

DeepSeek-V4.1-Flash-1017 調査報告書

性能・評判・知財業務への影響

2026 年 10 月 18 日時点

Claude Opus 5.5

要旨

- **前提**：名称と公開日（2026 年 10 月 17 日）は、DeepSeek 公式の一次ソースで確認できた^[1-3]。9 月 10 日公開の V4.1-Flash の「マイナーアップデート」（重みの更新版）で、API 名（deepseek-flash）と料金は変わらない^[1,2]。
- **性能**：公式表の全 11 項目で前版を 2.0～11.1 ポイント上回る（公式表から算出）^[2]。ただし数値はすべて自社測定で、1017 版の独立評価は 10 月 18 日時点で未公表である^[4]。比較相手の Claude Opus 5 には 4 勝 7 敗で、知識推論（HLE）では 13.7 ポイントの差がある^[2]。
- **評判**：安さ・速さ・エージェント性能の伸びが好評な一方^[4,5]、日本語の不自然さ、細かい指示への追従の弱さ、ハルシネーション、中国リスクへの批判がある^[6-9]。
- **知財業務**：公開特許文献のスクリーニング、中国語公報の一次読解、IP ランドスケープの下処理では費用対効果が高い（分析）。公式アプリ/API はデータを中国で保存し中国法に準拠するため^[10,11]、未公開発明・契約書・社内情報は入力せず、MIT ライセンスの重みを社内または承認済みホストで動かして検証するのが現実的である（分析）。

本書の表記と確認方法

- **【事実】** 一次ソース（DeepSeek 公式文書、公的機関の発表等）で確認した事項。
- **【二次情報】** 第三者の記事・集計・SNS 投稿に依拠し、一次ソースでは未確認の事項。
- **【分析】** 上記を踏まえた筆者の分析・推測。

本書作成時（2026 年 10 月 18 日）に、Hugging Face のモデルカード、API Change Log、Models & Pricing の記載を改めて確認した。上付きの番号は文末の参考文献に対応する。URL を確認できなかった資料はその旨を付記した。

1. 前提の検証

1.1 名称・公開日

【事実】 名称・公開日とも DeepSeek 公式の一次ソースで確認できた。

表 1 一次ソースでの確認結果

確認事項	結果	出典
正式名称	「DeepSeek-V4.1-Flash-1017」で正しい	[1,2]
公開日	2026 年 10 月 17 日。Web・アプリ・API が同日に新版へ切り替わった	[1]
オープンウェイト	Hugging Face 公式（deepseek-ai）で公開。MIT ライセンス	[2]
公式 SNS	公式 X で告知。性能とコストの「パレート改善」をうたう	[3]
公式ニュース記事	標題「DeepSeek-V4.1-Flash-1017: Wish-Driven Programming Made Real」。本文は取得時に別ページへ転送され、検索結果の抜粋でのみ確認	[12]
API 上の扱い	deepseek-flash で 1017 を提供。旧名 deepseek-v4-flash/deepseek-v4-flash-vision-exp も 1017 で処理され、Flash 料金で課金される	[13]

注：WeChat 公式アカウントの投稿は未確認。

1.2 9 月 10 日版「DeepSeek-V4.1-Flash」との関係

【事実】モデルカードは 1017 を V4.1-Flash の「minor update」と位置づけ、推論・コーディング・エージェント・視覚の各ベンチマークの向上と、タスク実行時間・出力長の短縮をうたう。引用される技術報告も 9 月版と同じ「DeepSeek-V4.1-Flash: Pushing the Limits of KV Cache Compression」である^[2]。API の呼び出し方法と料金は変わらない。DeepSeek Harness、Tencent CodeBuddy、WorkBuddy に最適化されており、Harness は 0.2.1-rc.1 以降が推奨されている^[1]。

【事実】公式 X によれば、自己フィードバックを用いた強化学習により、ゲーム制作から 3D モデリングまでのプロジェクト開発を一貫してこなす能力を高めた^[3]。【二次情報】The Decoder はこの手法を「RLSF」と呼び、モデル自身の確信度や自己生成した批評を報酬信号として使う方式と解説している^[4]。

