのスパースインデックス再利用を可能にする「IndexCache」を統合（64ヘッド）。

iHC (identity Hyper-Connections)

- 残差結合経路を4ストリームに拡張。
- 長文コンテキストにおける情報伝達のボトルネックを解消。



Native MTP (Multi-Token Prediction)

- ネイティブに組み込まれた投機的デコーディング (Speculative Decoding)。
- 自律的に未来のトークンを予測し、推論速度を劇的に向上。

デプロイメントとハードウェアの現実：本番環境への実装

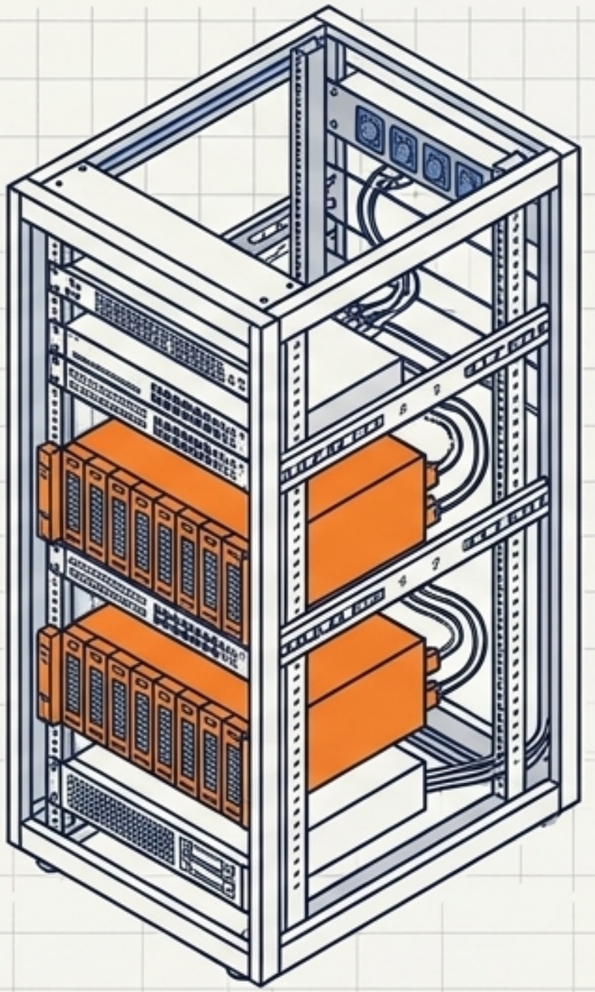
オープンウェイト仕様 (Open-Weight Specifications)

- フォーマット: BF16 / FP32, FP8量子化版 (AngelSlim対応)
- サービング環境: vLLM (v0.29.0+), SGLang
- テンソル並列度 (TP): 8

