ChatGPT

DeepSeek-V4.1-Flash 深掘り分析レポート

エグゼクティブサマリー

DeepSeekは**2026年9月10日**、新アーキテクチャ系列の最小モデルとして「**DeepSeek-V4.1-Flash**」を正式発表した。公式発表ページは日付を9月10日と明記しているが、**発表そのものの正確な時刻は未指定**である。Reutersは同日06:32 UTCにローンチを報じた。モデルは**552Bパラメータ級MoE（Mixture-of-Experts）**、最大**100万トークン**のコンテキスト、**ネイティブ画像理解**を備え、APIでは `deepseek-flash` として既に提供されている。モデル重みも公開され、ライセンスは**MIT License**である。 [1]

本モデルの最大の技術的特徴は、単に「画像が読めるV4 Flash」になったことではない。20層のCausal Encoderと20層のDecoderから成る新しい**Causal Encoder-Decoder（CED）**構造により、入力処理時の活性化パラメータを**8B**、デコード時を**16B**に非対称化した。さらにCSA2（Compressed Sparse Attention 2）、SWA Bounded Replay、FP4 KV Cache、Engram conditional memory、DSpark speculative decodingを組み合わせ、グローバルKVキャッシュを**890 bytes/token**まで圧縮した。DeepSeekはV4-Flash比でHBM需要を約4分の1、永続KVキャッシュを約8分の1、初代DeepSeek比ではKVキャッシュを437分の1にしたとしている。これは100万トークン級のエージェント処理で「計算能力」よりも「コンテキストを保持するメモリ帯域・容量」がボトルネックになりつつあることへの、かなり明確な設計回答である。 [2]

画像能力も後付けの画像エンコーダではなく、DeepSeek-ViTをスクラッチから学習し、画像埋め込みとテキスト埋め込みを言語モデルの事前学習開始時から共同処理したと説明されている。APIではJPEG、PNG、GIF、WebPを扱い、1画像あたり最大1024 visual tokens。大きな画像は総画素数がおおむね1300×1300相当に自動縮小され、API上の最大辺は通常8192 pxである。したがって「画像生成モデル」ではなく、**画像・スクリーンショット・図表・文書を読み、テキストを生成するimage-to-text型マルチモーダルモデル**と理解すべきである。 [3]

価格は極めて攻撃的である。100万トークン当たり、キャッシュミス入力は**$0.15 off-peak / $0.30 peak**、出力は**$0.60 / $1.20**、キャッシュヒット入力は**$0.003 / $0.006**。現在のV4 Proのpeak価格$1.32/$3.96と比べても大幅に安い。DeepSeekは9月14日12:00北京時間（**9月14日13:00日本時間**）以降、V4.1 Proが登場するまで `deepseek-v4-pro` をV4.1 Flashへルーティングするとしており、「廉価版Flashが旧世代Proを実質的に置換する」という異例の世代交代になっている。 [4]

性能は、**エージェント型ソフトウェア開発に非常に強い一方、「あらゆる知能ベンチマークでV4 Proを上回る」という意味ではない**。公式評価ではDeepSWE v1.1が74.2、Terminal-Bench 2.1が90.6、CyberGymが88.1、AutomationBenchが54.8で、直接比較されたClaude Opus 5、GPT-5.6 Sol、Kimi K3などに匹敵または一部で上回る。一方、GPQA Diamondは90.9でOpus 5の93.4、GPT-5.6 Solの94.1を下回り、基盤モデルのSimpleQA-Verifiedも42.3でV4 Proの55.2を大きく下回る。LongBench-V2も45.2対51.5でV4 Proが優勢である。したがってDeepSeekの「全面的にV4 Proを超えた」という表現は、**性能・コスト・速度・総処理時間を総合した製品価値の主張**と読むのが妥当で、個々の能力軸では明確なトレードオフが残る。 [5]

また、現時点は**公開初日**であり、本調査で確認できたV4.1 Flashの詳細定量評価はDeepSeek自身によるものが中心である。旧V4 ProについてはArtificial Analysisによる独立評価が存在するが、V4.1 Flashについて同程度に包括的な独立第三者検証は本調査時点で確認できなかった。したがって公式スコアは有望ではあるものの、企業調達の根拠としてはまだ**「ベンダー報告値」扱い**とするのが適切である。 [6]

総合すると、V4.1 Flashの重要性は「中国勢がまた高性能モデルを出した」こと以上に、**巨大モデルの競争軸を総パラメータ数から、活性化パラメータ、KVキャッシュ、エージェント総処理時間、API単価へ移そうとしていること**にある。MITのオープンウェイトと低価格APIを同時に提供するため、コードエージェント、長文書処理、企業内検索、視覚文書理解などでは強い価格圧力を生む可能性がある。一方、552B級モデルであるため「オープンウェイト＝手軽なオンプレミス」という意味ではなく、ハードウェア要件、訓練データ出所、安全性、法務・医療分野での検証は依然として大きな未公開領域である。 [7]

本稿の総合評価： 技術革新性 **高**、エージェント/コード用途の競争力 **非常に高い可能性**、コスト競争力 **非常に高い**、視覚理解 **有望**、一般知識・長文推論の絶対性能 **最上位とは未確認**、独立検証成熟度 **低〜中（公開初日のため）**、企業利用時のプライバシー・規制面 **導入方式によって大きく異なる**。