【二次情報】アーキテクチャの変更は発表されていないが、重みは別物で、条件付きメモリ「Engram」のテーブルも 1017 専用である。量子化版の配布者は、旧版の Engram を流用するとエラーを出さないまま不整合なモデルが動くと警告している^[14]。

【分析】1017 は別モデルではなく、同一アーキテクチャの更新チェックポイントである。「画像対応の新型モデル」という位置づけは 9 月版のもので、1017 はその改良版にあたる。

2. 性能

2.1 アーキテクチャ・仕様

表 2 主な仕様

項目	内容
方式	マルチモーダル MoE。「Causal Encoder-Decoder」構造（40 層 = エンコーダ 20 層 + デコーダ 20 層） ^[15,16]

項目	内容
パラメータ	本体 552B。アクティブは入力処理（prefill）時 8B、出力生成（decode）時 16B。各 MoE 層に共有エキスパート 1 + ルーティング 384、トークンあたり 6 エクスパートを起動 ^[15-17]
追加要素	条件付きメモリ Engram 196B（本体とは別枠）、投機デコードモジュール「DSpark」等。Hugging Face 上の重み総数は 763B ^[2,15]
コンテキスト	100 万トークン（API 最大出力 384K） ^[13]
入出力	テキスト + 画像を入力、テキストを出力。動画・音声・PDF のネイティブ入力は非対応 ^[13,16,18]
学習	45T トークンのマルチモーダルコーパスで事前学習 ^[15]
KV キャッシュ	1 トークンあたり 890 バイトで、V4-Flash の約 1/4（HBM 上）。永続 KV は約 1/8 ^[15]
ライセンス	MIT（商用利用・改変・再配布が可能） ^[2]
重みの形式	FP8 中心、リポジトリは約 510GB。Jinja チャットテンプレートは同梱せず、Python 参照実装と Rust 製「deepseek-recipe」で整形する ^[2,19]

2.2 公式ベンチマーク（1017 版、自社測定）

表 3 公式ベンチマーク

ベンチマーク	1017	9 月版	Kimi K3	GLM-5.3	GPT-5.6 Sol	Claude Opus 5
HLE（ツールなし）	42.6 (44.0 †)	36.8 (39.1 †)	43.5	42.0 †	44.5	56.3
Terminal-Bench 4.0	42.3	31.2	12.6	37.9	39.9	51.8
DeepSWE v1.1	77.0	74.2	67.5	66.9	73.0	74.0
ProgramBench	29.1	20.3	17.5	19.0	23.0	37.0
NL2Repo-Bench	68.2	65.4	58.0	58.0	56.8	75.3
Automation-Bench v1.0.6	60.5	54.8	46.7	48.8	45.8	50.3
Agents' Last Exam	35.4	31.8	27.6	28.5	26.7	28.6
Apex Bench v1.1	72.1	64.6	47.8	56.6	54.3	65.8
Chartography（ツールあり）	83.0	78.9	68.1	—	79.9	84.0
BabyVision（ツールあり）	92.5	89.6	85.7	—	88.9	94.1
ZeroBench-main（ツールあり）	51.0	49.0	41.0	—	53.0	52.0

注：数値は DeepSeek の自社測定。太字は各行の最高値。† は HLE のテキストのみのサブセット。Terminal-Bench 4.0 は Claude Code ハーネス、DeepSWE v1.1 は Mini-SWE-V2 ハーネス、ZeroBench-main は Pass@5 で評価^[1,2]。