⚠

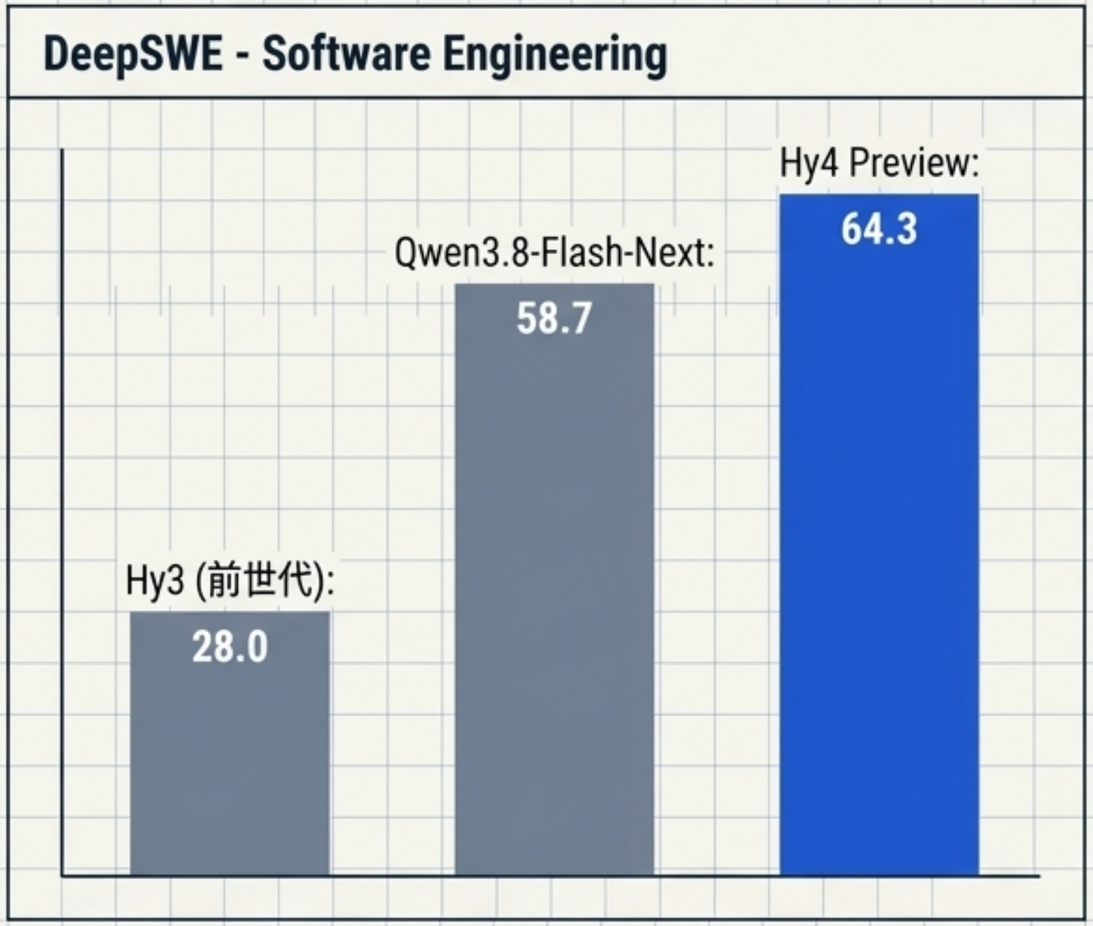
MTP（投機的デコーディング）有効化時の推奨クラスタ構成：

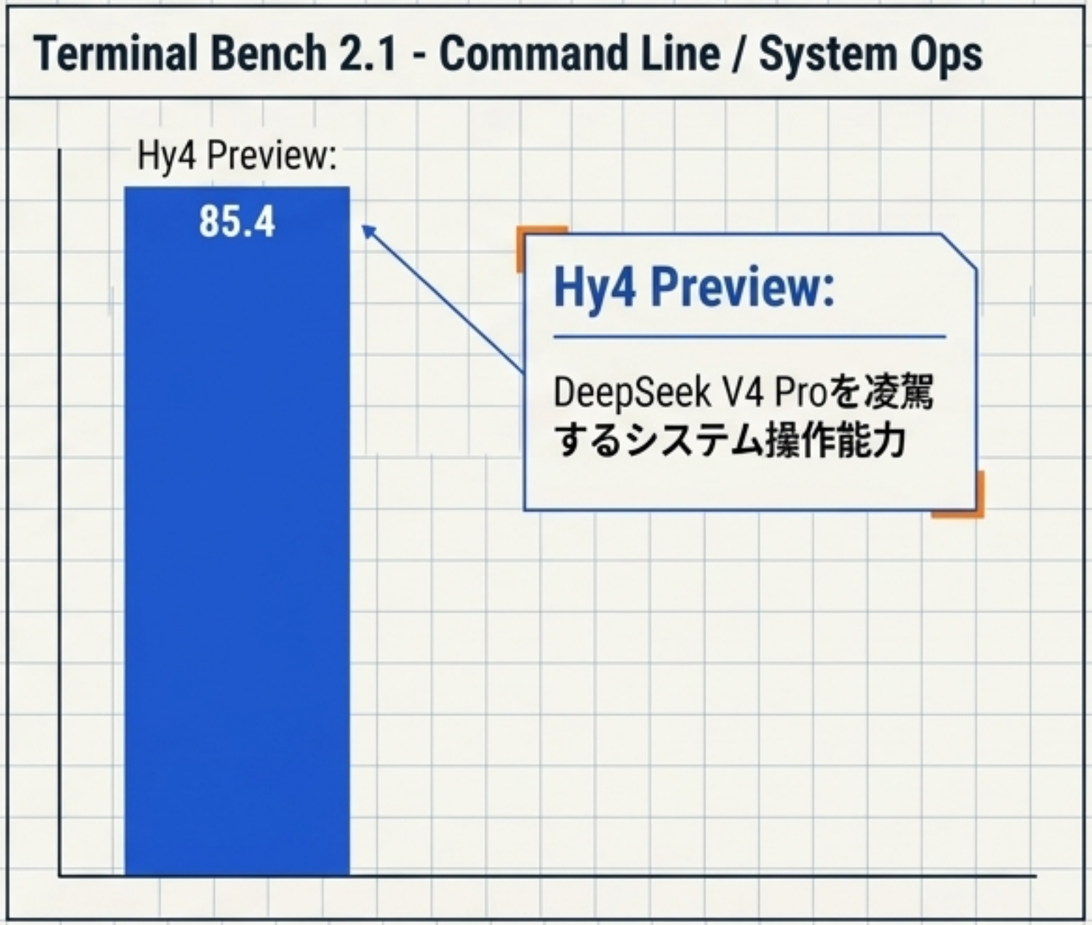
- Option A: 16x NVIDIA B200 GPU
- Option B: 8x NVIDIA B300 GPU



