DeepSeek-V4-Pro-0813の全貌：アーキテクチャ刷新・エージェント性能・コスト構造の包括的分析

Gemini 3.6 Flash

2026年8月13日、中国のAIスタートアップであるDeepSeekは、フラッグシップ大規模言語モデル「DeepSeek-V4-Pro」の正式版（General Availability: GA）にあたる最新ビルド「DeepSeek-V4-Pro-0813」の提供を開始した[1]。2026年4月に公開されたプレビュー版以降、継続的なポス・トレーニング（Post-training）の最適化が進められてきたが、本アップデートによりコード生成、ターミナル操作、自律型エージェントタスクにおける実用性能が飛躍的に向上している[2]。APIエンドポイントにおけるモデル呼出名称は従来のdeepseek-v4-proのまま維持されており、開発者は既存のシステム構成を変更することなく、自動的に最新の「0813」バージョンへアクセス可能となっている[2]。

開発背景と主要技術仕様

DeepSeek-V4-Pro-0813は、総パラメータ数1.6兆（1.6 Trillion）、推論時の有効（アクティブ）パラメータ数490億（49 Billion）を誇る超大規模Mixture-of-Experts（MoE）構造を採用している[2]。100万トークン（1M）の極めて広いコンテキストウィンドウをサポートし、1回の応答で最大384,000トークンの長文出力を生成することが可能である[2]。

項目	詳細仕様
モデル名称	DeepSeek-V4-Pro (API指定名: deepseek-v4-pro)[2]
内部バージョン	DeepSeek-V4-Pro-0813[2]
総パラメータ数	1.6兆 (1.6T)[2]
アクティブパラメータ数	490億 (49B)[2]
モデル構造	MoE (Mixture of Experts) + ハイブリッド・アテンション[3]
コンテキスト長	1,048,576 トークン (1M)[2]

最大出力トークン数	384,000 トークン[2]
量子化 / 精度	FP4 (MoEエキスパート層) + FP8 混合精度[3]
ライセンス	MIT License (Hugging Face公開重み)[3]

本アーキテクチャの核心は、長文コンテキスト処理における計算効率を極限まで高めた点にある[3]。Compressed Sparse Attention（CSA）とHeavily Compressed Attention（HCA）を組み合わせた独自のハイブリッド・アテンション機構により、100万トークンを処理する際の単一トークンあたり推論FLOPsを前世代モデル（DeepSeek-V3.2）比で27%に低減させ、KVキャッシュの占有容量を10%にまで圧縮することに成功している[3]。さらに、深層ネットワークにおける層間シグナル伝播の安定性を高めるManifold-Constrained Hyper-Connections（mHC）や、収束スピードを高速化するMuon最適化アルゴリズムがプレトレーニングおよびポス・トレーニング段階で投入されている[3]。
モデルの制御機能面においては、タスクの複雑性に応じて推論リソースの配分を動的に調整できる「Thinking Effort Control」がネイティブサポートされた[2]。思考モードは3段階で制御可能であり、低遅延での出力を優先する「low」設定は定常的な日常会話や単純な指示追従に適しており、デフォルト設定の「high」は通常のエージェントタスクや開発処理において論理的な思考プロセスを挟んでから回答を構成する[2]。最も高強度な「max」設定は、最高難易度の課題解決や広範な探索を要する研究タスク向けにモデルの能力を最大化するよう設計されている[2]。

ベンチマーク評価とエージェント性能の定量分析

AIコーディングエージェント開発元のClineをはじめとするコミュニティ検証およびDeepSeek公式発表のチェンジログにより、DeepSeek-V4-Pro-0813はエージェントタスクおよび開発者向けタスクにおいて旧バージョンから著しい性能向上を達成したことが確認されている[2]。最も象徴的な指標は、Linuxターミナル上での実システム操作や開発構築精度を測定する「Terminal-Bench 2.1」のスコアである[2]。4月に公開されたV4-Pro Preview版の72.1点から、0813版では87.9点へと大幅なスコア増加を記録した[2]。
メディアやSNS上の一部レポートでは「15.8%の向上」と記載される事例が散見されるが、元データの変化（72.1点から87.9点）は「15.8ポイントの上昇」であり、相対的な改善率を計算すると約21.9%の向上となる[3]。相対改善率は以下の計算式によって算出される。

$$\text{相対改善率} = \frac{87.9 - 72.1}{72.1} \times 100 \approx 21.91\%$$

主要なエージェント・ベンチマークにおける前世代および他社フロンティアモデルとの性能比較は下表の通りである[2]。

ベンチマーク指標	評価対象領域	**DeepSeek V4-Pro Preview (4**	**DeepSeek V4-Flash (7**	**DeepSeek V4-Pro 0813 (8月**	比較対照モデル **(Claude**

