DeepSeek-V4.1-Flash 徹底調査レポート

― 事実確認、仕様・性能の検証と今後に与える影響の考察 ―

2026 年 9 月 10 日

Claude Fable 5.1

要旨

本レポートは、DeepSeek が 2026 年 9 月 10 日に発表した新型モデル「DeepSeek-V4.1-Flash」について、発表の事実確認、仕様・性能、系譜と競合比較、各方面の反響を整理し、今後の影響を考察するものである。確認済みの事実、報道に基づく情報、筆者の分析・推論を明示的に区別して記述する。要点は次の 3 点である。

- **前提は概ね事実である。**DeepSeek は 2026 年 9 月 10 日（北京時間）、ネイティブ・マルチモーダル（画像理解）対応の新型 MoE モデル「DeepSeek-V4.1-Flash」を正式発表した。公式 API チェンジログ[1)]、Hugging Face モデルカード（MIT ライセンス、技術レポート付き）[2)]、ITmedia[3)]、Bloomberg[4)]で確認できる。ただし、9 月 8 日に先行した 48 時間限定ベータと混同されやすい点に注意を要する。
- **正体は「Flash」の名を持つ準フラッグシップである。**総バックボーン 552B・アクティブ 8B（prefill）／16B（decode）の新アーキテクチャ「Causal Encoder-Decoder（CED）」を採用し、KV キャッシュを 1 トークンあたり 890 バイト（V4-Flash の約 1/4）まで圧縮した[2)]。公表ベンチマークでは自社旗艦 V4-Pro（1.6T）や GPT-5.6 Sol、Claude Opus 5 をエージェント／コーディング系の一部で上回るとするが、Humanity's Last Exam（HLE）など難関推論では米フロンティアに劣後する[2)]。V4 Pro は 9 月 14 日以降 V4.1-Flash へ順次ルーティングされ、実質引退となる[1)]。
- **影響は「軽量マルチモーダルの低価格コモディティ化」の加速である。**同時に API 価格をキャッシュヒット入力で約 60%引き下げ[1,3)]、香港上場の MiniMax・Z.ai 株は 8%超下落した[4)]。日本の IP・文書処理業務では低コストな画像・PDF 読解の選択肢になり得るが、①中国サーバー・国家情報法によるデータガバナンスリスク、②標準的な文書ベンチマーク

（OCRBench／ChartQA 等）の非開示、③API 画像入力が実質 800×800 相当・384 トークン上限、の 3 点で機密・高精度用途には慎重な検証が必須である。

1. 前提の真偽検証

1.1 結論

【確認済み事実】「2026 年 9 月 10 日に DeepSeek が『DeepSeek-V4.1-Flash』を発表した」「画像を読める（画像理解・マルチモーダル対応）」という前提は、いずれも一次ソースで確認できる。

- **正式名称**：「DeepSeek-V4.1-Flash」（ハイフン表記が公式）。API 呼び出し名は新たに「deepseek-flash」となった [1,2]。
- **発表日**：DeepSeek 公式 API チェンジログおよび Hugging Face で 9 月 10 日に公開された [1,2]。ITmedia は「9 月 10 日（現地時間）」と報道 [3]。Bloomberg（2026 年 9 月 10 日、Saritha Rai 記者）は同日（木曜日）に発表されたと報じた [4]。新 Flash 系価格は北京時間 9 月 10 日 12:00（UTC 04:00）に発効した [1]。
- **発表形態**：①オープンウェイト公開（Hugging Face、MIT ライセンス、技術レポート PDF 付き）[2]、②公式 API 提供（deepseek-flash）[1]、③API 価格改定 [1]、④V4-Pro→V4.1-Flash ルーティング告知 [1]——の複合である。公式チェンジログは、2026 年 9 月 14 日北京時間 12:00 以降、V4.1 Pro のリリースまでの間、deepseek-v4-pro への全リクエストを V4.1 Flash にルーティングし、V4.1 Flash 価格で課金すると明記している [1]。