【分析】読み方

- 伸びが大きいのは、端末操作型のエージェント（Terminal-Bench +11.1）とプログラム構築（ProgramBench +8.8）である。一方、Opus 5 に対しては 4 勝 7 敗で、HLE では 13.7 ポイント差がある。「Opus 5 超え」はコーディング・エージェント系の一部指標に限られる。
- 比較相手の GPT-5.6 Sol と Opus 5 は最新世代ではない。Hacker News では、より新しい GPT-6.1 Sol との比較がないことが指摘されている^[5]。Claude Opus 5.5 も既に公開されており、比較対象の選び方には自社に有利なバイアスがあると見るべきである。
- 9 月版で公表された数学・推論系の値（GPQA Diamond 90.9、Codeforces 3471、MathArena Apex 65.6、CyberGym 88.1 等）^[1]は、1017 版の表に載っていない。1017 で数学や GPQA がどう変わったかは公式に示されていない。

2.3 既存 DeepSeek モデルとの比較

表 4 DeepSeek モデルの世代比較（公式値）

指標	V4-Flash-0731	V4-Pro GA (0813)	V4.1-Flash (9/10)	V4.1-Flash-1017
DeepSWE	54.4	62.7	74.2	77.0
Agents' Last Exam	25.2	25.7	31.8	35.4
HLE（ツールなし）	—	42.7	36.8	42.6
総／アクティブパラメータ	284B／13B	1.6T／49B	552B／8B・16B	同左

出典：^[1,15,20]。注：V4 系は DeepSeek Harness の minimal モードで測定されており^[1]、V4.1 系とはハーネス・ベンチマークの版が異なる可能性がある。厳密な同一条件比較ではない。

【事実】9 月の V4.1-Flash 公開時に旧 V4-Flash 系は退役した。V4-Pro については、9 月の公開告知で段階的な廃止を予告していたが^[21]、利用者の要望を受けて 9 月 14 日以降も API 提供を続けている^[1]。

2.4 API 価格・速度

表 5 公式 API 価格（1M トークンあたり、米ドル）

区分	1017 オフピーク	1017 ピーク	参考：V4-Pro オフピーク	参考：V4-Pro ピーク
入力（キャッシュヒット）	\$0.003	\$0.006	\$0.022	\$0.044
入力（キャッシュミス）	\$0.15	\$0.30	\$0.66	\$1.32
出力	\$0.60	\$1.20	\$1.98	\$3.96
画像入力	対応	対応	非対応	非対応

出典：^[13]。1017 の料金は 9 月版から据え置き^[1]。

- 【事実】ピーク時間帯は平日の 01:00–04:00 と 06:00–10:00（UTC）で、中国の祝日を除く^[13]。日本時間では 10:00–13:00 と 15:00–19:00 にあたり、日本の業務時間の大半がピーク料金帯に入る。

- **【事実】** 同日公開の Search/Fetch API は各\$0.75/1,000 リクエスト。DeepSeek Harness・Claude Code・Codex 経由でサーバー側から呼ばれる Search/Fetch は現在無料で、北京時間 2026 年 11 月 1 日 0 時から従量課金になる^[13,22]。
- **【二次情報】** OpenRouter では 1017 が 7 事業者から提供され、表示価格は入力\$0.08/出力\$0.50（第三者ホストを含む）^[23]。
- **【二次情報】** 速度：Artificial Analysis の計測では、9 月版の出力速度は 175.9 トークン/秒（77 モデル中 7 位）^[24]。1017 の独立計測はまだない。公式の「高速化」は主にタスク完了時間と出力長の短縮を指し、トークン生成速度の向上を示すものではない^[1]。中国メディアも同じ点を指摘している^[25]。

2.5 競合モデルとの比較

表 6 軽量モデルの比較（価格・仕様は公開値）

項目	DeepSeek V4.1-Flash (-1017)	Gemini 3.8 Flash	Qwen3.8 Flash
入力／出力価格 (1M)	\$0.15-0.30/\$0.60-1.20	\$0.75/\$3.75 (2026 年 12 月 31 日までの導入価格)	¥0.80/¥2.70 (北京リージョン)
入力モダリティ	テキスト・画像	テキスト・画像・動画・音声・PDF	未確認
重み	オープン (MIT)	クローズド	未確認
共通指標	9 月版：DeepSWE 74.2 / Terminal-Bench 4.0 31.2	DeepSWE 71 / Terminal-Bench 4.0 19.1	—