		月)	月)	GA)	Fable 5 / 他)
Terminal-Bench 2.1	ターミナル操作・システム開発	72.1	82.7	**87.9**	88.0 (Fable 5) / 83.8 (Claude Code)[2]
CyberGym	サイバーセキュリティタスク	52.7%	76.7%	**83.3%**	-[2]
DeepSWE	AIソフトウェアエンジニアリング	12.8%	54.4%	**62.7%**	-[2]
HLE (wo / w tools)	高度論理推論 (ツールなし/あり)	-	-	**42.7 / 60.0**	-[3]
NL2Repo	自然言語からのリポジトリ構築	-	54.2%	**61.5%**	-[3]
Toolathlon -Verified	複合ツール利用タスク	-	70.3%	**74.1%**	-[3]
DSBench-FullStack	フルスタック開発内部テスト	-	68.7%	**71.1%**	-[3]

ベンチマーク評価においては、評価を実行するエージェントフレームワークやプロンプト設定によって結果が動的に変動する点に注意を要する[3]。Terminal-Bench 2.1の公式リーダーボード上では、Claude Code（Fable 5構成）が83.8%、GPT-5.5（Codex構成）が83.1%を記録しており、DeepSeek V4-Pro 0813の「87.9点」という数値は独自ハーネス（DeepSeek Harness minimal mode）やClinePassの実行環境（high/max推論強度）で測定された最高評価値に属する[2]。

また、サードパーティの包括的評価インデックスである「BenchLM」の公開データによると、DeepSeek-V4-Pro-0813は217モデル中48位（総合スコア60.9/100）にランクインしている[6]。ドメイン別では「Knowledge（知識）」カテゴリで17位に位置し、MMLU-Proで87.5%、Short-Form Factualityを測るSimpleQAで57.9%の精度を獲得している[6]。一方で、応答速度に関しては生成スピードが約83 tokens/sec、ファーストトークン到達時間（TTFT）が25.67秒となっており、高度な思考モード（Thinking

Mode）の内部推論処理に起因する初期遅延が存在することが定量的に証明されている[6]。

コスト構造と動的価格モデルの多角的大解剖

DeepSeek-V4-Pro-0813の提供開始に伴い、大きな議論を呼んでいるのがその圧倒的なトークン単価の安さである[2]。開発ツールコミュニティからは「Claude Fable 5と同等のパフォーマンスを約57分の1のコストで実現した」と評されているが、この比較にはトークン種別による価格構造の詳細な分析が必要となる[2]。

現在提供されている標準価格体系（100万トークンあたり）におけるコストは、入力トークン（キャッシュミス時）が0.435ドル（約69.37円）、入力トークン（キャッシュヒット時）が0.003625ドル（約0.58円）、出力トークンが0.87ドル（約138.73円）に設定されている[2]。「57倍安価」という主張の実態を検証すると、これは同等性能とされるClaude Fable 5の推定出力単価（100万トークンあたり50ドル）とDeepSeek V4-Pro 0813の出力単価（0.87ドル）を比較した場合に成立する比例関係である[2]。

$$\text{出力単価比率} = \frac{\$50.00}{\$0.87} \approx 57.47\text{倍}$$

一方、入力トークン単価（キャッシュミス時）で比較した場合、Fable 5の推定入力単価（100万トークンあたり10ドル）に対してDeepSeekは0.435ドルであるため、削減比率は約23倍となる[3]。

$$\text{入力単価比率} = \frac{\$10.00}{\$0.435} \approx 22.98\text{倍}$$

実際の自律型エージェントの運用では、プロンプトの読み込み（入力）とコード生成（出力）が多段階で繰り返されるため、システム全体の総合費用は入力と出力の比率、およびコンテキストキャッシュのヒット率に強く依存する[3]。しかし、長文のコードベースを保持したままプロンプトを投げるケースでは、キャッシュヒット時の価格（$0.003625/1M）が適用されるため、実質的な実行コストを極めて低く抑え込める点が最大のアドバンテージとなる[2]。

さらに、DeepSeekはAPIトラフィックの平準化を目的として、2026年8月16日16:00 UTCより「ピーク / オフピーク時間帯別料金制（Peak / Off-Peak Pricing）」を導入した[2]。時間帯別の料金規定および適用時間帯は下表の通りである[2]。

時間帯区分	適用**UTC**時間 **(**日本時間 **JST)**	**1M**入力 **(**キャッシュヒット**)**	**1M**入力 **(**キャッシュミス**)**	**1M**出力トークン
リリース時標準価格	全時間帯共通 (〜8/16 16:00 UTC)[2]	$0.003625[2]	$0.435[2]	$0.87[2]
オフピーク料金	ピーク時間帯以外の全時間帯[2]	$0.022[2]	$0.66[2]	$1.98[2]

ピーク料金	01:00-04:00 & 06:00-10:00 UTC (10:00-13:00 & 15:00-19:00 JST)[2]	$0.044[2]	$1.32[2]	$3.96[2]

ピーク時間帯は1日の中で計7時間設定されており、日本時間（JST）では日中の業務時間帯（10:00～13:00、15:00～19:00）と直接重なっている[2]。オフピーク時間帯の価格はピーク時の半額に設定されているため、企業ユーザーや大規模バッチ処理を行う開発者は、タスクの実行タイミングをオフピーク時間帯へスケジューリングすることでインフラコストを大幅に削減することが可能となる[2]。