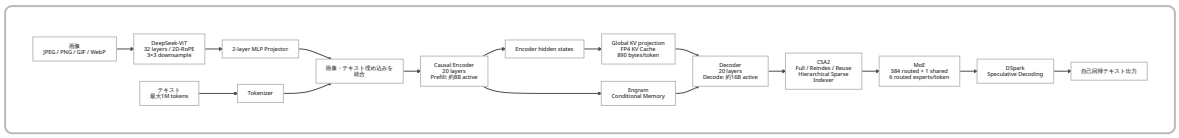
公式発表と技術仕様

発表・提供条件。 DeepSeek公式サイトは2026年9月10日付でV4.1 Flashを発表した。公式ページは「全く新しいモデル構造系列で最小サイズ」と位置付け、能力上限、推論速度、スループット、将来の大型化を設計目標に掲げている。APIの正式モデル名は `deepseek-flash`。旧 `deepseek-v4-flash` と `deepseek-v4-flash-vision-exp` は廃止されているものの、互換性維持のため当面はV4.1 Flashへルーティングされる。WorkBuddy/CodeBuddyおよびOpenCodeも公式パートナーとして統合済みである。 [8]

項目	DeepSeek-V4.1-Flashの公開仕様
正式発表日	2026年9月10日。**正確な公式発表時刻は未公開**。Reuters報道は06:32 UTC。 [9]
APIモデル名	`deepseek-flash`。旧Flash/vision-exp名は一時的な互換エイリアス。 [10]
モデル形態	マルチモーダルMoE、画像＋テキスト入力、テキスト出力。 [11]
公称パラメータ	552B backbone。入力時8B、デコード時16Bを活性化。 [11]
Transformer	40層＝20層Causal Encoder＋20層Decoder。 [11]
コンテキスト	1,048,576 positions ≒ 1M tokens。最大API出力384K。 [12]
MoE	384 routed experts＋1 shared expert、token当たり6 routed expertsを選択。 [13]
視覚モデル	DeepSeek-ViT、32層、hidden 1024、16 heads、patch 14、3×3 pixel-unshuffle。 [14]
画像入力	JPEG / PNG / GIF / WebP。base64、公開URL、Files API `file_id`。 [15]
画像トークン	最大1024 visual tokens/画像。 [16]
画像寸法	API最大辺8192 px。15枚以上では4096 px。大画像は推論前に約1300×1300総画素相当へ縮小。`detail=low` は512×512。 [15]
画像枚数	最大600画像/リクエスト。 [15]
キャッシュ	global KV 890 bytes/token、V4-Flash比約1/4。永続KVは約1/8。 [17]
推論モード	thinking / non-thinking。技術カード上は `reasoning_effort` 1〜100。 [18]
重み	Hugging Faceで公開。Transformers、vLLM、SGLang向け利用例あり。 [19]
ライセンス	MIT License。モデルカードはモデル重みとリポジトリをMITとしている。 [20]

項目	DeepSeek-V4.1-Flashの公開仕様
事前学習量	マルチモーダルコーパス45T tokens。64Kでsparse attentionを学習し、34T tokens時点で1M contextへ拡張。 11
学習データ内訳	**未公開**。Web、書籍、コード、画像、OCR、言語別割合、権利処理等の詳細は公開資料にない。 11
訓練ハードウェア	**V4.1 Flashについて未公開**。旧V4-Flashの一部訓練にHuawei Ascend 950が使われたとの情報はあるが、V4.1への外挿は不可。 21
raw推論速度	tokens/sec、TTFT、GPU別throughputは**未公開**。API同時実行上限は2500。 10
最小VRAM/GPU要件	**未公開**。公式は「2k GPUs＋storage cluster」を持つ大規模展開希望者には連絡を求めるが、これは最小要件の表明ではない。 8
公式API Fine-tuning	V4.1専用マネージドfine-tuning条件は**未公開**。オープンウェイトなので自前カスタム化はMITライセンス上可能。 20

CEDの意味。 従来のデコーダ専用LLMでは、長い入力について各デコーダ層が自身のKV状態を保持する設計がメモリを大量消費しやすい。V4.1では最初の20層をCausal Encoderとして入力処理に特化させ、Decoder側のglobal KVをencoder最終hidden stateから投影する。この構造が、入力の多いエージェント用途で8Bだけを活性化する仕組みの中核である。さらにCSA2ではFull/Reindex/Reuseという3種類のattention modeによりKVやTop-K indexを層間共有し、Hierarchical Sparse Indexerで長文ほど増えやすいindex探索コストを抑えている。 11



この設計のポイントは、**大モデルの全パラメータを各トークンで使わないだけでなく、長文推論のメモリ状態そのものを徹底的に削っている**点にある。特にAgentでは、コードベース、ツールログ、Web取得結果、スクリーンショットなどが会話履歴へ蓄積していくため、1M contextを「理論上入れられる」だけでなく「安価に維持できる」ことの方が商用価値を左右しやすい。DeepSeekがキャッシュヒット価格を100万tokens当たり最低$0.003に設定しているのは、このアーキテクチャ上の優位を直接料金戦略へ反映したものと解釈できる。 22

パラメータ数には公開情報上の小さくない不整合がある。 DeepSeek公式発表とモデルカード本文は552Bを明記する一方、Hugging Faceの自動表示では「Model size 485B params」と表示されている。またモデルカードは別途Engram conditional memoryを196B parametersと説明している。どの成分をheadline parameter countへ含めるかの完全な会計は公開資料だけでは一意に再構成できない。本稿では、DeepSeekが公式本文で繰り返している**552Bを正式なbackbone規模**として扱い、485BのHF表示やEngramとの関係は**未解明**とする。単純に552＋196＝748Bと断定することは避けるべきである。 23

価格・提供形態。 API価格は以下の通りで、DeepSeekは価格を将来変更する権利を留保している。Peakは平日01:00–04:00 UTCおよび06:00–10:00 UTCで、日本時間では概ね**平日10:00–13:00、15:00–19:00**に相当する。 10

100万tokens当たり	Off-peak	Peak
Input・cache hit	**$0.003**	**$0.006**

100万tokens当たり	Off-peak	Peak
Input・cache miss	**$0.15**	**$0.30**
Output	**$0.60**	**$1.20**