Apache 2.0ライセンスに基づくオープンウェイト形式。分散サービングを前提としたエンタープライズ・グレードの実装。

実務性能ベンチマーク：エンジニアリングとシステム制御





・GDPval-AA V2 (実務知的労働評価): 1,678 (Kimi K3の1,668を上回る)

専門家ブラインドテスト・バトルカード (Comparison Matrix I)

テンセント内部の専門家163名による、203件の実務工学タスク評価 (4点満点)

Model	Overall Score	Win Rate	Tie Rate	Loss Rate
Hy4 Preview	**2.99**	-	-	-
vs GLM-5.3	2.92	46.8% (勝)	12.8% (分)	40.4% (敗)
vs Kimi K3	2.94	51.2% (勝)	7.9% (分)	40.9% (敗)

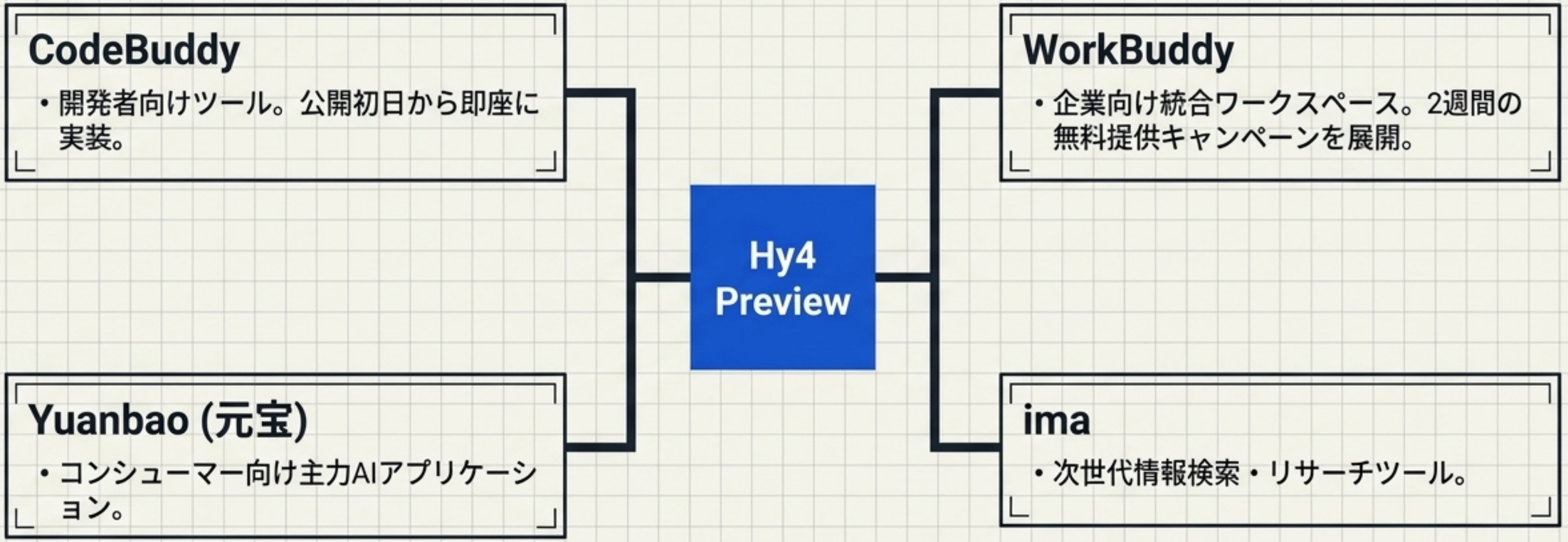
高付加価値な実務生産性領域（ソフトウェア工学、ドキュメント分析、科学研究）に特化したチューニングにより、国内主要フロンティアモデルを上回る評価を獲得。

「プレビュー版」の現実と透明性：現在の制約と課題

過剰検証ループ (Over-verification)	複雑なタスクにおいて、自らの出力を必要以上に検証・再確認し続ける現象が確認されている。
思考ステップの消費 (Latency in Logic)	推論過程において、不要に多い思考ステップを消費する傾向。
最高峰モデルとのギャップ (The Frontier Gap)	極めて複雑な多段階推論タスクにおいては、GPT-5.6 Solなどの最上位フロンティアモデルに対して未だ改善の余地を残す。

