ChatGPT

DeepSeek-V4-Pro-0813 深層調査レポート

調査基準時点：2026年8月13日（日本時間）

エグゼクティブサマリー

DeepSeekの公式APIドキュメントは現在、`deepseek-v4-pro` が **DeepSeek-V4-Pro-0813** に更新されたことを明記しており、「呼び出し方法は変更なし」と説明している。公式料金ページにもモデルバージョンとして **DeepSeek-V4-Pro-0813** が掲載され、1Mコンテキスト、最大384K出力、thinking/non-thinking、JSON、Tool Calls、Responses API、Anthropic APIをサポートする。したがって、**「0813というAPIモデルが正式に稼働している」ことは一次資料で確認済み**である。 [1]

一方で、「正式版V4-Proのリリース」という表現には留保が必要である。DeepSeekの公式Change Logは、調査時点でも最新項目が **2026年7月31日のV4-Flash-0731** であり、そこでは「V4-Pro正式版は近日リリース」とされているものの、**8月13日のV4-Pro-0813専用リリースノートはまだ掲載されていない**。公式Hugging Faceの `deepseek-ai/DeepSeek-V4-Pro` も本文が依然として「preview version」を説明しており、コレクションの最終更新表示は13日前である。このため、**APIの0813化は確定、専用リリースノートと0813固有の公開ウェイトは未確認**、という区別が最も厳密である。 [2]

技術的には、0813で総パラメータ数・活性パラメータ数・基本アーキテクチャ・事前学習データ量が変わったと示す一次資料は見つからない。公開済みV4-Proは **1.6T総パラメータ、49B活性、1Mコンテキスト**で、V4世代はCSA＋HCAのHybrid Attention、mHC、Muon、32T超トークンの事前学習、およびSFT＋GRPO＋on-policy distillationによる後学習を採用する。0813についてこれらを再設計したという技術報告はなく、巨大な改善がAgent系評価に集中していることから、**0813は新アーキテクチャというよりAgent/コード能力を重点的に強化した後学習更新である可能性が高い**。ただし、これは状況証拠からの推論であり、DeepSeekはFlash-0731については明確に「only re-post-trained」と述べた一方、Pro-0813について同じ説明をまだ公表していない。 [3]

性能面では、DeepSeekの公式コミュニティ由来と報じられる表で **Terminal Bench 2.1: 72.1→87.9、CyberGym: 52.7→83.3、DeepSWE: 12.8→62.7、NL2Repo: 38.5→61.5、DSBench-Hard: 31.1→67.2** と大幅改善が示されている。しかし、この0813表自体は公式Change Logにはまだ掲載されていない。日本語で検証したnote記事も、Terminal-Bench公式リーダーボード上で87.9を確認できず、Harnessや実行条件が不十分として慎重な扱いを求めている。したがって、**「Agent能力が大幅に向上した可能性」は強いが、87.9等を独立検証済みの絶対値として扱うべきではない**。 [4]

独立系Artificial Analysisでは、0813 Maxが **Intelligence Index 53、出力77.6 token/s、TTFT 1.71秒**を記録している。4月のV4-Pro Previewは当時の同社評価で52だったため、一般知能の総合指標ではAgentベンチほど劇的な差は出ていない。これは0813が「モデル全体を一から大型化した更新」よりも「実作業・Agentの成功率を重点的に上げた更新」であるという解釈と整合する。なお、評価インデックスの版が時間とともに更新されているため、52対53は厳密な同一条件比較ではない。 [5]

意思決定上の結論は、API利用なら「即時PoCに値するが、本番はカナリア＋自社Agent評価を経由」が妥当である。既存の `deepseek-v4-pro` 利用者はコード変更なしで0813へ切り替わるため、これは利点であると同時に**mutable aliasによる無通知モデル更新リスク**でもある。セルフホストについては既存V4-Proウェイトを運用できるものの、**それが0813 APIと同一のチェックポイントだと公式には確認できないため、0813そのものをセルフホストしたい組織はウェイト、commit、hashの明示を待つべき**である。 [6]

判定項目	2026-08-13時点の結論	確度
`DeepSeek-V4-Pro-0813` API	**確認済み**	高
「Previewから正式版へ移行」の意図	**強く支持される**	高〜中
0813専用公式リリースノート	**未公開/確認不能**	高
0813固有Hugging Faceウェイト	**未確認**	高
0813専用実行バイナリ/Release asset	**未確認**	高
1.6T / 49B / 1M基本仕様	**V4-Proとして継続とみられる**	高〜中
新アーキテクチャ	**公表なし**	高
新しい事前学習データ/32T超からの更新	**公表なし**	高
Agent性能の大幅向上	**vendor/community報告は強い、独立検証は初期段階**	中
API移行コスト	**非常に低い**	高
0813のセルフホスト同一性	**未証明**	高