出典：DeepSeek は^[1,13]。Gemini・Qwen の値は第三者集計・比較記事による（一次ソースでは未確認）^[18,26-28]。Gemini 3.8 Flash は 2027 年 1 月に値上げ予定と報じられている^[26]。Claude Haiku 5.5 等との比較記事もあるが^[29]、信頼できる一次データが見つからないため本書では扱わない。

2.6 第三者評価（公式値との差）

- **【二次情報】 Artificial Analysis（9 月版、Max 設定）**：Intelligence Index は 39 で、120 モデル中 9 位（同規模オープンモデルの中央値は 19）。1 タスクあたりのコストは\$0.27。一方、指数の評価で消費した出力は 2.5 億トークンで、比較対象 46 モデル中**最も冗長**だった（中央値 1.3 億）^[24]。**【分析】** トークン単価が安くても、冗長さがタスクあたり費用を押し上げる。1017 の「出力短縮」がどこまで効くかが焦点になる。
- **【二次情報】 非ハルシネーション率 3.5% (AA-Omniscience)**：OpenRouter が Artificial Analysis の値として転載しているもので、AA 本体では未確認^[9]。**【分析】** 正しければ、知らないことを知らないと言えずに誤答する傾向が強いことを示し、知財用途では最も重要な注意点となる。
- **【二次情報】** Terminal-Bench 4.0 の AA 独自測定は 26.8%（9 月版）で、公式値 31.2 を下回る^[24]。

- **【二次情報】** 1017 版の独立スコアは 10 月 18 日時点で未公表^[4]。AA 指数を 45～46 と予測する記事があるが、推測にとどまる^[30]。
- **【二次情報】** LMArena（9 月版）：Text 1474（141 モデル中 28 位）、WebDev 1620（94 モデル中 17 位）。集計サイト経由の値で、LMArena 本体では未確認^[31]。対人の好みの評価では上位勢と差がある。
- **【二次情報】** OpenRouter の週間トークン数で V4.1-Flash が 1 位（36.8 兆トークン）^[32]。
【分析】 性能の裏付けというより、コスト優位の表れと読むべきである。
- **【事実】** 米 NIST/CAISI は V4-Pro（2026 年 5 月）について、フロンティアから約 8 か月遅れており、独自評価は DeepSeek の自己申告より低いと結論づけた^[33,34]。1017 は評価対象外だが、自社ベンチと独立評価が乖離する傾向には留意すべきである。**【分析】** CAISI は米政府機関であり、評価方法の妥当性には議論の余地がありうる。

3. 評判・評価

3.1 好意的な評価

- **技術メディア**：The Decoder は、より大きな競合をエージェントタスクで上回り、タスクあたりコストをさらに大きく下げたと評した。DeepSWE で 77.0%を 1 タスク約\$0.20 未満で達成し、前版（74.2%、約\$0.50）や競合（\$2～\$10 超）を大きく下回るとするが、DeepSeek 作成のグラフに基づく^[4]。
- **Hacker News**（507 ポイント、480 コメント）：処理速度と初回応答の速さ、思考過程（CoT）が見えること、サイバー関連の審査で突然ブロックされない開放性が好評だった。独自計測で、各推論強度で出力トークンが約 2 割減り、最低強度ではトークンが 2.5 分の 1 でも知能を維持したとの報告もある^[5]。
- **中国コミュニティ**：知乎のホットリストで 1 位となり、「許願式プログラミング（願うだけでプログラムができる）」の完成度を評価する声が多い^[25,35]。第三者ベンチ「Bug Hunt Bench」では、1017 が 105 バグ中 34 件を修復し（前版 21.7 件）、費用は\$0.9 で、Opus 5.5 の 41.7 件・\$58.5 と比べて費用対効果が際立つとの報告がある（二次情報）^[25]。
- **日本メディア**：PC Watch はコード系で Opus 5 を超えたと報じ、MIT ライセンスでの無償公開と商用利用可を紹介した^[36]。他の国内メディアも公開直後から取り上げている^[37]。

