テンセントによる7700億パラメータMoEモデル「Hy4 preview」の技術構造・性能評価および産業的意義

Gemini 3.7 Flash

開発背景とエコシステム戦略における位置づけ

大規模言語モデル（LLM）の研究開発競争は、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、コンテキスト長の拡張性、および実用的なタスク遂行能力を重視するフェーズへと移行している[1]。この潮流の中、テンセント（騰訊・Tencent）は混元（Hunyuan）シリーズの次世代フラッグシップモデル「Hy4 preview」を発表し、Apache 2.0ライセンスに基づくオープンウェイト形式で一般公開した[2]。本モデルは総パラメータ数7700億（$770\mathrm{B}$）、トークンあたりのアクティブパラメータ数490億（$49\mathrm{B}$）を誇る疎結合Mixture-of-Experts（MoE）構造を備え、最大100万（$1\mathrm{M}$）トークンのコンテキスト処理を可能にしている[2]。

テンセントは2026年第2四半期において、研究開発費を前年同期比35%増の273億人民元、設備投資を同176%増の528億人民元へと大幅に拡大し、計算資源とモデルアーキテクチャの刷新を進めてきた[5]。同年2月のインフラ再構築以降、混元開発チームは約2か月ごとに大規模アップデートを実施する高密度なリリースサイクルを維持している[1]。今回の「preview先行公開」戦略は、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迅速に吸い上げ、正式版（GA）の最適化へと繋げる反復型アプローチを体現したものである[1]。

モデルアーキテクチャと計算効率化メカニズム

Hy4 previewのバックボーンは78層のトランスフォーマー層で構成されており、第1層に標準的な高密度（Dense）フィードフォワードネットワーク（FFN）を配置し、第2層から第78層までの77層にMoE構造を採用している[3]。各MoE層は256個のルーティング対象エキスパートと1個の共有エキスパート（Shared Expert）を備え、トークンごとに上位8個のエキスパートと共有エキスパートが同時に活性化される[3]。総パラメータ数の約6.4%のみを動的に演算へ投入する設計により、膨大なパラメータ容量を確保しながら推論時の計算負荷を大幅に低減している[3]。

超長文コンテキスト処理と推論のスループット向上を実現するため、複数の先進的なアーキテクチャ技術が統合されている[3]。アテンション層には、層間でのスパースインデックス再利用を可能にする「IndexCache」を統合した「Gated DeepSeek Sparse Attention (Gated DSA)」が採用されている[3]。さらに、残差結合経路を4ストリームに拡張して情報伝達を円滑化する「identity Hyper-Connections (iHC)」が組み込まれている[3]。また、投機的デコーディング（Speculative Decoding）を自律的に実行するためのマルチトークン予測（Multi-Token Prediction: MTP）レイヤー（総パラメータ$10\mathrm{B}$、アクティブ$0.7\mathrm{B}$）がネイティブに組み合わされている[3]。

構成要素・モジュール	アーキテクチャ仕様・数値
アーキテクチャ分類	Mixture-of-Experts (MoE)
総パラメータ数（Backbone）	7700億（$770\mathrm{B}$）[3]
活性化パラメータ数（トークンあたり）	490億（$49\mathrm{B}$）[3]
レイヤー構成	78層（第1層: Dense FFN / 第2〜78層: MoE）[3]
エキスパート構成	256 ルーテッドエキスパート ＋ 1 共有エキスパート（Top-8 選択）[3]
投機的デコーディング（MTP層）	総パラメータ $10\mathrm{B}$ / 活性化パラメータ $0.7\mathrm{B}$ [cite: 3, 10]
コンテキスト長	$1,048,576$ トークン（$1\mathrm{M}$）[3]
アテンション機構	Gated DSA（IndexCache統合、64ヘッド、Indexer Top-k: 2048）[3]
隠れ層次元 / 残差ストリーム数	Hidden Size: 6144 / Residual Streams: 4[3]
ボキャブラリサイズ	120,832 トークン[3]
配布フォーマット / 量子化	BF16/FP32、FP8量子化版（AngelSlim対応）[3]

推論デプロイメント環境においては、vLLM（v0.29.0以降）およびSGLangに対応しており、テンソル並列度（TP）8を前提とした分散サービング環境が提供されている[10]。MTPを有効化した本番環境での推奨ハードウェア構成としては、16基のNVIDIA B200 GPU、または8基のNVIDIA B300 GPUクラスタが指定されている[10]。

実務性能ベンチマークと定量的評価

Hy4 previewの設計思想は、一般的な対話能力にとどまらず、ソフトウェアエンジニアリング、オフィスドキュメント分析、ゲームプロトタイピング、科学研究といった高付加価値な実務生産性領域に特化している[3]。

テンセント内部の専門家163名が203件の実務工学タスクを対象に実施したブラインド評価（4点満点）において、Hy4 previewは平均2.99点を獲得した[3]。このスコアは、中国の主要なフロンティアオープンウェイトモデルであるZ.aiの「GLM-5.3」（平均2.92点）やMoonshot AIの「Kimi K3」（平均2.94点）