1.2 経緯（混同に注意）

【確認済み事実／報道ベース】9 月 8 日午後（北京時間約 15:00）、DeepSeek は 48 時間で失効する内部ベータ「deepseek-v4.1-flash-expires-on-0910」を公式ユーザーコミュニティで静かに告知した（1 アカウント 20 同時実行に制限）[5,6]。DEV Community も、9 月 8 日にベータが静かに開始され、開発者には 48 時間のテスト期間しか与えられなかったと記録している [7]。9 月 9 日に TechNode 等が「限定ベータ」を報道し [5]、公式が正式リリースを予告、9 月 10 日に正式版・オープンウェイト・技術レポートが公開された [1,2]。したがって「9 月 8 日ベータ／9 月 10 日正式発表」の 2 段階を 1 つに丸めると誤りが生じる。

1.3 既存モデルとの取り違えに関する注意

V4-Pro-0813（8 月 13 日 GA）、V4-Flash-0731（7 月 31 日）、V4-Flash-Vision-Exp（8 月 21 日公開の実験的マルチモーダル。視覚エンコーダを「外付け」した構成）[8,9,10]は別物であり、V4.1-Flash はこれらを置換・統合する[1,2]。一部の日本語ブログ[11]は「公式 API に V4.1 というモデル名はない」と 9 月 9 日時点の情報で断じているが、これはベータ段階の観測であり、9 月 10 日の正式版で覆っている。

2. モデルの仕様・性能

2.1 アーキテクチャ

【確認済み事実】以下は Hugging Face 公式モデルカード[2]および技術レポートの要約[12,13]に基づく。

- MoE モデル。バックボーン 552B パラメータに加え、Engram 条件付きメモリ 196B パラメータを持つ。1 トークンあたり prefill で 8B、decode で 16B をアクティブ化する。1 shared expert＋384 routed experts のうち 6 routed experts をアクティブ化[2,12]。
- **Causal Encoder-Decoder（CED）**：40 層を「20 層 causal encoder＋20 層 decoder」に分割し、デコーダのグローバル KV を最終エンコーダ隠れ状態から射影する（YOCO 着想）。プロンプトトークンはエンコーダで止まり、prefill 計算をほぼ半減させる[2,12,13]。
- **Compressed Sparse Attention 2（CSA2）**：各層を Full／Reindex／Reuse の 3 モードに静的割当てし、主 KV・indexer K・Top-K インデックスを層間で共有する。FP4 主 KV キャッシュ（NVFP4 系、E2M1＋16 チャネルあたり E4M3 スケール）と併せ、グローバル KV を 1 トークンあたり 890 バイトに圧縮（V4-Flash の約 1/4、V1 の約 1/437）。HBM 要件を約 1/4、SSD 要件を約 1/8 に削減する（SWA Bounded Replay 併用）[2,12,13]。
- **コンテキスト長**：100 万トークン。推奨 max_tokens は 256K 以上[2]。
- その他：Single-Pass mHC（Mega-mHC カーネル）、DSpark 投機的デコード、head-wise Muon オプティマイザ[2,12]。
- **視覚**：DeepSeek-ViT（スクラッチ学習、2D-RoPE、3×3 pixel-unshuffle ダウンサンプリング）＋2 層 MLP プロジェクタ。言語モデル事前学習の最初からテキストと同時に処理する真のネイティブ・マルチモーダルであり、V4-Flash-Vision-Exp の「外付け」構成とは異なる[2,12]。準一次ソースの技術レポート要約では 32 層 ViT・約 1344×1344 対応とされるが、一次確認は未了である[13]。動画対応・画像生成は非対応（テキスト自己回帰生成のみ）[2,14]。

2.2 学習手法

- スクラッチから 45T トークンのマルチモーダルコーパスで事前学習（テキスト：マルチモーダル≒7：1）。スパースアテンションは 64K 系列長で学習し、34T トークン時点で 1M へ拡張 [2,12]。
- 事後学習は SFT→RL→on-policy distillation（OPD）。アルゴリズム自体の変更はなく、検証可能なエージェントタスクの大規模自動合成と 40 超の教師モデルからの蒸留が中核 [2,12]。
- 連続可変の「reasoning effort」（1〜100）で推論コストと精度をトレードオフ可能 [2]。