公式アーティファクトとリリース時系列

V4の一次資料を時系列に並べると、0813が「API上では確実に切り替わったが、通常のオープンモデルリリースに付随する全アーティファクトが同時に更新されていない」という非対称なリリースであることが分かる。[7]

日付	アーティファクト	内容・意味	0813との関係
2026-04-24	DeepSeek公式 V4 Preview / API	V4-Pro、V4-FlashをOpenAI ChatCompletions/Anthropic APIで提供開始	Pro Previewの起点。[8]
2026-04-24	公式HF `DeepSeek-V4-Pro`	1.6T/49B、1M、Hybrid Attention、mHC、Muon、32T超学習などを公開	現在も本文はPreview説明。[9]
2026-04-26	DeepSeek-V4 Technical Report	V4のアーキテクチャ・長コンテキスト効率を論文化	調査時点で0813固有改訂は確認できない。[10]
2026-07-24	旧alias終了予定日	`deepseek-chat` / `deepseek-reasoner` の移行期限	V4系aliasへの移行を促進。[8]
2026-07-31	V4-Flash-0731 正式版	Agent性能向上、Responses API/Codex対応。「アーキテクチャとサイズはPreview同一、再後学習のみ」と公式明記	同日にProは未変更、「正式版は近日」と明記。[11]

日付	アーティファクト	内容・意味	0813との関係
2026-08-13 JST確認	API Quick Start	`deepseek-v4-pro` が **DeepSeek-V4-Pro-0813** に更新されたと公式明記	**0813稼働の最も強い一次証拠**。[12]
2026-08-13 JST確認	Models & Pricing	0813、1M、384K、Responses/Anthropic API、料金を掲載	**正式API仕様を確認可能**。[13]
2026-08-13 JST確認	Change Log	最新記事は依然7月31日	**0813専用release noteなし**。[11]
2026-08-13 JST確認	Hugging Face	`DeepSeek-V4-Pro` safetensorsは存在するが、カード本文はPreview、コレクションは13日前更新	**0813固有ウェイトとは確認できない**。[14]
2026-08-13 JST確認	GitHub	DeepSeek公式にはAgent統合用 `awesome-deepseek-agent` は存在	V4専用の0813リリースリポジトリ/Release assetは確認できない。[15]

配布物について重要な区別がある。 公式Hugging Faceには通常の `DeepSeek-V4-Pro` モデルウェイトがSafetensorsとして存在し、ローカル実行・weight conversion用の `inference` フォルダへの案内もある。モデルカードはMITライセンスを明記している。したがって「V4-Proの公開ウェイトがない」という意味ではない。問題は、**8月13日のAPI版0813とそのウェイトがbit-identicalであることを示す0813名のrepo、commit、revision、hashが確認できない**点である。[16]

従来型ソフトウェアの `.exe`、`.so`、tarballのような**「DeepSeek-V4-Pro-0813配布バイナリ」は確認できなかった**。公開形態は基本的にモデルチェックポイント＋推論ランタイムであり、既存HFページはDocker Model Runner、量子化派生、ローカル推論を案内している。0813専用バイナリが必要という意味では「現時点では無し」と扱うのが安全である。[17]