APIはOpenAI互換形式に加えAnthropic互換endpoint、Responses API、JSON output、tool callsをサポートする。これは既存のOpenAI/Anthropic系エージェントスタックからの切り替えコストを抑える重要な商業上の特徴である。 24

性能評価と競合比較

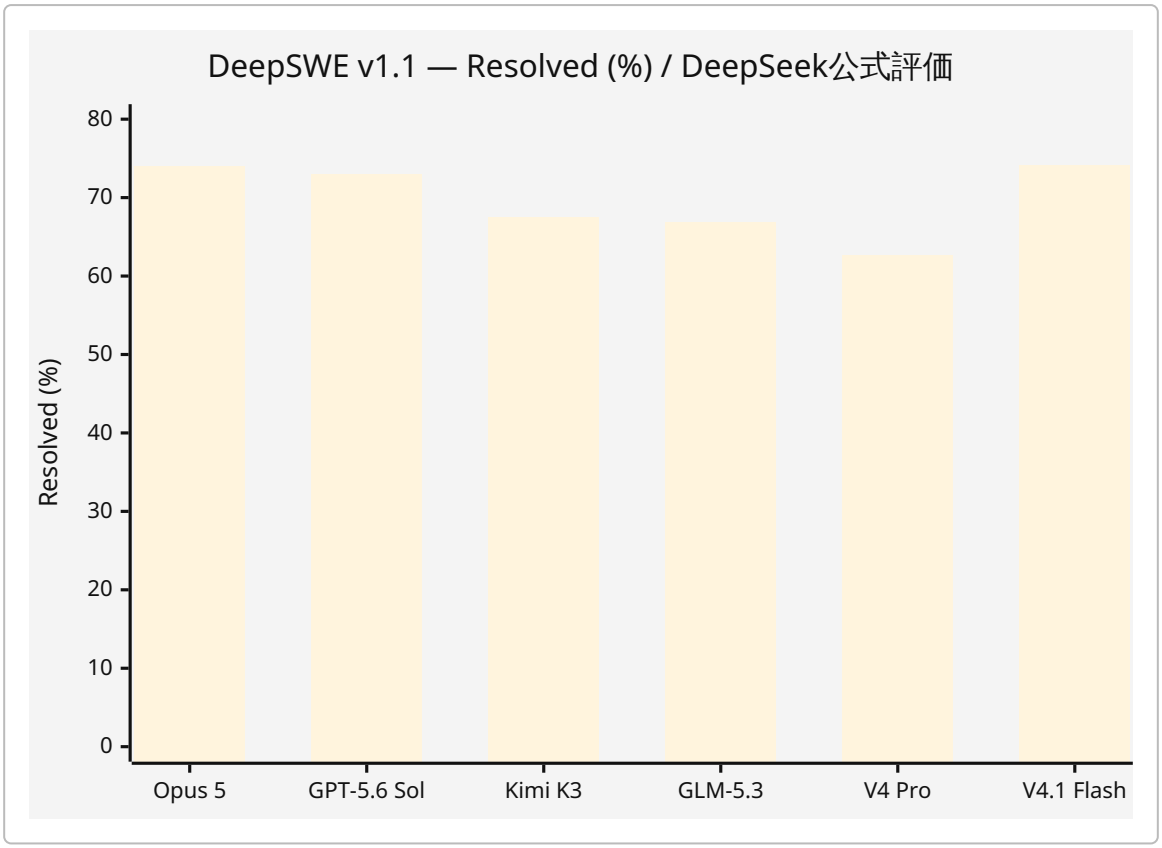
DeepSeekの基盤モデル評価をV4 Proと比べると、V4.1 Flashはコード系で明確に向上している。BigCodeBenchは59.2→60.6、HumanEvalは76.8→79.4、GSM8Kは92.6→93.0、MMLU-Proは73.5→74.1。一方、SimpleQA-Verifiedは55.2→42.3、MATHは64.5→61.1、LongBench-V2は51.5→45.2であり、**世界知識・一部数学・長文理解では旧Proの方が強い指標が残っている**。 25

ベンチマーク	V4 Flash	V4 Pro	V4.1 Flash	V4.1 vs Pro	解釈
MMLU-Pro	68.3	73.5	**74.1**	+0.6	一般高度知識はほぼPro以上
SimpleQA-Verified	30.1	**55.2**	42.3	**−12.9**	事実QAは重要な弱点
BigCodeBench	56.8	59.2	**60.6**	+1.4	コード生成改善
HumanEval	69.5	76.8	**79.4**	+2.6	コード能力が強い
GSM8K	90.8	92.6	**93.0**	+0.4	初中等数学は高水準
MATH	57.4	**64.5**	61.1	−3.4	高度数学ではPro優位
LongBench-V2	44.7	**51.5**	45.2	**−6.3**	1M context ≠ 長文理解性能
DocVQA	—	—	**95.6**	—	視覚文書理解の有望な結果

評価条件はDeepSeek内部フレームワークで統一され、base modelでは0.3以内を同等とみなしている。したがってMMLU-Proの+0.6やGSM8Kの+0.4は実質的には小差である一方、SimpleQAやLongBenchの差は無視しにくい。 25

より特徴的なのが**agentic benchmark**である。最大 `reasoning_effort=100` ではTerminal-Bench 2.1が90.6、DeepSWE v1.1が74.2、CyberGymが88.1、AutomationBenchが54.8、Agent's Last Examが31.8となり、DeepSeekが直接比較したモデル群の最高値を複数項目で記録する。一方Terminal-Bench 3.0/4.0ではClaude Opus 5に大きく劣り、GPQA DiamondでもGPT-5.6 Sol、Opus 5、K3を下回る。つまり**「短〜中距離の実務エージェント/コード実行が特に強く、最高難度の一般推論では絶対首位ではない」**というプロファイルが見える。 26