2.3 ベンチマーク

【確認済み事実】以下は DeepSeek 公式値（max reasoning effort 時）である [2]。

項目	V4.1-Flash	比較対象（DeepSeek 公表）
Terminal-Bench 2.1	90.6	V4-Pro 87.9／GPT-5.6 Sol 88.8／Kimi K3 88.3／Opus-5.0 89.1
DeepSWE v1.1	74.2	—
AutomationBench	54.8	—
CyberGym	88.1	—
Agent's Last Exam	31.8	—
Codeforces rating	3471	—
Humanity's Last Exam（HLE）	36.8	Opus-5.0 56.3／GPT-5.6 Sol 44.5
GPQA Diamond	90.9	Opus-5.0 93.4／GPT-5.6 Sol 94.1
MMMU-Pro（base）	56.5	—
CVBench（base）	77.9	—
DocVQA（LLM-Judge, base）	95.6	—（ANLS 値ではないため他社値と直接比較不可）
RefCOCO-avg（base）	86.0	—
Chartography（instruct）	78.9	—
BabyVision（instruct）	89.6	—
ZeroBench-main（instruct）	49.0	—

出典：Hugging Face モデルカード [2]。「—」は同一資料に比較値の記載がないことを示す。

Bloomberg は、スリム化したモデルでありながら Moonshot の Kimi K3 などを上回ったと主張しつつ、大幅な割引価格を提示していると要約している [4]。留意点として、OCRBench・

ChartQA・標準 MMMU・MathVista・AI2D・InfoVQA・TextVQA は非開示であり、DocVQA は通常の ANLS ではなく LLM-Judge 評価のため他社公表値と直接比較できない [2)]。第三者ベンチマーク（Artificial Analysis 等）は執筆時点で V4.1-Flash を未測定である。

2.4 価格・提供条件

【確認済み事実】新 Flash 料金（100 万トークンあたり、オフピーク、UTC 04:00 発効）は以下のとおり [1,3,15,16)]。

区分	旧 V4 Flash	V4.1-Flash（新）
キャッシュヒット入力	$0.007	$0.003
キャッシュミス入力	$0.22	$0.15（≒1 円）
出力	$0.66	$0.60（≒4 円）

出典：DeepSeek API Docs Change Log[1)]、ITmedia[3)]、Costgoat[15)]、Intelligent Living[16)]。

- ピーク時（UTC 01:00〜04:00、06:00〜10:00）は 2 倍 [1,15)]。キャッシュヒット入力で約 60%、出力で約 1 割の引き下げとなる。8 月 13 日の値上げから一転しての値下げである点を ITmedia が指摘している [3)]。
- V4 Pro ユーザーは 9 月 14 日以降、コード変更なしで V4.1-Flash（Flash 価格）に自動移行する。V4.1-Pro 登場までの暫定措置である [1,16)]。
- **ライセンス**：MIT ライセンスでオープンウェイト公開（vLLM／SGLang／Transformers 対応）[2)]。ベータ時の同時実行上限は 1 アカウント 20 だったが [5)]、正式版は「高同時実行」を掲げる（具体値は執筆時点で未公表）[17)]。
- **画像 API 制約**：入力形式は JPEG／PNG／GIF／WebP（PDF 直接入力は不可、クライアント側でのラスタライズが必要）。画像は実質 800×800 相当・1 枚最大 384 トークンにリサイズされる。detail=low は 512×512。1 リクエスト最大 600 枚、最大辺 8192px（15 枚以上の場合は 4096px）[14)]。

3. 系譜と競合比較

3.1 DeepSeek 自社系譜

V3（671B／37B）→V3.2→V4 Preview（4 月 24 日：Pro 1.6T／49B、Flash 284B／13B）→V4-Flash-0731→V4-Pro-0813 GA→V4-Flash-Vision-Exp（8 月 21 日）→V4.1-Flash（9 月 10 日）という系譜である [10,18)]。V4.1-Flash は「新アーキテクチャ・ファミリーの最小モデル」と位置づけ

られ、より大型の V4.1-Pro へのスケールが前提となる[1,2]。V4 系が CSA＋HCA のハイブリッドであったのに対し、V4.1 は CSA2 に一本化し CED を導入した点が構造的な断絶である[12,18]。

「Flash」の名だが実は大型化しており、284B から 552B へほぼ倍増した。Hacker News では「もはや Flash ではない」「オープンウェイトの新王者かもしれない」との反応が見られた[19]。