リリース差分とアーキテクチャ

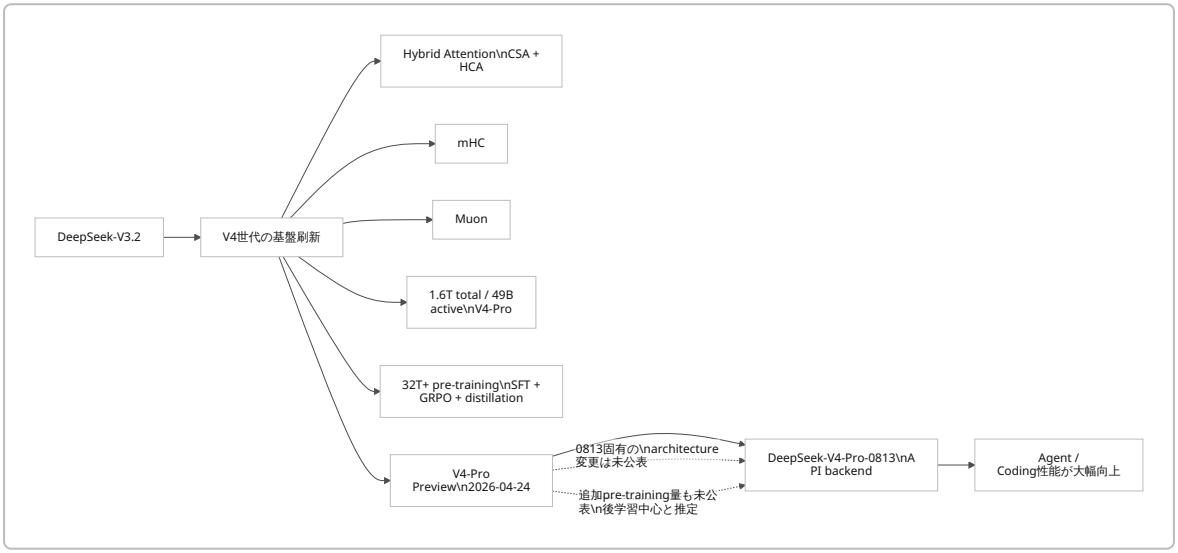
V4世代そのものはV3.2に対して大きなアーキテクチャ変更を行った。DeepSeekの公式モデルカードによれば、V4-ProはMoEで1.6T総パラメータ・49B活性、Hybrid Attentionとして **Compressed Sparse Attention（CSA）＋Heavily Compressed Attention（HCA）**、残差経路にmHC、学習にMuonを導入した。1Mコンテキスト時には、V3.2比で単一トークン推論FLOPsを27%、KV cacheを10%まで低減するとしている。事前学習は32T超の「diverse and high-quality tokens」、後学習はdomain expertをSFT＋GRPOで育成し、その後on-policy distillationで統合する二段階方式である。[18]

しかし、**Preview→0813の差分として新しいCSA/HCA、MoE構成、パラメータ増大などが発表されたわけではない**。現在のAPI版も外部測定・配布情報では1.6T/49B/1Mとして扱われ、公式技術報告に0813固有の新設計はない。したがって「V4アーキテクチャの変更」と「V4-Pro-0813の変更」を混同しないことが重要である。[19]

項目	V4-Pro Preview	V4-Pro-0813	判定
総パラメータ	1.6T	1.6Tとみられる	変更公表なし。[20]

項目	V4-Pro Preview	V4-Pro-0813	判定
活性パラメータ	49B	49Bとみられる	変更公表なし。 20
Context	1M	1M	継続。 21
Max output	V4 API系	**384K**	現行仕様で明示。 13
Attention	CSA＋HCA Hybrid	同一とみられる	0813固有変更なし。 9
mHC / Muon	採用	同一とみられる	変更公表なし。 9
事前学習量	32T超	新規更新量非開示	**0813追加事前学習データは未公表**。 9
後学習	SFT＋GRPO＋distillation	詳細非開示	Agent特化再後学習が有力だが**推定**
Responses API	Preview時点ではPro対応を確認できず、7/31にはFlashのみ明示	✓	実務上の明確なAPI機能差。 22
Anthropic API	✓	✓	継続。 23
FIM	—	non-thinkingのみ	現行仕様で明示。 13
Thinking	対応	thinking / non-thinking、thinkingがdefault	継続・現行仕様明確化。 13

Flash-0731についてDeepSeek自身が「same model architecture and size」「only re-post-trained」と明記したことは、Pro-0813を考える上で有力な手掛かりである。ただし、**Proにも同じ文章を自動的に適用することはできない**。現状の最も保守的な結論は、「0813にarchitecture/pretraining変更を示す証拠はなく、後学習中心と推定されるが、公式なtraining-delta disclosureを待つ」である。 11



この図のV4基盤部分はDeepSeek公式モデルカード・技術報告に基づく。一方、Preview→0813を「後学習中心」とする部分は、公開仕様が不変であること、Agent系の伸びが突出していること、Flash-0731が後学習のみだったことからの**分析上の推定**である。 [3]

性能・ベンチマーク

0813について最も注目されているAgentベンチ表は、DeepSeekの公式グループ/コミュニティから共有されたものと中国メディアが報じているが、**DeepSeekの公開Change Logにはまだ載っていない**。36Kr系報道も明示的にこの点を留保している。一方、この表のV4-Flash-0731列はDeepSeek公式7月31日Change Logの9指標と一致しており、完全な出所不明データよりは信頼度が高い。以下では「DeepSeek報告値・公開release note未収録」として扱う。 [24]