以下は同一指標で比較できるDeepSWE v1.1の公式値である。これは**DeepSeek自身の評価値で、独立第三者評価ではない**。 26



ただし、エージェント評価には重大な注意点がある。同じV4.1 FlashでもDeepSWE v1.1はscaffoldを変えるだけで**65.5〜74.2**、Terminal-Bench 2.1は**84.1〜90.6**まで変動した。DeepSWEでは8.7ポイント、Terminal-Benchでは6.5ポイントもの振れ幅である。この結果は、「どのLLMか」だけでなく、**tool loop、system prompt、コンテキスト管理、停止条件、shell実行環境などを含むagent scaffoldそのものが製品性能の一部**であることを示す。企業PoCでは公開ランキングよりも、自社で実際に使うagent framework込みでA/Bテストする必要がある。 20

主要競合との比較。 下表の「性能」は主としてDeepSeekが同じ表で直接比較した値、「コスト」は各社公開API価格を基礎にしている。料金体系、cache discount、通貨、推論強度が異なるため、単純な$/tokenだけで総所有コストを比較することはできない。Anthropic Opus 5は標準料金$5/MTok input、$25/MTok output、1M contextで全現行モデルが画像入力対応である。Kimi K3は2.8T parameters、1M context、ネイティブvision、open weightsを公式に掲げる。 27

モデル	公開性能の位置付け	標準的APIコスト	API・導入容易性	マルチモーダル	プライバシー / セキュリティ上の特徴
DeepSeek-V4.1-Flash	DeepSWE 74.2、TB2.1 90.6。Agent系で最上位級。ただしGPQA/HLE等は絶対首位でない。 26	**$0.15/$0.60 off-peak、$0.30/$1.20 peak**（cache-miss input/ output） 10	**非常に高い**。OpenAI/ Anthropic互換＋open weights。	**画像**◯、1画像最大1024 visual tokens。動画・音声は未公開。 15	自前ホストならデータ制御性は高い。一方、公式クラウドサービス利用時は過去のDeepSeekサービスに海外規制当局のプライバシー懸念があり、別途データフロー審査が必要。 28
Claude Opus 5	DeepSWE 74.0、TB2.1 89.1、TB4.0 51.8。難しい長期agentで非常に強い。 26	**$5 input / $25 output / MTok**。 29	**高い**。Claude APIに加えAWS Bedrock、Google Cloud、Microsoft Foundry。 30	**画像**◯、1M context。 29	closed modelで自前weight保持不可。ただし複数企業クラウド経由の選択肢を持つ。 30
GPT-5.6 Sol	GPQA 94.1、DeepSWE 73.0、CyberGym 84.5など。V4.1が直接比較したOpenAIモデル。 26	公開価格ベースでDeepSeekより大幅に高い。	OpenAI APIエコシステムの成熟度が強み。	比較表ではvisual-agent評価対象。	closed model。データ統制は契約/API設定依存。
Kimi K3	GPQA 92.9、DeepSWE 67.5、MathArena Apex 65.6。 26	中国国内API課金。DeepSeekとの直接ドル比較は為替・cache条件調整が必要。	OpenAI SDK互換で導入可能。 31	**画像◯、動画入力も公式例あり**。1M context。 31	**open weights**だが2.8T規模のため自前運用コストは非常に大きいと考えられる。 31
DeepSeek-V4 Pro	一般知識ではV4.1より強い項目が残るが、Agentでは大幅に劣るものが多い。 32	$0.66/$1.98 off-peak、$1.32/$3.96 peak。 10	APIあり。ただし9/14以降V4.1 Flashへ暫定ルーティング予定。	画像×。 10	事実上世代交代局面。

重要な留保として、DeepSeekのV4.1評価表で使われるOpenAIモデルは**GPT-5.6 Sol**だが、2026年9月10日時点のOpenAI Developersトップページは既に**GPT-6 Astraを「most capable model」**として掲載している。

このため「V4.1 Flashが現在の全フロンティアモデルより優れる」という結論は、DeepSeek公開表だけからは導けない。Googleの最新Gemini系列についてもV4.1の直接比較データは公開されていない。 [33]

コスト面では差がより明確である。例えばOpus 5の$5/$25（input/output）に対し、V4.1 Flashはpeakでも$0.30/$1.20で、単純な非cache token価格では**入力約16.7分の1、出力約20.8分の1**である。もちろん実際のagentコストにはthinking tokens、tool execution、cache率、再試行回数が入るが、同程度のタスク成功率を維持できる場合、価格差はかなり大きい。 [34]

応用可能性と商業・産業的影響

最も即効性のある領域は**ソフトウェア開発エージェント**である。DeepSWE、Terminal-Bench、CyberGym、NL2Repoなどが高く、1M contextによって大規模repository、issue、test log、terminal historyを同一セッションへ投入しやすい。また公式APIはtool callsをサポートし、OpenCode、WorkBuddy/CodeBuddyも既に統合している。実際の企業価値は「コードを一回書く」より、issue理解→repository検索→編集→test→エラー解析→修正という反復を低コストで回せる点にある。 [35]

文書・バックオフィスAIも有力である。DocVQA 95.6という公式結果、スクリーンショット文字読み取り、図表分析、最大600画像/リクエスト、1M contextを組み合わせれば、請求書・契約書・決算資料・マニュアル・スキャン帳票・スライドなどを横断するdocument agentを構築できる。ただしDocVQAはLLM-Judge評価であり、会計・法務で要求されるセル単位・金額単位の100%に近い正確性を証明するものではない。 [36]