エコシステム統合と開発運用上の留意点

DeepSeek-V4-Pro-0813のリリースに伴い、APIの機能面および開発ツールとの統合環境も拡充された[2]。APIエンドポイントはOpenAI互換のChatCompletions形式（https://api.deepseek.com）およびAnthropic互換形式（https://api.deepseek.com/anthropic）の両方を標準提供しており、開発者は既存SDKの接続先URLを変更するだけでモデルを差し替えることができる[2]。

今回の更新では、Codexをはじめとする高度なコード補完・エージェント環境向けに、OpenAI Responses APIフォーマットへのネイティブ対応が追加された[2]。リクエストボディ内に思考モードのパラメータ（thinking: {"type": "enabled"}）および思考強度（reasoning_effort: "low" | "high" | "max"）を指定することで、複雑な技術課題に応じた推論制御が行える[2]。また、非思考モード時限定で、コードの途中で挿入補完を行うFIM（Fill-In-the-Middle）Completion機能も提供されている[2]。

オープンソースコミュニティへの貢献として、Hugging Face上にはdeepseek-ai/DeepSeek-V4-Proのリポジトリを通じてモデル重み（BF16, FP8, FP4 mixed等）がMITライセンス下で公開されている[3]。ただし、実運用においては注意が必要である[3]。Hugging Face上のモデルカードおよびリポジトリ説明文は2026年4月のアナウンス（arXiv: 2606.19348）に基づくプレビュー仕様の記述が残されており、ダウンロード可能な公開重みが現在API側で提供されている「0813」ポス・トレーニング済み重みと完全に一致するかは公式ドキュメントで明記されていない[3]。したがって、オンプレミス環境やローカル推論サーバー（vLLM, SGLang等）でモデルを展開する場合は、「API提供の0813版」と「Hugging Face公開重み」の間で微小な挙動差分が生じる可能性を念頭に置き、個別に検証を行うことが推奨される[3]。

結論と業界への波及効果

DeepSeek-V4-Pro-0813の正式版リリースは、単なる一企業のモデル更新にとどまらず、生成AI産業全体の経済構造および開発プラクティスに大きな変化をもたらしている[1]。

第一の波及効果は、フロンティア級AIエージェントにおける価格破壊の加速である[2]。Claude Fable 5やGPT-5.5といった米国製の最上位モデルが100万トークンあたり数十ドルの単価を設定するなか、同等レベルのターミナル操作能力・コード修復能力を持つモデルが極めて低い単価で提供されるインパクトは大きい[2]。これにより、大企業のAI推進部門やSaaSベンダーは、バックエンドモデルをDeepSeekへ直接移行するか、あるいは推論処理をDeepSeekで行い最終出力を他モデルで整形する「DeepClaude」のようなハイブリッド構成を採用する動きを強めている[2]。

第二の波及効果は、自律型AIエージェント（Agentic Workflows）の経済的実行可能性（Economic Feasibility）の拡大である[2]。従来、コンテキスト長100万トークンを維持しながら数十回に及ぶ試行錯誤やツール呼出を行うエージェントシステムは、1タスクあたりのAPI費用が高額になりがちであった[2]。DeepSeek-V4-Pro-0813の低いトークン単価、高度なコンテキストキャッシュ、そしてオフピーク料金制の組み合わせは、夜間バッチによる全自動コードベースリファクタリングや常時モニタリング型エージェントの実運用化を現実的なコストで可能にする[2]。
総合的に見て、DeepSeek-V4-Pro-0813は現在の市場において最高水準の価格対性能比（Price-to-Performance）を誇るモデルであり、今後のエンタープライズAIインフラ構築における極めて重要な選択肢として定着していくことが見込まれる[2]。

引用文献

1. DeepSeek V4 Pro 0813がリリース、ユーザーは「最高」「ついに」歓喜の声 (2026/08/13), https://search.yahoo.co.jp/realtime/search/matome/996e31e840af4570a878b1da69eb5f61-1786552800?rkf=1&ifr=matome_mtmrltl
2. 中華AIスタートアップのDeepSeekが「DeepSeek V4 Pro 0813」をひっそりとリリース - GIGAZINE, https://gigazine.net/news/20260813-deepseek-v4-pro-0813/
3. DeepSeek V4-Pro 0813が静かにリリース？公式情報を検証してみた - note, https://note.com/kazu_t/n/n4a2958c942de
4. 【2026年7月版】DeepSeek-V4-Flashとは？性能とAPIでの利用方法について解説！ - note, https://note.com/kazu_t/n/n5f5e11c8ebb8
5. DeepSeek V4 Pro 0813 - API Pricing & Benchmarks | OpenRouter, https://openrouter.ai/deepseek/deepseek-v4-pro-0813
6. DeepSeek V4 Pro 0813 Benchmarks, Pricing & Speed | BenchLM.ai, https://benchlm.ai/models/deepseek-v4-pro-0813
7. DeepSeek-V4-Proの正式版がひっそり公開？料金は据え置き - PC Watch, https://pc.watch.impress.co.jp/docs/news/2132548.html