Benchmark	V4-Pro **0813**	V4-Flash 0731	V4-Pro Preview	Opus 4.8	Fable 5*
HLE without / with tools	**42.7 / 60.0**	37.8 / 51.5	37.7 / 48.2	49.8 / 57.9	**53.3 / 63.0**
Terminal Bench 2.1	**87.9**	82.7	72.1	85.0	**88.0**
NL2Repo	**61.5**	54.2	38.5	**69.7**	—
CyberGym	**83.3**	76.7	52.7	78.3	83.1
DeepSWE	**62.7**	54.4	12.8	58.0	**70.0**
Toolathlon-Verified	**74.1**	70.3	55.9	76.2	**77.9**
Agents' Last Exam	**25.7**	25.2	16.5	**25.7**	—
AutomationBench Public	**31.8**	25.1	12.8	27.2	29.1
DSBench-FullStack	**71.1**	68.7	41.8	71.6	**77.2**
DSBench-Hard	**67.2**	59.6	31.1	**71.7**	68.3

* Fable 5はfallback構成。表は同一の独立運営者による完全統制実験ではなく、Harness・effort・tool環境に依存する「reported scores」である。元表はHacker Newsでも全文再掲され、別媒体でも同じ数値が報じられている。 [25]

Previewから0813への差は、Terminal Benchで**+15.8ポイント（相対+21.9%）**、NL2Repo +23.0、CyberGym +30.6、DeepSWE +49.9、Toolathlon +18.2、Agents' Last Exam +9.2、AutomationBench +19.0、DSBench-FullStack +29.3、DSBench-Hard +36.1ポイントとなる。特にDeepSWEは12.8→62.7で、単純なチャット品質の微調整では説明しにくい大きなAgent/Software Engineering改善を示唆する。元スコア自体は前述の「報告値」である。 [26]

なおClineが「Terminal Benchで15.8%向上」と表現したことが広く転載されたが、72.1→87.9は正確には**15.8 percentage points**、相対改善率なら約21.9%である。日本語noteの初期検証もこの点を指摘するとともに、87.9が調査時点のTerminal-Bench公式verified leaderboardに掲載されていないこと、Harness・試行回数・詳細ログが不足していることを警告している。 [27]

独立評価ではより穏当な差が見える。 Artificial Analysisの現行0813 MaxはIntelligence Index v4.1.1で53、出力77.6 token/s、TTFT 1.71秒である。同IndexはGDPval-AA v2、τ^3-Banking、Terminal-Bench v2.1、

SciCode、HLE、GPQA Diamond、CritPt、AA-Omniscience、AA-LCRの9評価を含む。4月時点のPreview Maxは同社の当時のIndexで52、V3.2は42だったため、0813の最も特徴的な差は「総合知能の桁違いのジャンプ」ではなくAgent実行能力にあるように見える。ただしIndexの版が異なるため52→53は厳密な回帰比較ではない。 [5]

運用指標	V4-Pro-0813	V4-Flash-0731	解釈
Artificial Analysis Intelligence Index	**53**	52	Proの品質優位は小さい。 [28]
API output speed	**77.6 tok/s**	120.6 tok/s	throughput優先ならFlashが明確に速い。 [28]
TTFT	**1.71 s**	—	0813は同クラス中央値より速いとAA評価。 [29]
Input $/1M cache miss	**$0.435**	$0.14	Proは約3.1倍。 [13]
Output $/1M	**$0.87**	$0.28	Proは約3.1倍。 [13]
Cache-hit input $/1M	**$0.003625**	$0.0028	長いAgent loopで極めて安価。 [13]
Context	**1M**	1M	同一。 [13]
Max output	**384K**	384K	同一。 [13]

Cline等による「Fable 5の約57分の1」という表現は、**出力単価 $50 / $0.87 ≈ 57.5** を比較したものであって、実タスク総コストが常に57分の1という意味ではない。Fable 5の公開価格は入力$10/M・出力$50/M、DeepSeek現行価格は$0.435/M・$0.87/Mであるため、価格優位そのものは非常に大きいが、キャッシュ率、思考トークン、再試行回数、pass@1成功率を含めたTCOで評価すべきである。 [30]

メモリについて0813固有の公式実測値は公開されていない。 V4アーキテクチャとしては1Mコンテキスト時のKV cacheがV3.2の10%と公式に説明されている。公開V4-Proチェックポイントを LambdaがHGX B200上でvLLM運用した参考値では、8192入力/2048出力、32並列、512 promptでaggregate generation throughput 911.92 tok/s、総throughput 4,561.38 tok/s、平均TTFT 1.186秒、per-user generation 28.5 tok/sだった。ただしこれは**公開V4-Proチェックポイントのself-host benchmarkであり、API版0813を測った値ではない**。 [31]