プレビュー先行公開戦略：これらの課題をオープンソースコミュニティや実運用環境からのフィードバックを通じて迅速に吸収し、正式版（GA）の最適化へと繋げる反復型アプローチ。

エコシステムの浸透：Day 1からの完全統合



モデル単体のオープンソース化と並行し、Tencent自社インフラを用いた即時商用展開を実施。
理論上のモデル発表にとどまらず、即日での大規模ユーザー獲得（エコシステム・ロックイン）を完了。

商業的ディスラプター：価格破壊マトリクス (Comparison Matrix II)

	入力:	出力:	キャッシュヒット:
Hy4 Preview	**$0.834**	**$2.501**	**$0.042**
Z.ai GLM-5.3	$4.40	N/A	
Moonshot Kimi K3	$15.00	N/A	

- Deep Reasoning (思考プロセスの保持)
- 構造化出力
- Function Calling

Tencent Cloud TokenHub (シンガポール) および OpenRouter経由で提供。MoEの極限の計算効率が、競合を完全に突き放す圧倒的なコストパフォーマンスを可能にしている。

統合的洞察：The Autonomous Agent Equation

6.4%（490億）の低アクティブパラメータ

＋

85.4 Terminal Bench (高次元のシステム制御能力)

＋

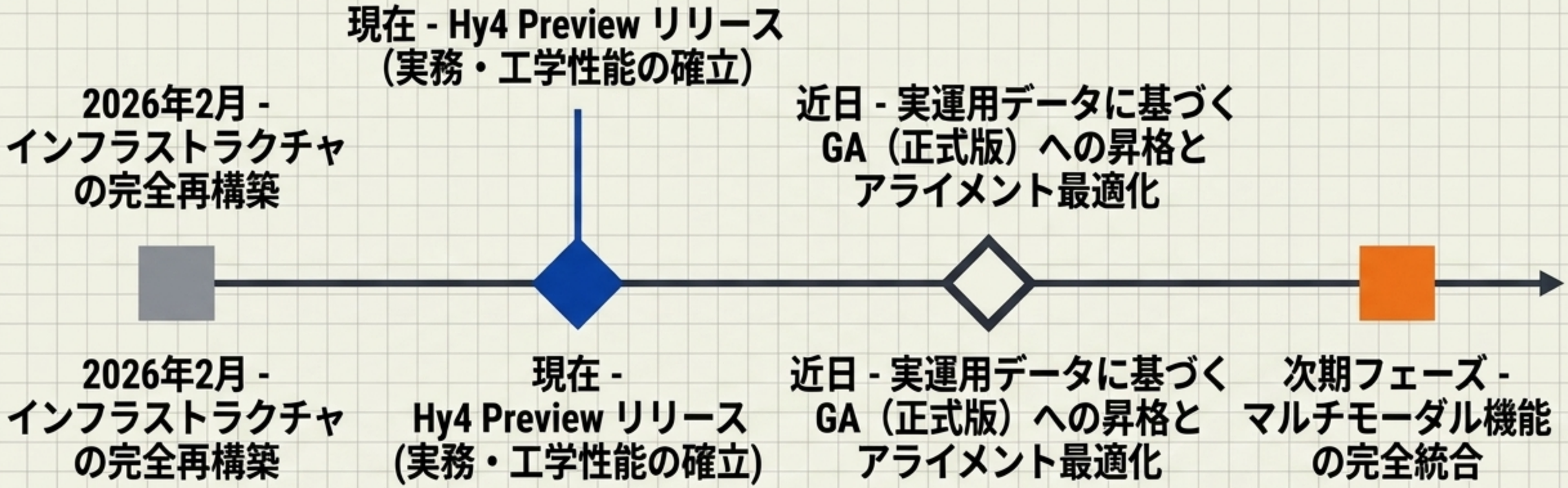
$0.834 入力コスト (極限の経済性)

＝

究極の自律型エージェント・エンジン (The Ultimate Agent Engine)

反復推論や検証ループを多用し、大量のトークンを消費する「自律型AIエージェント」の運用において、これら3つの要素の結合は、運用コストの壁を破壊し、実用化への最大のブレイクスルーとなる。

今後の展望：GAリリースとマルチモーダルへの進化



巨額な設備投資と継続的なインフラ刷新を背景に、Hy4シリーズはオープンソースAIコミュニティとエンタープライズ生成AIエコシステム、双方の中心的なエンジンの座を確固たるものにする。