をわずかに上回る水準である[3]。
標準的な公開ベンチマークにおいても、前世代モデル（Hy3）からの劇的な進化が確認されている[5]。ソフトウェア工学ベンチマークの「DeepSWE」スコアは前世代の28.0から64.3へと急伸し、コマンドラインやシステム操作能力を評価する「Terminal Bench 2.1」では85.4点を記録してDeepSeek V4 Proを上回った[5]。

評価指標・ベンチマーク	**Hy4 preview**	比較対象モデル / 前世代スコア
社内工学ブラインド評価（163名 / 203タスク）	**2.99** / 4.00[3]	GLM-5.3: 2.92 / Kimi K3: 2.94[3]
対 GLM-5.3 勝敗比率（工学ブラインドテスト）	勝率 **46.8%** / 引分 12.8%[3]	敗率 40.4%[3]
対 Kimi K3 勝敗比率（工学ブラインドテスト）	勝率 **51.2%** / 引分 7.9%[3]	敗率 40.9%[3]
DeepSWE（ソフトウェア工学）	**64.3** [cite: 5, 15, 16]	Hy3（前世代）: 28.0 / Qwen3.8-Flash-Next: 58.7[5]
Terminal Bench 2.1	**85.4** [cite: 5, 15, 16]	DeepSeek V4 Pro を凌駕[5]
GDPval-AA V2（実務知的労働評価）	**1678** [cite: 1, 17]	Kimi K3: 1668[17]

一方で、モデルカードに明記されている通り、本モデルはプレビュー段階特有の課題を抱えている[3]。具体的には、複雑なタスクにおいて必要以上の思考ステップを消費する傾向や、自らの出力を過度に検証・再確認し続ける現象（Over-verification）が指摘されている[3]。また、極めて複雑な多段階推論タスクを対象とする一部の指標では、最上位の超巨大フロンティアモデル（GPT-5.6 Sol等）に対して未だ改善の余地を残している[1]。

エコシステム統合と商用コスト構造

テンセントはモデル単体のオープンソース化と並行して、自社サービスへの直接実装およびクラウドインフラを通じた商用APIの提供を即座に開始した[9]。開発者向けツール「CodeBuddy」や企業向け統合ワークスペース「WorkBuddy」には公開初日からHy4 previewが組み込まれ、2週間の無料提供キャンペーンが展開された[9]。さらに、コンシューマー向けAIの「元宝（Yuanbao）」や情報検索ツール「ima」にも本モデルが統合されている[9]。
商用API提供基盤としては、Tencent Cloud TokenHub（シンガポール拠点のグローバルエンドポイント）およびOpenRouterが採用された[18]。価格設計は極めて競争力が高く設定されており、競合するフ

ロンティア級モデルと比較して出力トークンのコスト効率が際立っている[8]。

項目・提供チャネル	仕様および価格体系
オープンソース配布	Hugging Face, ModelScope, GitHub, GitCode, CNB（Apache 2.0）[3]
API 通常入力価格	$0.834 / 100万トークン（中国国内: 6元 / 100万トークン）[5]
API 出力価格	$2.501 / 100万トークン（中国国内: 18元 / 100万トークン）[5]
キャッシュヒット価格	$0.042 / 100万トークン[4]
トークン許容量	最大入力 960k トークン / 最大出力 64k トークン（総枠 1M）[20]
API 機能対応	Deep Reasoning（思考プロセスの保持）、構造化出力、Function Calling[20]
提供エンドポイント	Singapore Base URL（Global Resource Scheduling）[20]

この価格水準は、Moonshot AIのKimi K3（入力 $3.00/\text{出力}$ 15.00）やZ.aiのGLM-5.3（入力 $1.40/\text{出力}$ 4.40）と比較して大幅に安価であり、反復推論や検証ループを多用する自律型AIエージェントの運用コストを大幅に抑制する[21]。

結論と今後の展望

テンセントによるHy4 previewのリリースは、オープンウェイトLLMのフロンティアを7700億パラメータ規模へと拡張し、実務指向のAIモデル開発における重要なベンチマークを確立した[3]。
Gated DSA、IndexCache、iHC、そしてMTPレイヤーによる投機的デコーディングといった複合的なアーキテクチャ革新は、7700億パラメータの巨大容量と490億パラメータの実稼働計算量を両立させ、100万トークン長文コンテキストの高速推論を実現している[2]。過剰検証や思考遅延といったプレビュー段階特有の挙動は残るものの、WorkBuddyやCodeBuddyをはじめとする実サービスでの運用データが、今後の事前学習およびアライメント調整をさらに加速させるとみられる[3]。
今後の展開としては、プレビュー版から正式安定版への昇格、さらにはマルチモーダル機能の統合が予定されている[7]。テンセントの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エコシステム双方において中心的な役割を