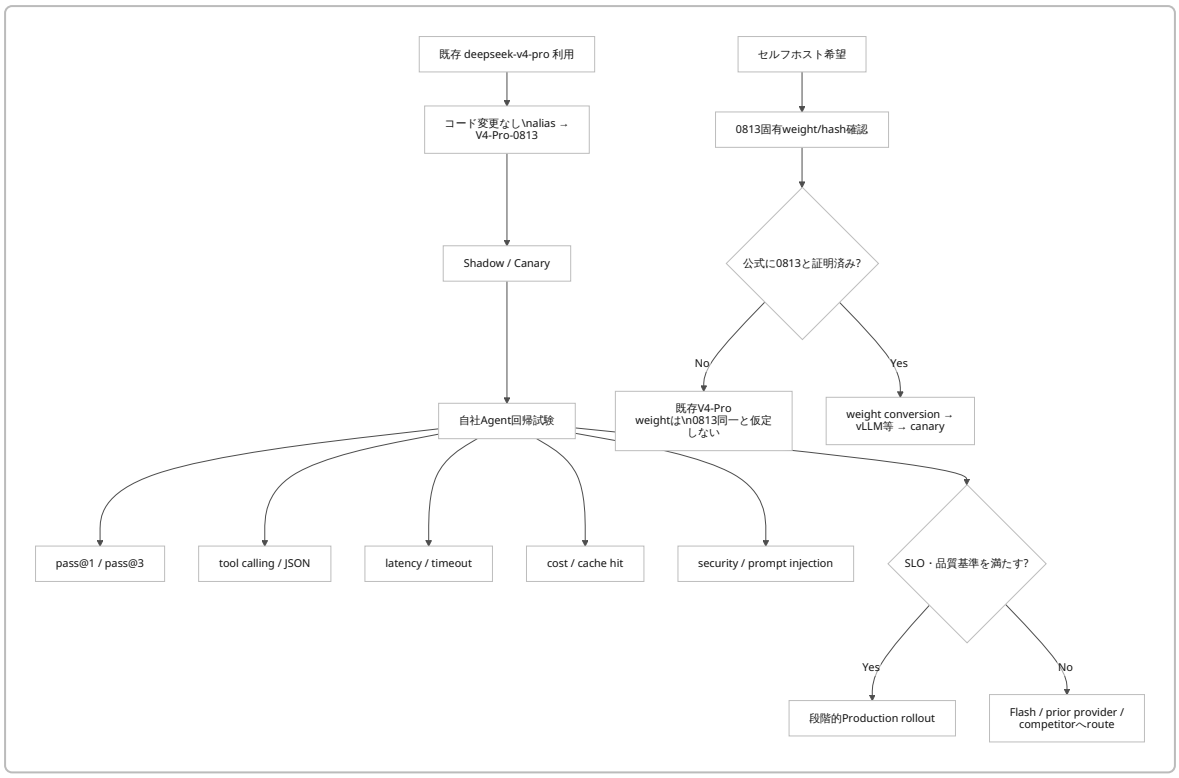
LambdaはフルT1.6モデルについてHGX B200ノードを最低構成として挙げ、FP4 expert＋FP8その他、FP8 KV cache、expert parallel、DeepGEMM MegaMoE等を用いている。この情報はセルフホスト容量設計の参考にはなるが、0813ウェイト同一性が確認されるまでは「0813のハードウェア要件」と断言できない。 [32]

互換性・移行・導入

API利用者にとって最大の利点は、**モデルIDを変更しなくてよい**点である。公式Quick Startは `deepseek-v4-pro` を使えば現在のDeepSeek-V4-Pro-0813へ到達すると明記している。OpenAI形式のbase URLとAnthropic形式のbase URLも維持される。既にV4-Pro Previewを利用しているアプリケーションでは、実装上は「移行」ではなくバックエンドモデルの差し替えに近い。 [12]

ただしこれは、同じaliasの意味が変わる**mutable alias**でもある。本番システムでは、コードに変更がなくてもモデル挙動が変わり得るため、 `deepseek-v4-pro` を固定仕様と考えるべきではない。特にAgent、

function calling、JSON schema、拒否判定、長いmulti-turn loopでは回帰試験が必要である。公式規約もサービス内容が更新・変更され得ることを前提としている。 33



DeepSeekは現在、Responses API、Anthropic API、Tool Calls、JSON OutputをPro/Flash双方で提供する。公式`awesome-deepseek-agent`リポジトリにはClaude Code、Cline、Codex、GitHub Copilot、OpenCode、Qwen Code、Reasonix等の統合ガイドがあるため、**Coding Agent/CLI用途が0813の最も自然な導入先**である。 34

V4-Pro Previewからの具体的な機能差として特に重要なのはResponses APIである。7月31日の公式Change LogではResponses API/Codex対応がFlash正式版の特徴として記され、その時点ではPro APIは「unchanged」だった。現在のPro-0813料金・機能表ではResponses APIが✓になっているため、**Proの正式版化に伴うAPI側の実用上の追加機能**と判断できる。 22