製造・物流では、作業手順書と現場写真の照合、設備画面やエラーログの読み取り、検査写真の説明、部品カタログ検索などが候補になる。RefCOCO-avg 86.0は画像内対象物のgrounding能力を示す一定の材料だが、専用外観検査システムと同じ欠陥検出精度を保証する指標ではない。安全停止や品質合否の最終判定をV4.1単体へ委ねるのは現状のエビデンスからは正当化できない。 [25]

研究用途では、論文、実験ノート、表・図、コードを一つの長いcontextへ集約し、解析コードをtool executionで作らせる構成が有望である。GPQA 90.9や高いコード能力は強みだが、SimpleQA-VerifiedがV4 Proを大幅に下回るため、文献に存在しない事実をモデル自身の知識だけで補わせる使い方には注意が必要である。「検索・論文DB→原文取得→出典付き回答」というgrounded workflowとの組み合わせが適する。 [32]

法務では契約書の条項抽出、版間差分、due diligence資料の初期整理、証拠資料への索引付けが現実的である。一方、法律解釈や裁判見通しを最終回答として利用するには、管轄法、判例の最新性、引用の実在確認が不可欠である。V4.1専用の法務benchmarkや法律家によるvalidationは**未公開**であり、自律的な法的判断システムとしての適格性は確認されていない。 [32]

医療・ライフサイエンスについても、医学文献整理、非診断的な文書分類、研究データのコード化、画像付き資料の説明補助などには可能性があるが、V4.1 Flashについて臨床画像診断、患者アウトカム、安全性、医療機器規制対応を裏付ける公式臨床試験・医学benchmarkは**未公開**である。したがって診断、投薬、トリアージなど患者安全に直接影響するタスクへの自律投入は、現状の公開エビデンスを超える。これは「モデル性能が低い」という評価ではなく、**用途固有validationが存在しない**という問題である。 [32]

コンシューマ・クリエイティブでは、スクリーンショットを渡して操作法を聞く、写真の内容説明、デザインやWeb UIのレビュー、グラフ読解、画像ベースの学習支援などが自然な用途になる。ただしV4.1自身は画像を**生成・編集するモデルではなく、画像を入力してテキストを返すモデル**であるため、画像生成市場に直接参入したと見るのは誤りである。 [37]

商業的には、V4.1 Flashがもたらす最も大きな変化は**「1回の質問の価格」ではなく「長時間稼働するagentの限界費用」**の低下だと考えられる。典型的なagentは、同じrepositoryや企業文書を何十ターンも参照する

ためcache hit率が高くなりうる。そこで入力cache hit $0.003/MTokという単価とKV cache圧縮が組み合わさると、従来はコスト的に成立しにくかった常時稼働型agent、全社文書横断agent、コードレビューagentなどのビジネスモデルが成立しやすくなる。これはアーキテクチャと料金体系から導く本稿の推論である。[22]

さらにDeepSeekは、8月時点ではV4 ProをFlashより高価なプレミアム製品として展開し、独立評価Artificial AnalysisでもV4 Proは旧V4 Flashを上回っていた。それから約1カ月で「FlashがProを性能・費用・速度・総時間で全面的に上回った」としてProを事実上置換する方針へ転じた。これは、大型AIの価格階層が「Pro＝大きく高価、Flash＝廉価で弱い」という固定構造ではなく、**アーキテクチャ効率の世代差が、旧premium tierを一気にコモディティ化し得る**ことを示している。[38]

DeepSeekがopen weightsを継続していることも市場への圧力になる。ただし552B級backboneなので、中小企業が一般的な数枚のGPUで容易にfull modelを運用できるとは考えにくい。公式自身が大規模展開希望者に「2k GPUsとstorage cluster」を持つ場合の連絡を呼びかけている。したがって市場機会は、モデル販売そのものだけでなく、**量子化・推論最適化、GPU/AIアクセラレータクラスタ、MaaS、オンプレ運用支援、agent orchestration**へ広がる可能性が高い。[39]

中国産AIインフラとの関係も注目点である。旧V4はHuawei Ascend 950向けに最適化され、HuaweiはV4-Flashの訓練の一部にも自社チップを使用したとしていた。OmdiaのLian Jye Suは、この連携が中国企業による国内AIスタック構築の障壁を下げる一方、HuaweiとNvidiaの技術・エコシステム差はなお残るとReutersに述べている。**ただしV4.1 Flashの訓練にどのGPU/acceleratorを何台・何時間使用したかは未公開**であり、「V4.1もHuaweiで訓練された」と断定する根拠はない。[21]

社会・倫理・規制上の論点

誤情報・ハルシネーション。 V4.1について直接利用できる一つの警告シグナルはSimpleQA-Verifiedである。V4 Proの55.2に対しV4.1 Flashは42.3で、大量のagent benchmark改善と事実的world knowledgeの改善が同義ではない。このため、企業検索、ニュース、金融、法務、医療などでは「モデルの内部知識だけで答えさせない」「検索元・社内文書・DBレコードを出典として固定する」という設計が重要になる。[25]

著作権・訓練データ。 DeepSeekは45T-tokenのmultimodal corpusでスクラッチ学習したことを明らかにしているが、データセットの名称、画像の権利処理、Web crawl比率、出版社データの有無、opt-out、重複除去、各国語構成などは公表していない。したがって著作権上の適法性について、本公開資料だけから肯定も否定もできない。企業がモデル重みを自社配布物へ組み込む場合、**MITライセンスがモデル重みの利用を許すことと、学習データや特定出力について第三者権利が不存在であることは別問題**として扱う必要がある。[23]

バイアス・政治的中立性。 過去のDeepSeekモデルを対象にした研究では、中国語・英語間で政治的/地政学的回答傾向が変化するとの報告や、中国語コンテキストで安全性上の弱点を報告する研究が存在する。ただし、これらはR1/V3等の旧モデルを対象としており、**V4.1 Flashへその数値を直接転用することはできない**。V4.1については、日本語・簡体字中国語・繁体字中国語・英語を同じ意味のpromptで比較する新しいcross-lingual bias auditが必要である。[40]