3.2 批判的な評価

- **日本語品質と指示追従**：ML_Bear 氏は激安と認めつつ、細かい指示を読み取れない点を致命的とし、精度は Sol・Opus に遠く及ばない、日本語が不得手、JSON Schema の strict モード非対応が不便と指摘し、精度を問わない大量処理向きではないかと評した^[6]。IT navi 氏は、日本語エッセイの表現が過剰で単調だと評価した^[7]。きしだ氏はローカル環境（GGUF）で試し、思考過程が長すぎて回答がなかなか出ない、質問に中国語で答える場面があったと報告している^[8]。

- **実務での信頼性**：Hacker News では、レビュー作業で GPT-6.1 Sol より約 4 倍時間がかかり、検出した問題も少なかったとの声がある^[5]。中国メディアも、月初の障害に起因する安定性への懸念や、既存コードの保守での信頼性が未検証である点を指摘している^[25]。
- **ベンチマークの信頼性**：公表値はすべて自社測定で、比較対象は旧世代、独立評価は未公表である^[2,4]。CAISI の V4-Pro 評価でも、自社申告と独立評価の乖離が指摘された^[33]。
- **安全性・検閲**：【二次情報】NIST/CAISI は 2025 年 9 月 30 日公表の評価で旧モデル（R1、R1-0528、V3.1）を調べた。R1-0528 ベースのエージェントがハイジャック指示に従う確率は米国モデルの平均 12 倍、一般的なジェイルブレイク手法の下で悪意ある要求に応答した割合は 94%（米国モデルは 8%）、中国共産党の不正確な主張を繰り返す頻度は米国モデルの 4 倍だった。報告は検閲がモデル自体に組み込まれていると結論づけ、Hugging Face から取得したモデルでも確認したとしている^[38]。【分析】1017 固有の評価はないが、ローカルで動かしても検閲バイアスは消えない前提で扱うべきである。
- **データ保護・政府の利用制限（日本）**：個人情報保護委員会は 2025 年 2 月 3 日、取得データは中国のサーバに保存され中国の法令が適用される旨を情報提供し、翌 4 日に林官房長官が会見で留意を呼びかけた^[39]。2 月 6 日には、デジタル社会推進会議幹事会事務局が政府機関等に注意喚起の事務連絡を出し、機密情報を取り扱わないこと、利用する場合は NISC とデジタル庁に助言を求めることを求めた^[40]。
- **データ保護・政府の利用制限（海外）**：WSJ を引く PYMNTS によれば、豪州、イタリア、台湾が利用禁止を実施した^[41]。台湾は公的機関での使用を全面禁止し^[42]、韓国は新規ダウンロードを一時停止した^[43]。米国では商務省の局が政府支給端末での利用を広く禁止したと報じられ^[44]、ニューヨーク、テキサス、バージニアの各州も政府端末で禁止し、21 州の司法長官が議会に全国的な禁止を求めている^[43]。米 FY2026 国防授權法による国防総省での利用禁止も報じられているが、法文は未確認である^[43]。

3.3 半導体・輸出規制の論点

- **学習チップ**：【事実】V4.1-Flash の技術報告は、学習に使ったハードウェアのベンダーを開示していない。記載は「GPU」という一般的な表現と、複数ハードウェアへの対応を意図した MXFP4/NVFP4 互換形式の採用にとどまる^[15]。【二次情報】V4 世代については、Huawei が学習の一部に自社チップが使われたと表明した^[45]。2026 年 9 月には、創業者の梁文鋒氏が投資家に、学習での国産チップ利用を拡大する意向を示したと報じられた^[46]。米国側は密輸された Nvidia チップの利用や米国モデルからの蒸留を疑うが、DeepSeek は否定または沈黙しており、事実関係は未確定である^[47,48]。
- **中国側の輸出管理（新たな論点）**：【二次情報】公開前日の 10 月 16 日、中国商務部などが AI 関連 23 技術を輸出管理リストに加える草案を公表した。対象は、学習計算量 10 の 26 乗 FLOP 以上、2 兆パラメータ以上、または「世界トップ 10」級のモデルの重み等である。公開済みのオープンソースは除外され、オープンウェイトの公開は登録で自動承認する仕組みが想定されている。意見募集は 11 月 15 日まで^[49,50]。【分析】1017 自体は公開済みのため