一方、Tool Callingには旧Preview時代から互換性上の注意点がある。DeepSeek公式GitHub上のissueでは、thinking modeで`tool_choice="required"`や特定function強制指定がHTTP 400になるとの報告があり、`auto`は動作するとされている。これはユーザー報告であり0813で完全に再検証された公式仕様ではないため、**OpenAI互換＝全parameter semanticsまで完全互換ではない**と考えてテストするべきである。 35

Claude Codeのauto classifierではPreview時代にthinkingによるタイムアウトが報告されていたが、同issueの8月13日追記では**0813がlow thinking levelをサポートし、thinking enabled＋lowがauto modeで良好**と報告されている。これは興味深い初期改善シグナルだが、公式release noteではなく単一ユーザーのissue更新である。 36

セルフホストでは、公式HFがweight conversionを含む`inference`フォルダを案内している。実運用例としてLambdaはvLLM、DeepGEMM MegaMoE、expert parallel、FP8 KV cache、FP4 indexer cache、DeepSeek V4 tokenizer/tool/reasoning parser、MTP speculative decodingを組み合わせている。したがって、一般的なTransformersモデルを単純にロードするより、**DeepSeek-V4固有のtokenizer、tool parser、MoE kernel、mixed precisionを正しく揃えることが重要**である。 37

RTX 5090 / RTX Pro 6000 Blackwell等のSM120については、4月にDeepSeek公式DeepGEMM issueでV4用の一部SM120 kernel不足が報告されている。0813導入時点で全framework経路が解消済みだと本調査では確認できていないため、consumer/professional Blackwellを採用する場合は**現在使用するDeepGEMM/vLLM revisionで実機検証してからGPUを確定**すべきである。 [38]

導入ユースケースの優先順位は、コードベース横断修正、terminal agent、CI失敗修復、リポジトリ生成/移行、長文ドキュメント分析、低コストのmulti-agent orchestrationが高い。公式統合エコシステムとAgentベンチ改善がこれを支持する。一方、Artificial Analysisは0813を**text-only、image input非対応**として測定しており、画像・スクリーンショット・GUI視覚操作を必須とするAgentには別のVLMを組み合わせる必要がある。 [39]

0813固有の成熟した企業本番導入事例は、リリース当日のためまだ確認できない。 Cline等は即日利用可能としており、OpenRouterも0813の提供開始を告知しているが、これは「配信・ツール統合」であって長期的なproduction case studyではない。 [40]

セキュリティ・ライセンス・運用リスク

調査したNVD情報では、**DeepSeek-V4-Pro-0813モデル/APIそのものに固有と確認できるCVEは特定できなかった**。ただしV4エコシステムのサードパーティ製 `@arikusi/deepseek-mcp-server` には2026年7月公開の脆弱性が存在する。CVE-2026-55604は認証主体にbindされていないsession IDにより他セッションのconversation contextを取得できる問題で、1.7.0で修正された。CVE-2026-55605はself-host HTTP transportの認証欠如により、ネットワーク到達可能な攻撃者がMCP toolやサーバー側DeepSeek API keyを利用し得る問題で、1.8.0で修正された。**いずれもDeepSeek本体ではなく第三者MCPサーバーの脆弱性**である。 [41]

この区別は重要である。Agent性能の向上は同時に「モデルにより多くの権限を渡したときのblast radius」を広げる。報告値とはいえCyberGym 83.3という強いサイバーAgent能力を考えると、productionではshell、filesystem、Git、cloud IAM、database、MCPへのアクセスを**sandbox、least privilege、allowlist、短寿命credential、human approval**で制限するのが妥当である。CyberGymの高さは脆弱性そのものではないが、強力なtool-using modelに通常以上の権限を持たせない理由にはなる。 [42]

ライセンスについて、公式HFモデルカードは **repositoryおよびmodel weightsをMIT License** と明記している。一般にMITは改変・配布・商用利用を許容する非常に緩いライセンスであり、したがって**現在公開されているV4-Proウェイトの商用利用は原則可能**と読める。ただし前述の通り、その公開チェックポイントが0813そのものだという公式同一性は未確認なので、「V4-Pro open weightsのライセンス」と「0813 API buildの配布権」を混同すべきではない。 [43]

API側についても、DeepSeek Open Platform TermsはAPIを下流システム、アプリケーション、内部・外部ユーザー向け機能に統合することを明示的に許容している。Inputの権利は利用者に残り、DeepSeekがOutputについて有する権利は利用者へassignするとされ、Outputは個人利用、研究、派生製品開発、他モデルのtraining/distillation等に利用できるとしている。したがって、**通常の商用SaaS組み込み自体を禁止する条項は見当たらない**。もちろん適用法、利用規約、第三者知財等への適合は利用者責任である。 [44]