3.2 競合（2026 年 8〜9 月の Flash 級ラッシュ）

【報道ベース】Alibaba Qwen3.8 Flash（8 月 26 日）、Z.ai GLM-5.3／GLM-5.3 Flash（8 月 14 日・8 月 26 日）、Moonshot Kimi K3、Google Gemini 3.8 Flash（9 月 2 日）、Meta Muse Spark、OpenAI GPT-6 Astra（9 月 4 日）、Anthropic Claude Fable 5.1（9 月 1 日）が相次いで登場した。価格面では、9 月時点の比較で GPT-6 Astra と Claude Fable 5.1 が入力$10／出力$50、Gemini 3.8 Flash が$0.75／$3.75 とされ[18]、V4.1-Flash はこれらを桁違いに下回る。

3.3 米中格差論への示唆

【分析】V4.1-Flash は、エージェント／コーディング／長文の実務系ベンチマークではフロンティア級に肉薄する一方、HLE や ProgramBench 等の難関推論では米トップに明確に劣後する[2]。「中国は用途を絞れば追いついたが、最難関推論のフロンティアではなお差がある」という混在した状況を示唆する。SCMP は、Artificial Analysis のベンチマークチャートを引き、同等製品でより高い価格を付けるか、同価格でより低性能であるなら参入する意味がない、いわゆる「DeepSeek death zone」が出現したと報じている（独立評価では複雑な実世界ワークロードの実行コストが約 US$0.03 とされる）[20]。

4. 反響・評価

4.1 市場

【報道ベース】Bloomberg によれば、MiniMax Group と Z.ai の株価は香港市場で 8%超急落し、Alibaba も 2%超下落した[4]。Bloomberg は、DeepSeek の低価格戦略を「十分に良い性能を超低価格で」提供し AI 採用の次段階を狙う動きと評価している[4,21]。

4.2 開発者コミュニティ

発表前ベータには開発者が殺到した[6,19]。実測で出力 300〜500 トークン/秒、ピーク 400 超の報告があるが、これはコミュニティ実測でありベンダー公表値ではない[17,19]。速度は高く評価される一方、「トークンが速く出る分、残高の減りも速い」という声もある[19]。第三者ベンチマーク

（Artificial Analysis 等）は V4.1-Flash を未測定である。

4.3 チップ／インフラ・資金調達

【報道ベース】DeepSeek は訓練を NVIDIA、推論を Huawei Ascend で行う分業を継続している[22)]。TechNode（IBTimes Singapore 経由）の報道では、内モンゴルのデータセンターに少なくとも 16 万個の Huawei Ascend 950DT アクセラレータを、主にモデル訓練ではなく推論用途として配備する計画とされる（額面約 26 億ドル、GW 級）[22,23,24)]。並行して上海 STAR Market での IPO 準備が報じられており、SCMP によれば CITIC 証券を含む 4 社を引受幹事に起用し、2026 年内の手続き開始を目指す[20)]。資金調達については、Reuters 報道（Business Standard 経由）で評価額 5,000 億元（約 750 億ドル）のラウンドの最中とされ、6 月ラウンドではポストマネー500 億ドル超・約 74 億ドルを調達（創業者梁文峰が 200 億元、Tencent が 100 億元、CATL が 50 億元を拠出）と報じられている[25)]。

4.4 日本の反応

ITmedia が「『DeepSeek-V4.1-Flash』発表 値上げから一転『より低コストでフラッグシップ超え』うたう」と報道した[3)]。GIGAZINE は前世代の V4-Flash-Vision-Exp を詳報している[8)]。複数の日本語技術ブログが速度と価格を評価している[11)]。

4.5 安全性・ガバナンス

【報道ベース】中国の消費者向けサービスはデータを中国国内サーバーに保存し、2017 年国家情報法・データ安全法・サイバーセキュリティ法の適用下にある[26)]。米国・イタリア・韓国・オーストラリア・台湾・インドなどが政府端末等での利用を制限している[26,27)]。ただし、オープンウェイトを自社インフラで自己ホストすればプロンプトを DeepSeek に送らずに済むため、リスク構造は変わる[26)]。