担っていくと考えられる[1]。

引用文献

1. 腾讯混元Hy4来了！7700亿参数直接开源, https://finance.sina.com.cn/tech/roll/2026-08-29/doc-inipwxpi9008421.shtml
2. Tencent Hunyuan - AIモデル「Hy4 Preview」を公開 - Phemex, https://phemex.com/ja/news/article/tencent-hunyuan-unveils-hy4-preview-with-770b-parameters-94831
3. Tencent Hy4 Preview: Inside the 770B Open-Weight Flagship Model, https://www.mindstudio.ai/blog/tencent-hy4-preview-open-weight-model
4. Tencent Hy4 preview — 770B open-weights model… | AI/TLDR, https://ai-tldr.dev/releases/tencent-hy4-preview/
5. Tencent Hunyuan、新世代大規模モデル「Hy4 preview」を発表 オープンソース化し複数製品に同時搭載, https://finance.biggo.jp/news/439ad16c-57ce-4efc-bfd0-83f079cfdc9c
6. Tencent Hunyuan releases next-generation Hy4 preview model, open-sourced and launched across multiple products, https://finance.biggo.com/news/439ad16c-57ce-4efc-bfd0-83f079cfdc9c
7. Tencent lança novo modelo e diz superar Z.ai e Moonshot, https://exame.com/inteligencia-artificial/tencent-lanca-novo-modelo-e-diz-superar-z-ai-e-moonshot/
8. Tencent Releases and Open-Sources Tencent Hy4 preview, https://www.tencent.com/tencent-releases-and-open-sources-tencent-hy4-preview/
9. Tencent HunYuan releases and open-sources the Hy4 preview with, https://www.kucoin.com/news/flash/tencent-hunyuan-releases-and-opens-source-hy4-preview-with-770b-total-parameters
10. tencent/Hy4-preview - vLLM Recipes, https://recipes.vllm.ai/tencent/Hy4-preview
11. Tencent Activates Just 6.4% of Its 770 Billion-Parameter AI, https://www.gurufocus.com/news/9058240/tencent-activates-just-64-of-its-770-billionparameter-ai
12. Tencent releases new AI model, says it beat ZAI and Moonshoot in testing, https://www.indiatoday.in/technology/news/story/tencent-releases-new-ai-model-says-it-beat-zai-and-moonshoot-in-testing-2981956-2026-08-28
13. [閒聊] 騰訊混元Hy4預覽版發布- 看板AI_Art - 批踢踢實業坊, https://www.ptt.cc/bbs/AI_Art/M.1787918987.A.BF7.html
14. 騰訊混元發布新一代大模型Hy4 preview 開源並同步上線多款產品, https://finance.biggo.com.tw/news/439ad16c-57ce-4efc-bfd0-83f079cfdc9c
15. Hy4 preview Benchmarks & Context (August 2026) | BenchLM.ai, https://benchlm.ai/models/hy4-preview
16. Hy4 preview vs Qwen3.8-Flash-Next: Benchmarks & Cost - BenchLM, https://benchlm.ai/compare/hy4-preview-vs-qwen3-8-flash-next
17. GDPval-AA Leaderboard & Scores - Benchmarks - BenchLM.ai, https://benchlm.ai/benchmarks/gdpvalaa

18. Tencent open-sources Hy4 preview with 770B parameters and a 1M-token context, https://technode.com/2026/08/28/tencent-open-sources-hy4-preview-with-770b-parameters-and-a-1m-token-context/
19. Tencent releases new AI model, says it beat ZAI and Moonshoot in testing, https://www.indiatoday.in/amp/technology/news/story/tencent-releases-new-ai-model-says-it-beat-zai-and-moonshoot-in-testing-2981956-2026-08-28
20. Hy4 Preview FAQ Specifications, API, Pricing, Regions, and Limits, https://www.tencentcloud.com/techpedia/148044?lang=en
21. Tencent's Hy4 Outperforms GLM-5.3 and Kimi K3 in Internal Tests, Reduces Output Costs by 82%, https://www.kucoin.com/news/flash/tencent-s-hy4-outperforms-glm-5-3-and-kimi-k3-in-internal-tests-cuts-output-costs-by-82
22. Hy4 Preview: 770B Open Weights at $0.83/M Input | explainx.ai Blog, https://explainx.ai/blog/tencent-hy4-preview-770b-moe-1m-context-august-2026
23. 技术文摘- 程序员工具箱- 在线工具, http://tool.pfan.cn/article
24. Repository-Level Code Analysis with Hy4 Preview - Tencent Cloud, https://www.tencentcloud.com/techpedia/148038?lang=en
25. 腾讯混元大模型_大语言模型 - 腾讯云, https://cloud.tencent.com/product/tclm
26. 腾讯Q2财报披露混元大模型进展Hy4预告将推出 - 淘AI工具, https://bot.taobao.com/tool-today_discovery/detail/da477e60c98b766930d0dd17982fca32.html
27. 騰訊Q2売上高11%増 AI投資が自由キャッシュフロー悪化の要因に, https://finance.biggo.jp/news/fdcf980d-f23b-4dc2-a99c-d58b7d0a589f