プライバシーとデータ主権。 過去のDeepSeekオンラインサービスについては、2025年に韓国のPersonal Information Protection Commissionが個人データ保護法への対応不備を理由に新規アプリダウンロードを一時停止させ、イタリア当局もデータ処理の説明を巡りサービスを制限した経緯がある。これはV4.1のモデル重みそのものに脆弱性があることを意味しないが、**DeepSeek公式APIに機密情報を送信するリスク評価では無視できないベンダー履歴**である。[28]

一方でV4.1 FlashはMITで重みが公開されているため、自社管理クラスタに完全self-hostし、外部APIへ機密promptを送らない構成も技術的に選択できる。したがってDeepSeekのプライバシー評価を「安全/危険」と一律に分類するより、**公式SaaS API利用とself-host deploymentを明確に分けるべき**である。ただしself-hostではmodel supply chain、weight integrity、runtime vulnerability、アクセス制御、監査ログ等を利用企業自身が負担する。 [20]

サイバーセキュリティのdual-use。 CyberGym 88.1、SEC-Bench Pro 62.8など、V4.1はセキュリティ関連コードタスクでも旧V4を改善している。これは脆弱性診断、防御コード、patch生成などに有用である一方、攻撃用途への転用能力も一般論として上がり得る。オープンウェイトモデルではクラウド事業者側の利用監視だけに依存できないため、企業内導入ではtool permissionsやnetwork accessをモデル能力とは別レイヤーで制御する必要がある。 [26]

規制。 EUではAI Act、GDPR、セクター別法規が重なり、GPAI、高リスク用途、透明性、著作権、データ処理などに異なる義務が生じる。AI Actの実施時期は2025〜2026年に段階的に進められる一方、後続の制度簡素化提案も存在するため、2026年9月時点では利用地域と用途別の最新法務確認が不可欠である。特に医療、採用、人事評価、信用判断、重要インフラなどは「汎用モデルを使っているだけ」では規制分析を免れない。 [41]

最大の透明性上の課題は、**モデルの性能仕様は驚くほど詳細なのに、訓練データの構成、V4.1固有の安全性評価、安全post-trainingデータ、安全ケース、red-team結果について同等レベルの開示が確認できないこと**である。DeepSeekは透明度センターとセキュリティ報告窓口を設置しているが、2026年9月10日の確認時点で同センターの主要モデル一覧にはV4.1の情報がまだ十分反映されていなかった。公開初日の更新ラグである可能性もあり、これ自体を不透明性の決定的証拠とみなすべきではない。 [42]

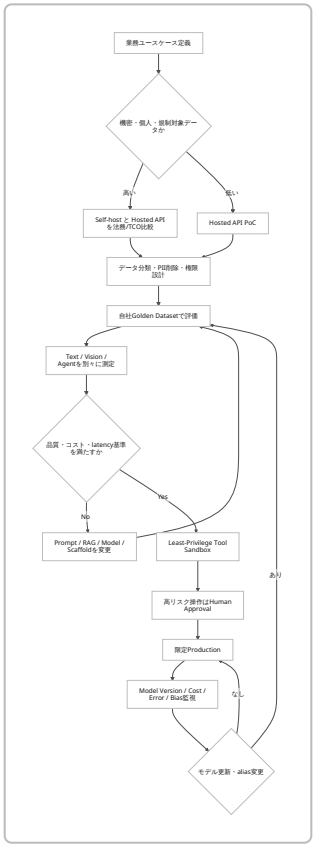
導入リスクと緩和策

企業導入では、単に「APIを既存モデルから `deepseek-flash` に差し替える」アプローチは避けるべきである。OpenAI/Anthropic互換性により技術的な切替は容易だが、モデル挙動、vision prompt injection、長文context、tool use、データ移転、モデルalias更新といった新しい攻撃面が同時に生じる。 [43]

リスク	根拠・影響	推奨緩和策
ハルシネーション/事実誤認	SimpleQAはV4 Proを大きく下回る。 [25]	RAG、DB照合、URL/文献引用必須化。高リスク回答はhuman approval。
画像/OCR誤認	DocVQAは高いがvendor benchmarkであり完全性を保証しない。画像も自動resizeされる。 [36]	OCRとの二重照合、数値・日付・口座番号等をdeterministic parserで検証。
画像・文書経由prompt injection	画像とtextを同じagent contextで処理しtool callsも可能。 [44]	文書内容を「命令」ではなくuntrusted dataとして分離。tool allowlist、least privilege、sandbox。
機密データ流出	過去のDeepSeek hosted serviceは海外当局のprivacy scrutinyを受けた。 [28]	API利用前にDPA、保存場所、retention、subprocessorを確認。高度機密用途はself-hostを比較検討。
モデル更新による挙動変化	legacy model namesがV4.1へ自動routeされ、V4 Proも9/14以降route予定。 [10]	aliasを信用せずmodel-version telemetryを保存。更新時は自動regression evalを実行。

リスク	根拠・影響	推奨緩和策
scaffold依存性	同じモデルでもDeepSWEが最大8.7pt変動。 20	自社prompt/tool/scaffoldを固定した評価セットで比較。公開leaderboardのみで調達しない。
巨大モデルの運用難度	552B級で、公式も大規模展開では2k GPUs＋storage cluster保有者への連絡を案内。 8	API対self-hostのTCO比較、quantization検証、fallback provider、容量計画。
ピーク料金・context肥大化	peak/off-peak価格制、最大1M context。 10	context pruning、summary、cache設計、batchはoff-peakへ。task単位cost ceilingを設定。
著作権・第三者権利	45T corpusの出所詳細が未公開。 11	商用出力を人間/類似性検査で確認。vendor indemnityや契約条件を別途審査。
医療・法務等の高リスク利用	V4.1固有の臨床/法務validationは未公開。	advisory用途に限定し、qualified professionalを最終決定者にする。用途別validation完了まで自動実行禁止。
サイバー/agent暴走	高いcyber・coding能力とtool useの組合せ。 26	network egress制限、read-only default、金銭・削除・deploy等は人間承認、詳細audit log。