5. 今後に与える影響の考察

本章は筆者の分析・推論であり、事実関係（第 1〜4 章）とは区別して読まれたい。

5.1 AI 業界

ネイティブ軽量マルチモーダルのオープンウェイト化と価格破壊が加速する。競争軸は「モデル選定」から「推論インフラ戦略」へ移行しつつある（NVIDIA 技術ブログも同旨を指摘[28)]）。CED／CSA2／FP4 KV による KV キャッシュ削減は、長文・エージェント（入力偏重）ワークロ

ードの推論経済を大きく変え得る。

5.2 米中 AI 競争・地政学

訓練は NVIDIA、推論は国産という分業と自社推論チップ開発、Ascend の大量調達は、米輸出規制下でも推論を国産で回す実証である。オープンウェイト＋低価格は、米クローズドモデルへの価格圧力を一段と強める。

5.3 日本企業・知財実務への示唆

可能性：1M コンテキスト＋ネイティブ画像理解＋低価格は、明細書 PDF・図面・チャート・製品写真の一次スクリーニング、先行技術調査の下読み、IP ランドスケープの大量文書処理に魅力的である。MIT オープンウェイトのため、自己ホストにより機密を外部送信せずに運用できる。

リスク：①標準文書ベンチマーク（OCRBench／DocVQA-ANLS／ChartQA 等）が非開示であり、特許図面の微細な符号・化学構造式・小さな図表ラベルの読解精度が未検証である。②API 経由では画像が実質 800×800 相当・384 トークンに圧縮され [14]、高精細な図面・構造式は情報欠落の恐れがある（ROI を切り出して送る運用が必要）。③ホスト版 API は中国法上のデータリスクがあり、侵害鑑定・出願前情報など機密性の高い業務には自己ホストか非中国系ホスティングが前提となる。④検閲・バイアスの懸念がある [26]。

推奨スタンス：非機密の大量下読み・分類には有力候補だが、権利化・鑑定に直結する判断は人間および専用文書 AI（専用 OCR／文書抽出モデル）との二重化が必要である。

5.4 今後の見通し

DeepSeek は公式に V4.1-Pro を予告している（V4 Pro ルーティングの暫定措置の説明内）[1]。V5 は未発表・未確認であり、噂の多くは旧エイリアス廃止告知の誤読とみられる [29,30,31]。次期は V4.1 系のスケールアップが本命と見られる。

6. 推奨アクション

1) **用途を「非機密の大量下読み」に限定して PoC**：自己ホスト（vLLM／SGLang、MIT ライセンス）で V4.1-Flash を構築し、先行技術調査の一次スクリーニングや IP ランドスケープの文書分類に適用する。プロンプトを外部送信しない構成とする。判断基準：自社の代表的な特許図面・化学構造式・チャート 100 件で、符号読み取り・要素抽出の精度を人手正解と照合し、実務許容水準（例：抽出再現率 90%以上）を満たすか。

2) **画像前処理の標準化**：API 利用時は 800×800・384 トークン上限を前提に、図面・構造式は ROI（関心領域）を切り出して個別送信し、ページ全体の一括投入を避ける。微細ラベルの誤読率が許容を超える場合は、専用 OCR／文書抽出モデルとの二段構えに切り替える。

3) **機密・権利化直結業務では中国ホスト版 API を使わない**：侵害鑑定・出願前情報・営業秘密を含む処理は、自己ホストまたは非中国系クラウドでのオープンウェイト運用に限定し、法務・情報セキュリティ部門によるデータ越境レビューを必須化する。

4) **第三者ベンチと技術レポートの精読を待って本番採用を判断**：Artificial Analysis 等の独立評価と DeepSeek 技術レポート PDF の文書系スコアが出そろってから、本番トラフィック移行の可否を決定する。

5) **V4.1-Pro／価格改定の監視**：V4.1-Pro のリリースと更なる価格改定はコスト・性能の前提を変えるため、四半期ごとにモデル・価格を再評価する。