また規約はサービスを“as is / as available”で提供し、無停止、安全、error-free、出力の正確性・最新性・非侵害性・安全性を保証しないとする。金融、医療、法務、採用、アクセス制御等の高影響用途では、LLM出力を最終決定として直接実行する設計は避け、deterministic validationまたはhuman reviewを置くべきである。 [45]

データ所在は企業導入上の重大論点である。 DeepSeek Privacy Policyは、収集したPersonal Dataが利用者の居住国外のサーバーに保存され得ることに加え、サービス提供のためPersonal Dataを**中華人民共和国で直**

接収集・処理・保存すると明記している。機密ソースコード、顧客PII、輸出管理データ、規制業界データを公式APIへ送る場合は、法務・セキュリティ・データガバナンス部門による事前評価が必要である。 [46]

さらに、現行公式料金ページは「近い将来API全体を**significantly**値上げする予定」と明記している。現在の驚異的な価格差を3年TCOにそのまま外挿するのは危険であり、vendor lock-inを避けるためrouting abstractionを維持すべきである。 [13]

コミュニティ検証、リスクと推奨事項

日本語一次検証に近い情報として、GIGAZINEは8月13日に公式API料金ページで0813を確認し、ClineやChrisGPTが報告したTerminal Bench、CyberGym、DeepSWEの改善を紹介している。ただしそれらのベンチ値そのものはDeepSeek公開Change Log由来ではない。 [47]

より慎重な日本語検証としてnoteのkazu_t氏は、公式APIで実際の応答を確認しつつ、**「0813 API availability」と「0813 weights release」を分けて考えるべき**と指摘している。また87.9が当時のTerminal-Bench公式リーダーボードで確認できないこと、「15.8%」ではなく15.8ポイントであること、「57倍安い」は出力token単価だけの比較であることを検証しており、現時点では非常に有用なファクトチェックである。 [48]

中国語圏の初期報道でも同じ構図が見える。36Kr系記事は、DeepSeek公式API文書がPreviewから0813へ切り替わったことは確認しつつ、性能表は公式グループから流通したもので、**公開update logがまだ存在しないため厳密にはofficial public benchmark releaseではない**と明記している。これは本調査の証拠評価と一致する。 [49]

コミュニティの実使用評価はまだ割れている。Hacker Newsではベンチ表が広く共有される一方、ある初期利用者は0813を **“most unreliable model I have tried”** と評し、pass@3では強いもののpass@1で失敗しやすいという印象を報告している。単一ユーザーの逸話であり一般化はできないが、Agentでは平均benchmarkだけでなく**first-attempt success rateと再試行コスト**を測るべきという重要な警告である。 [50]

逆にGitHubのClaude Code統合issueでは、Previewで問題だったauto-classifier timeoutに対し、0813ではlow thinking levelが使えるようになったとの初期報告がある。つまりコミュニティ観測を総合すると、**Agent capabilityは向上したが、安定性・thinking設定・Harnessとの相互作用はまだ初日評価段階**とみるのが合理的である。 [36]

意思決定者・エンジニア向けの推奨を優先度順に整理すると次の通りである。

優先度	推奨アクション	理由
P0	`deepseek-v4-pro` をproductionへ即時100%投入せず、shadow→5–10% canary→段階拡大	aliasが既に0813へ切り替わっており、コード変更なしに挙動が変わる。 [12]
P0	Agent評価を**pass@1、pass@3、再試行数、wall-clock、実コスト**で再測定	vendor benchmarkの大幅上昇とcommunityのpass@1懸念を同時に検証できる。 [51]
P0	shell/MCP/cloud権限をsandbox＋least privilege化	Agent能力が高く、第三者MCPでは実際の認証系CVEも確認されている。 [41]

優先度	推奨アクション	理由
P0	機密データ投入前にPRCデータ処理を法務・情報管理部門で承認	Privacy Policyが中国での直接処理・保存を明記。 46
P1	Tool Callで `auto` だけでなく `required`、specific function、JSON schema、stream/non-streamを回帰試験	OpenAI互換でもthinking modeのparameter semantics差異が報告済み。 35
P1	0813セルフホストは**0813固有repo/revision/hashが出るまで「既存HF＝0813」とみなさない**	HFカードが依然Preview説明で、0813専用artifactが確認できない。 14
P1	高品質タスクをPro、volume/latency-sensitiveタスクをFlashへroute	AAでPro 53 vs Flash 52だが、Flashは120.6 vs Pro 77.6 tok/sで約55%高速。 28
P1	価格は現在値ではなく2–5倍シナリオでもTCO試算	DeepSeek自身が近い将来の大幅値上げを予告。 13
P2	1M contextを「全部詰め込む」ためでなくrepo/document retrieval＋cache reuseに活用	長文処理能力とcache価格は強力だが、長コンテキストはlatency/attention品質も実測すべき。 21