推奨される導入フローは次のようになる。



特に重要なのは、**V4.1 Flashと他モデルを「同じpromptへの単発回答」で比較しないこと**である。Agent製品では「成功タスク1件当たり総コスト」「完了までのwall-clock time」「平均tool calls」「再試行数」

「human escalation率」を測るべきである。DeepSeek自身がV4.1の優位を「performance、cost、speed、total time」の組合せとして説明しているため、この評価方法の方がベンダー自身の設計思想にも合っている。 [45]

今後の展望と推奨アクション

短期、2026年9月〜2027年初頭。 最も可能性が高いのはコードエージェントとdocument intelligenceでの急速なPoC拡大である。OpenAI/Anthropic API互換性と低価格により、既存アプリケーションのshadow trafficをV4.1へ流して比較する障壁が低い。9月14日には旧V4 Proが実質V4.1 Flashへ置換され、DeepSeek自身も推論コミュニティによる新アーキテクチャ対応を支援するとしているため、vLLM/SGLang等を含むserving stackの成熟度が短期の焦点になる。 [46]

同時に、短期で最も必要なのは**独立検証**である。2026年8月のV4 ProについてはArtificial Analysisの独立評価があり、同社IndexでV4 Pro 53、旧V4 Flash 40という差が確認されていた。V4.1 Flashについても、Artificial Analysis等の第三者がintelligence、latency、tokens/sec、TTFT、price-performance、long-context reliabilityを再測定するまでは、企業は公式benchmarkを暫定値と扱うべきである。 [47]

中期、2027年前後。 DeepSeekが発表文で「より大きなパラメータモデルへ拡張可能」と明記し、V4.1 Proの将来投入も予告していることから、CED/CSA2を大型モデルへ拡張するシナリオが最も自然である。もしFlashの8B/16B active designと890-byte KV cacheを保ちながら能力上限を引き上げられるなら、従来の「frontier性能には高価な密集計算が必要」という経済性へさらに圧力がかかる。これは公式ロードマップの確約ではなく、発表された設計目標からの推論である。 [48]

中期には、競争指標そのものも変わる可能性がある。V4.1では552Bという総規模より、8B prefill、16B decode、890 bytes/token cacheという「実際に毎tokenで使う資源」が重要になっている。今後の企業調達では**parameter countではなく、task-success-per-dollar、task-success-per-joule、GB-per-million-context-tokens、agent completion time**がより有用なKPIになると考えられる。 [49]

長期。 もしopen-weightのマルチモーダルモデルが、closed frontierのagent性能へ低価格で継続的に接近するなら、基盤モデルAPI自体の利幅は縮小し、価値は「独自データ」「業界workflow」「tool integration」「監査・セキュリティ」「UX」へ移りやすい。DeepSeek自身が旧premiumモデルを新Flashで置換しようとしていることは、そのコモディティ化を自社製品内で先取りする動きとも読める。 [50]

企業向けには、現時点での推奨は**「全面移行」ではなく「評価系を先に作り、低リスク領域からdual-model運用する」**ことである。特にソフトウェア開発、社内文書検索、請求書・帳票処理、スクリーンショット解析ではV4.1を候補に入れる価値が高い。一方、外部公開情報の事実QA、高度数学、法務、医療ではV4 Proや他社frontier modelを含むrouter構成を維持し、自社evalの結果でmodel selectionする方が合理的である。根拠はV4.1の性能プロファイルがuniformではないことにある。 [32]

研究者向けには、優先すべきテーマはCED/CSA2の再現性、KV cache圧縮による実測latency/energy削減、1M contextでのneedle retrievalだけでない実用長文理解、vision-language grounding、scaffold sensitivity、日中英cross-lingual bias、安全性である。DeepSeekはDeepSWE再現手順まで公開しているため、まず公式結果の再現を行い、その後にvendor-independent scaffoldで評価することが望ましい。 [20]

政策立案者には、モデルの国籍や総parameter数だけを規制proxyにするより、**データがどこへ送られるか、open weightsかhostedか、外部toolを何の権限で操作するか、人間への影響がどの程度か**を中心にしたdeployment-based評価が適する。V4.1のように同一モデルを中国の公式APIでも、自社完全self-hostでも使える場合、両者のプライバシー・サイバーリスクは大きく異なるからである。訓練データsummary、安全評価、モデル更新履歴、重大incident報告の標準化も、ベンダー横断比較を可能にするため重要になる。 [51]

最終的には、V4.1 Flashの成功を決めるのは552Bという数字ではなく、**「従来の高価なfrontier modelと同程度の実務タスクを、どれだけ少ないactive compute・cache memory・ドル・秒で完了できるか」**である。公開初日の証拠では、DeepSeekはコード/agent領域でその命題をかなり強く示した。一方、一般知識、最高難度推論、長文理解、安全性、訓練データ透明性では結論を出すには早い。したがって現時点で最も厳密な評価は、**「frontier intelligenceを全面的に塗り替えたモデル」ではなく、「frontier級agent性能のコスト構造を大きく下げる可能性を示した、新しいメモリ効率重視アーキテクチャ」**である。 52

主要ソース

以下は本調査で特に重視した一次資料・主要報道である。URLは2026年9月10日の調査時点。

- **DeepSeek公式発表 — DeepSeek V4.1 Flash：更强、更快、更普惠**
  https://www.deepseek.com/news/deepseek-v4-1-flash/ 8