7. 留意事項

- 本レポートは発表当日（2026 年 9 月 10 日）時点の情報に基づく。第三者による独立ベンチマーク（Artificial Analysis 等）は V4.1-Flash を未測定であり、性能・速度の主張の多くはベンダー公表値かコミュニティ実測で、独立検証は未了である。
- 技術レポート PDF 本文は本調査でアクセス確認できなかった（Hugging Face 上のリンクは存在するが取得時に権限エラー）。CED／CSA2／890 バイト等の数値はモデルカード [2] と技術レポート要約 [12,13] に基づく。視覚エンコーダの「32 層・約 1344×1344 対応」は準一次 [13] の要約であり、一次確認は未了である。
- 標準的な文書／視覚ベンチマーク（OCRBench・ChartQA・MathVista 等）は DeepSeek 公式資料に非開示であり、特許図面・化学構造式の実務精度は本レポート時点で客観的に評価できない。
- 価格・同時実行上限・ルーティング日程は DeepSeek が短期間で変更する実績があり（8 月の値上げ→9 月の値下げ）、記載値は執筆時点のものである。
- 市場反応（株価下落率等）は Bloomberg 報道 [4] に基づく発表当日の速報値である。
- IPO 評価額（約 5,000 億元／750 億ドル）・Ascend 調達規模（16 万個・約 26 億ドル）・訓練／推論分業は報道ベース [20,22,23,24,25] であり、DeepSeek の監査済み開示ではない。

- 参考文献のうち、URL を本調査で直接確認できなかったもの（SCMP、DEV Community、IBTimes Singapore、Reuters／Business Standard、NVIDIA 技術ブログ）には「URL 未確認」と付記した。

参考文献

1) DeepSeek, “Change Log”, DeepSeek API Docs, 2026 年 9 月 10 日更新. https://api-docs.deepseek.com/updates/

2) DeepSeek-AI, “deepseek-ai/DeepSeek-V4.1-Flash”（モデルカード・技術レポート）, Hugging Face, 2026 年 9 月 10 日. https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash

3) ITmedia NEWS, 「『DeepSeek-V4.1-Flash』発表 値上げから一転『より低コストでフラッグシップ超え』うたう」, 2026 年 9 月 10 日. https://www.itmedia.co.jp/aiplus/article/2609/10/2000001374/

4) Saritha Rai, “DeepSeek’s New Low-Cost Model Deals a Fresh Blow to OpenAI, Z.ai”, Bloomberg, 2026 年 9 月 10 日. https://www.bloomberg.com/news/articles/2026-09-10/deepseek-s-new-low-cost-model-deals-a-fresh-blow-to-openai-z-ai

5) TechNode, “DeepSeek begins limited-time beta of V4.1 Flash multimodal model”, 2026 年 9 月 9 日. https://technode.com/2026/09/09/deepseek-v4-1-flash-multimodal-limited-beta/

6) Hacker News, “DeepSeek v4.1 Flash is now available for internal beta testing”, 2026 年 9 月. https://news.ycombinator.com/item?id=49607094

7) DEV Community, DeepSeek V4.1 Flash の 48 時間限定ベータに関する記事, 2026 年 9 月. ※URL 未確認

8) GIGAZINE, 「DeepSeek が画像認識に対応した『DeepSeek-V4-Flash-Vision-Exp』をオープンモデルとして公開、画像認識を含むタスクで Claude Opus 4.8 と同等の性能」, 2026 年 9 月 2 日. https://gigazine.net/news/20260902-deepseek-v4-flash-vision-exp/

9) Developers Digest, “DeepSeek V4 Flash Vision Exp: Experimental Vision, Limits, and How to Run It in OpenCode”, 2026 年. https://www.developersdigest.tech/blog/deepseek-v4-flash-vision-exp-opencode-guide

10) AI Release Tracker, “DeepSeek Models — 22 Releases & Benchmarks”, 2026 年. https://aireleasetracker.com/company/deepseek

11) ZIDOOKA!, 「DeepSeek V4.1 Flash は本当に賢くなったのか——公式情報と使用感で確かめる」, 2026 年 9 月 9 日. https://www.zidooka.com/archives/5031

12) MarkTechPost, “DeepSeek AI Released DeepSeek-V4.1-Flash with 1M Context, FP4 KV Cache, and Cross-Layer Attention Reuse”, 2026 年 9 月 10 日. https://www.marktechpost.com/2026/09/10/deepseek-ai-released-deepseek-v4-1-flash-with-1m-context-fp4-kv-cache-and-cross-layer-attention-reuse/