採用判断としては「API PoC＝Go」「一般production＝条件付きGo」「規制・高機密データのDeepSeek公式API＝要個別審査」「0813そのもののセルフホスト＝現時点ではHold」を推奨する。品質/価格比は非常に魅力的で、特にTerminal、コード修復、repository-scale Agent、多段tool useでは有力な候補である。一方、正式release note、0813-specific weights/hash、再現可能なbenchmark logsが揃っていないことから、今日の段階で「Fable 5相当」「Previewから完全互換」「公開ウェイトも0813」と断定するのは証拠を超えている。 52

主要一次資料・検証ソース： DeepSeek公式「Your First API Call」 12 、公式「Models & Pricing」 13 、公式Change Log 53 、公式Hugging Face `DeepSeek-V4-Pro` 54 、DeepSeek-V4 Technical Report / arXiv 10 、公式 `awesome-deepseek-agent` 15 、DeepSeek Open Platform Terms 55 、DeepSeek Privacy Policy 46 、Artificial Analysis 0813独立評価 29 、Lambda V4-Pro self-host benchmark 32 、日本語GIGAZINE 47 、日本語noteファクトチェック 48 、36Kr/Phoenix系初期報道 49 、Hacker News初期議論・ベンチ表 51 、NIST NVD CVE-2026-55604 / CVE-2026-55605 41 。

---

1 6 12 33 Your First API Call | DeepSeek API Docs
https://api-docs.deepseek.com/

2 3 7 8 11 22 53 Change Log | DeepSeek API Docs
https://api-docs.deepseek.com/updates

4 24 49 Liang Wenfeng Challenges Elon Musk: DeepSeek V4 Pro Outperforms & Pressures the World's Most Powerful AI Model
https://eu.36kr.com/en/p/3937236905344390?utm_source=chatgpt.com

5 19 28 29 DeepSeek V4 Pro 0813 (max) - Intelligence, Performance & Price Analysis
https://artificialanalysis.ai/models/deepseek-v4-pro

9 14 16 17 18 20 31 37 43 54 deepseek-ai/DeepSeek-V4-Pro · Hugging Face
https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro

10 [2606.19348] DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence
https://arxiv.org/abs/2606.19348

13 21 23 34 52 Models & Pricing | DeepSeek API Docs
https://api-docs.deepseek.com/quick_start/pricing/

15 39 GitHub - deepseek-ai/awesome-deepseek-agent · GitHub
https://github.com/deepseek-ai/awesome-deepseek-agent

25 26 42 50 51 DeepSeek V4 Pro 0813
https://news.ycombinator.com/item?id=49274600&utm_source=chatgpt.com

27 40 47 中華AIスタートアップのDeepSeekが「DeepSeek V4 Pro ...
https://gigazine.net/news/20260813-deepseek-v4-pro-0813/?utm_source=chatgpt.com

30 Claude Fable 5 (with fallback)
https://artificialanalysis.ai/models/claude-fable-5?utm_source=chatgpt.com

32 deepseek-ai/DeepSeek-V4-Pro
https://lambda.ai/inference-models/deepseek-ai/deepseek-v4-pro

35 [BUG] DeepSeek V4 rejects tool_choice="required" and ...
https://github.com/deepseek-ai/DeepSeek-V3/issues/1376?utm_source=chatgpt.com

36 Claude Code with Deepseek Pro V4, classifier gets time-out or failed constantly in auto mode · Issue #284 · deepseek-ai/awesome-deepseek-agent · GitHub
https://github.com/deepseek-ai/awesome-deepseek-agent/issues/284

38 DeepSeek-V4 on SM120: tf32_hc_prenorm_gemm and ...
https://github.com/deepseek-ai/DeepGEMM/issues/317?utm_source=chatgpt.com

41 NVD - CVE-2026-55604
https://nvd.nist.gov/vuln/detail/CVE-2026-55604

44 45 55 DeepSeek Open Platform Terms of Service
https://cdn.deepseek.com/policies/en-US/deepseek-open-platform-terms-of-service.html

46 DeepSeek Privacy Policy
https://cdn.deepseek.com/policies/en-US/deepseek-privacy-policy.html

48 Did DeepSeek V4-Pro 0813 Release Quietly? Verifying the ...
https://note.com/kazu_t/n/n4a2958c942de?hl=en&utm_source=chatgpt.com