- **DeepSeek公式 Hugging Face — DeepSeek-V4.1-Flash model card / technical report content**
  https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash 53

- **DeepSeek-V4.1-Flash config.json — architecture configuration**
  https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash/blob/main/config.json 54

- **DeepSeek API — Models & Pricing**
  https://api-docs.deepseek.com/quick_start/pricing/ 10

- **DeepSeek API — Vision documentation**
  https://api-docs.deepseek.com/guides/vision/ 15

- **DeepSeek Transparency Center**
  https://www.deepseek.com/transparency/ 55

- **Reuters — China’s DeepSeek launches V4.1-Flash model, 2026-09-10**
  https://www.reuters.com/world/asia-pacific/chinas-deepseek-launches-v41-flash-model-2026-09-10/ 56

- **Reuters — DeepSeek launches V4 Pro at prices up to 14 times higher than V4 Flash, 2026-08-13**
  https://www.reuters.com/world/china/deepseek-releases-official-v4-pro-model-it-steps-up-expansion-2026-08-13/ 47

- **Reuters — DeepSeek-V4 and Huawei accelerator adaptation, 2026-04-24**
  https://www.reuters.com/world/china/deepseek-v4-chinese-ai-model-adapted-huawei-chips-2026-04-24/ 21

- **Anthropic — Claude Models & Pricing**
  https://platform.claude.com/docs/en/models/overview 29

- **Moonshot AI — Kimi K3 official documentation**
  https://platform.kimi.com/docs/guide/kimi-k3-quickstart 31

- **OpenAI Developers — current model lineup**
  https://developers.openai.com/ [57]

- **Reuters — South Korean regulator action concerning DeepSeek privacy, 2025-02-17**
  https://www.reuters.com/technology/south-koreas-data-protection-authority-suspends-local-service-deepseek-2025-02-17/ [58]

- **Huang et al. — Analysis of LLM Bias in DeepSeek-R1 vs. ChatGPT, arXiv, 2025**
  https://arxiv.org/abs/2506.01814 [59]

- **Ying et al. — Towards Understanding the Safety Boundaries of DeepSeek Models, arXiv, 2025**
  https://arxiv.org/abs/2503.15092 [60]

情報の確度に関する最終注記： V4.1 Flashは本レポート作成日である**2026年9月10日当日に公開されたモデル**である。アーキテクチャ、価格、公式benchmark、画像制限についてはDeepSeek一次資料を優先した。一方、実GPU上のtokens/sec、TTFT、必要VRAM、訓練GPU構成、45T-token corpusの詳細な出所・著作権処理、V4.1固有の包括的safety/red-team結果、医療・法務validation、マネージドfine-tuning条件は**未公開または本調査で公式情報を確認できなかった**。また、性能値の多くは現時点でDeepSeek自身による評価であり、公開初日のため第三者による包括的再現検証がまだ不足している。この点が、現段階でV4.1 Flashを評価する際の最大の不確実性である。 [61]

---

[1] [7] [8] [9] [22] [35] [39] [45] [46] [48] [50] [61] https://www.deepseek.com/news/deepseek-v4-1-flash/
https://www.deepseek.com/news/deepseek-v4-1-flash/

[2] [3] [5] [11] [14] [17] [19] [20] [23] [25] [26] [32] [36] [49] [51] [52] [53] https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash
https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash

[4] [10] [18] [24] [43] [44] https://api-docs.deepseek.com/quick_start/pricing/
https://api-docs.deepseek.com/quick_start/pricing/

[6] [56] https://www.reuters.com/world/asia-pacific/chinas-deepseek-launches-v41-flash-model-2026-09-10/
https://www.reuters.com/world/asia-pacific/chinas-deepseek-launches-v41-flash-model-2026-09-10/

[12] [13] [54] https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash/blob/main/config.json
https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash/blob/main/config.json

[15] [16] [37] https://api-docs.deepseek.com/guides/vision/
https://api-docs.deepseek.com/guides/vision/

[21] https://www.reuters.com/world/china/deepseek-v4-chinese-ai-model-adapted-huawei-chips-2026-04-24/
https://www.reuters.com/world/china/deepseek-v4-chinese-ai-model-adapted-huawei-chips-2026-04-24/

[27] [29] [34] https://platform.claude.com/docs/en/models/overview
https://platform.claude.com/docs/en/models/overview

[28] [58] https://www.reuters.com/technology/south-koreas-data-protection-authority-suspends-local-service-deepseek-2025-02-17/
https://www.reuters.com/technology/south-koreas-data-protection-authority-suspends-local-service-deepseek-2025-02-17/

30 https://docs.anthropic.com/
https://docs.anthropic.com/

31 https://platform.kimi.com/docs/guide/kimi-k3-quickstart
https://platform.kimi.com/docs/guide/kimi-k3-quickstart

33 57 https://developers.openai.com/
https://developers.openai.com/

38 47 https://www.reuters.com/world/china/deepseek-releases-official-v4-pro-model-it-steps-up-expansion-2026-08-13/
https://www.reuters.com/world/china/deepseek-releases-official-v4-pro-model-it-steps-up-expansion-2026-08-13/

40 59 https://arxiv.org/abs/2506.01814
https://arxiv.org/abs/2506.01814

41 https://www.reuters.com/world/europe/artificial-intelligence-rules-go-ahead-no-pause-eu-commission-says-2025-07-04/
https://www.reuters.com/world/europe/artificial-intelligence-rules-go-ahead-no-pause-eu-commission-says-2025-07-04/

42 55 https://www.deepseek.com/transparency/
https://www.deepseek.com/transparency/

60 https://arxiv.org/abs/2503.15092
https://arxiv.org/abs/2503.15092