影響を受けにくい、将来のモデルが同じ条件で公開され続ける保証は弱まった（草案段階で運用は未確定）。

4. 知財業務への影響

4.1 業務別の評価

【分析】 以下は公開ベンチマークと評判からの推論である。知財タスクに特化した公開評価は見つからなかった。

表 7 知財業務別の適性

業務	効く点	限界・リスク	推奨度
先行技術調査・無効資料調査のスクリーニング	100 万トークンの文脈と低価格で、母集団の一次振り分けを大量に回せる。1 万件×2 万トークン（入力 2 億トークン）＋出力 500 万トークンで、オフピーク約\$33、ピーク約\$66（推論トークンは別途） ^[13]	ハルシネーション傾向が強い疑い（2.6 参照） ^[9] 。検索 DB の原文から根拠段落を示させ、人が監査する。記憶に頼って特許番号を挙げさせない	◎（公開文献に限る）
FTO（侵害予防調査）	クレームチャートの下書きを安く作れる	最終判断は法的評価そのもの。HLE の差が示すとおり、精緻な推論は上位モデルに劣る	△
IP ランドスケープ・特許分析	大量公報の要約・分類・技術要素の抽出を低コストで実行できる。オープンウェイトのため社内パイプラインに組み込みやすい	分類の一貫性・再現性は社内評価セットで検証が必要	○
中国語特許文献の読解	中国企業の開発で、中国語処理に強みがあると考えられる。CNIPA 公報の一次理解に向く	日本語出力の不自然さが指摘されている ^[6,7] 。「中国語→英語・要点抽出」は本モデル、「日本語の最終訳」は別モデルか人手、という分業が無難	◎
明細書作成支援	従属クレーム案や実施例展開の下書き程度	日本語品質と指示追従に批判がある。未公開発明を外部 API に入れるのは原則不可	×～△
中間処理（拒絶理由対応）	引用文献との相違点表の下書き	法的論理の質とハルシネーションが課題。審査官の引用箇所は原文での確認が必須	△
契約・ライセンス業務	英中二言語の契約の要約・比較	機密性が高く、公式 API は不適。ローカル運用が前提	△（ローカルのみ）

推奨度：◎＝推奨、○＝条件付きで推奨、△＝限定的な利用、×＝非推奨。

4.2 画像対応の可能性（図面・意匠・商標）

【事実】 チャート読解（Chartography 83.0）、視覚推論（BabyVision 92.5）、難問の視覚ベン

チ (ZeroBench 51.0) で前版より向上した^[2]。ただし、特許図面・意匠・商標に特化した評価は公表されていない。

【分析】図面の符号と明細書記載の突き合わせ、フローチャートやブロック図の構造化、意匠公報の図の説明文生成には有望である。一方、商標・意匠の類否検索には画像の埋め込み検索インデックスが別途必要で、本モデル単体では代替できない。PDF のネイティブ入力に対応しないため、公報 PDF は画像・テキストへの前処理が要る。この点では Gemini 3.8 Flash が優位である^[18]。