13) The Neuron, “DeepSeek V4.1 Flash explained: how it cuts AI memory 8x”, 2026 年 9 月. https://www.theneuron.ai/explainer-articles/deepseek-v41-flash-explained-how-it-cuts-ai-memory-8x/

14) DeepSeek, “Vision”, DeepSeek API Docs, 2026 年. https://api-docs.deepseek.com/guides/vision/

15) Costgoat, “DeepSeek API Pricing Calculator & Cost Guide (Sep 2026)”, 2026 年 9 月. https://costgoat.com/pricing/deepseek-api

16) Intelligent Living, “DeepSeek V4.1 Flash Released: Benchmarks, Pricing, and Pro Retirement”, 2026 年 9 月. https://www.intelligentliving.co/deepseek-v41-flash-pricing-release/

17) Tech Insider, “How to Use DeepSeek V4.1 Flash API: 12 Steps [2026]”, 2026 年 9 月. https://tech-insider.org/how-to-use-deepseek-v4-1-flash-api-2026/

18) OrcaRouter, “DeepSeek V4.1 Flash vs V4 Flash: Should You Switch?”, 2026 年 9 月. https://www.orcarouter.ai/blog/deepseek-v4-1-flash-vs-deepseek-v4-flash

19) Hacker News, “DeepSeek v4.1 Flash”, 2026 年 9 月 10 日. https://news.ycombinator.com/item?id=49639090

20) South China Morning Post, DeepSeek V4.1 Flash（Causal Encoder-Decoder、Terminal-Bench、「DeepSeek death zone」）および STAR Market IPO 準備に関する報道, 2026 年 9 月. ※URL 未確認

21) Deccan Chronicle, “China’s Deepseek Unveils V4.1-Flash Model”, 2026 年 9 月 10 日. https://www.deccanchronicle.com/technology/chinas-deepseek-unveils-v41-flash-model-1986393

22) XenoSpectrum, “DeepSeek Plans 160,000 Huawei AI Chips for Inference, Still Trains on NVIDIA”, 2026 年 9 月. https://xenospectrum.com/en/deepseek-huawei-ascend-950dt-160k-order/

23) Big Hat Group, “China AI Weekly: DeepSeek’s 160K Huawei Chip Megaproject”, 2026 年 9 月 6 日. https://www.bighatgroup.com/blog/china-ai-weekly-2026-09-06/

24) TechNode（IBTimes Singapore 経由）, DeepSeek の Huawei Ascend 950DT 配備計画に関する報道, 2026 年 9 月. ※URL 未確認

25) Reuters（Business Standard 経由）, DeepSeek の資金調達・評価額に関する報道, 2026 年. ※URL 未確認

26) Tech Jacks Solutions, “Security and Privacy Risks of DeepSeek: Censorship, Jailbreaks, and Bans (2026)”, 2026 年. https://techjacksolutions.com/ai-tools/deepseek/deepseek-security-and-privacy-risks/

27) Aitechtonic, “DeepSeek AI Banned Countries 2026: Complete List & Reasons”, 2026 年. https://aitechtonic.com/deepseek-ai-banned-countries/

28) NVIDIA Technical Blog, 推論インフラ戦略（モデル選定から推論インフラへの競争軸の移行）に関する記事, 2026 年. ※URL 未確認

29) DeepSeek AI Guide, “DeepSeek Roadmap 2026: V4 Preview, V5 Outlook & Plans”, 2026 年. https://deepseekai.guide/guides/deepseek-roadmap/

30) Chat-deep, “DeepSeek V5 Release Date and 2026 Roadmap: Official Status”, 2026 年. https://chat-deep.ai/guide/deepseek-roadmap-rumors/

31) Thomas Wiegold, “When Is DeepSeek V5 Coming Out? The Honest 2026 Answer”, 2026 年. https://thomas-wiegold.com/blog/when-is-deepseek-v5-coming-out/

※「URL 未確認」の文献は、調査過程で参照されたが本調査において URL を直接取得・確認できなかったものである。引用に際しては原典の確認を推奨する。