4.3 企業知財部門の利用リスク

1. **公式アプリ/API のデータ取扱い【事実】**：運営主体は杭州 DeepSeek 社で、Open Platform 規約は中国本土法を準拠法、杭州の裁判所を管轄と定める (2026 年 4 月 29 日発効)^[10]。プライバシーポリシーは、個人データを中国で直接収集・処理・保存すること、モデル学習への利用 (オプトアウト権あり)、アカウントが存続する限り入力データを保持することを定める^[11]。API 入力を学習に使わない旨の API 固有の約束は、一次ソースで確認できなかった。出力の権利は利用者に帰属するとされる^[10]。
2. **未公開発明の入力【分析】**：外部サービスへの入力が「公知」にあたるかは規約や状況次第で、法的に未確定の論点である。少なくとも守秘義務・営業秘密管理の観点から、公式 API への入力は避けるべきである。
3. **ローカル／オンプレ運用**：MIT の重みを社内・国内クラウドで動かせば、データの越境は回避できる。必要リソースは、FP8 原版でリポジトリ約 510GB^[19]、公式構成は 8GPU 以上との報道 (二次情報)^[25]、2.9bpw 量子化版で約 196GiB^[51]。Hacker News では Mac (M5 Ultra) で約 50 トークン/秒との報告もある^[5]。【分析】ただし、検閲バイアスや安全性の弱さは重みに内在するため、ローカル化しても解消しない (CAISI の旧モデル評価に基づく推論)^[38]。技術文書の処理に限れば実害は小さいと見られるが、交渉相手が中国企業である場合の分析等は、別モデルでクロスチェックすべきである。
4. **第三者ホスティング【分析】**：OpenRouter 等は複数の事業者に分けるため^[23]、処理する国・事業者を契約で固定する必要がある。
5. **社内ガイドライン【分析】**：日本政府の注意喚起 (2025 年 2 月)^[40]に沿い、「公式アプリ/API への業務情報の入力禁止」「ローカル・承認済みホストでの公開情報の処理は条件付きで許可」という二段構えが現実的である。

4.4 中長期的な影響【分析】

- **モデル層のコモディティ化**：オープンウェイトの軽量モデルが、数か月前のフロンティア級のエージェント性能 (の一部) を、タスクあたり 1 桁以上安い費用で提供すると試算も出ている^[4]。知財 AI ツールベンダーの差別化要因は、「どの LLM を使うか」から、特許 DB・分類／引用データ・ワークフロー・説明責任 (根拠提示、監査ログ) へ移る。ツール価格への下押し圧力も強まる。

- **内製化の現実味**：MIT の重みと低価格 API により、大手知財部門は社内 RAG で公報スクリーニングやランドスケープの下処理を内製しやすくなる。ただし、評価セットの整備と運用体制がボトルネックになる。
- **マルチモデル・ルーティングの標準化**：「大量処理は安価なオープンモデル、最終判断と日本語文書は上位モデルと人」という分業が合理的である。単一ベンダーへの依存は、価格改定、中国の輸出管理、米国の調達規制のいずれでも頓挫しうるため、モデルを差し替えられる設計にしておくことが重要である。
- **地政学リスクの常態化**：米国の政府調達規制と中国の重み輸出管理草案に挟まれ、中国製オープンウェイトを使い続けられるか自体が不確実になっている。採用するなら、公開済みバージョンの重みとハッシュ値を社内に保全しておく価値がある。

5. 推奨アクション

1. **当面の方針**：公式アプリ/API への未公開発明・契約書・社内情報の入力禁止とする。公開特許文献の処理に限り、ローカルまたは国内の承認済みホストで試験利用する。
2. **2〜4 週間の PoC**：自社の過去案件から約 100 件の評価セット（先行技術スクリーニング、中国語公報の要約、図面の符号照合）を作り、1017 と上位モデル（Claude・GPT・Gemini 系）を、精度・根拠提示率・1 件あたり費用で比較する。日本語の最終出力は別モデルまたは人が担う。公式 API でのバッチ処理は、日本時間の夜間・早朝・週末（オフピーク）に回せば料金は半額になる^[13]。
3. **監視項目**：Artificial Analysis 等による 1017 の独立スコア、Search/Fetch API の課金開始（北京時間 2026 年 11 月 1 日）、中国の輸出管理草案（意見募集は 11 月 15 日まで）、上位モデルの動向、日本政府のガイドライン更新。

6. 留保・未確認事項

- 公式ニュースページ（news261017）の本文は検索結果の抜粋でのみ確認した。WeChat 公式投稿は未確認。
- 1017 版の独立ベンチマーク・速度・ハルシネーション率は未公表である。本書の第三者値の多くは 9 月版 V4.1-Flash の値である。
- 非ハルシネーション率 3.5%、LMArena の順位、米国防授權法の規定、8GPU 構成、Bug Hunt Bench の結果は二次情報で、一次ソースでは未確認である。
- 参考文献のうち、ITmedia、Reuters、PYMNTS の記事は URL を確認できていない。標題を確認できなかった資料は、その旨を付記した。
- 学習に使ったチップは技術報告で開示されていない。米国の密輸・蒸留疑惑も事実関係は未確定である。

- 未公開発明の外部 AI への入力と新規性喪失の関係は、法的に確立した結論がなく、個別に専門家の判断が必要である。
- 第 4 章の知財業務評価は、公開ベンチマークとコミュニティの評判からの推論である。知財タスクに特化した公開評価は見つからなかった。

参考文献

本文中の上付き番号に対応する（初出順）。閲覧日はいずれも 2026 年 10 月 18 日。*は標題を URL から復元したもの。※は確認状況に関する注記。

- [1] DeepSeek「Change Log」 DeepSeek API Docs、2026 年 10 月 17 日・9 月 10 日・8 月 13 日・7 月 31 日の各項。<https://api-docs.deepseek.com/updates/>
- [2] DeepSeek-AI「DeepSeek-V4.1-Flash-1017（モデルカード）」 Hugging Face、2026 年 10 月 17 日公開。
<https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash-1017>
- [3] DeepSeek (@deepseek_ai)「DeepSeek-V4.1-Flash-1017 公開の告知投稿」 X、2026 年 10 月 17 日。
https://x.com/deepseek_ai/status/2111375640717033777
- [4] The Decoder「DeepSeek-V4.1-Flash-1017 beats larger rivals in agentic tasks while once again massively cutting per-task costs」* 2026 年 10 月。<https://the-decoder.com/deepseek-v4-1-flash-1017-beats-larger-rivals-in-agentic-tasks-while-once-again-massively-cutting-per-task-costs/>
- [5] Hacker News「DeepSeek-V4.1-Flash-1017 に関する議論スレッド（item 50129045）」 2026 年 10 月。
<https://news.ycombinator.com/item?id=50129045>
- [6] ML_Bear (@MLBear2)「DeepSeek-V4.1-Flash-1017 の使用感に関する投稿」 X、2026 年 10 月。
<https://x.com/MLBear2/status/2111408613143322888>
- [7] IT navi (@itnavi2022)「DeepSeek-V4.1-Flash-1017 の日本語文章評価に関する投稿」 X、2026 年 10 月。
<https://x.com/itnavi2022/status/2111474062841319448>
- [8] きしだなおき「DeepSeek-V4.1-Flash-1017 をローカルで試した記事」 はてなブログ、2026 年 10 月 18 日。
<https://nowokay.hatenablog.com/entry/2026/10/18/100508> ※標題未確認
- [9] OpenRouter「DeepSeek V4.1 Flash（モデルページ）」。<https://openrouter.ai/deepseek/deepseek-v4.1-flash>
※Artificial Analysis の値の転載を含む
- [10] DeepSeek「DeepSeek Open Platform Terms of Service」 2026 年 4 月 29 日発効。
<https://cdn.deepseek.com/policies/en-US/deepseek-open-platform-terms-of-service.html>
- [11] DeepSeek「DeepSeek Privacy Policy」。https://cdn.deepseek.com/policies/en-US/deepseek-privacy-policy.html?locale=en_US
- [12] DeepSeek「DeepSeek-V4.1-Flash-1017: Wish-Driven Programming Made Real」 DeepSeek API Docs News、2026 年 10 月 17 日。<https://api-docs.deepseek.com/news/news261017/> ※本文未確認（取得時に別ページへ転送。標題は検索結果の抜粋で確認）
- [13] DeepSeek「Models & Pricing」 DeepSeek API Docs、本書作成時に再確認。https://api-docs.deepseek.com/quick_start/pricing/
- [14] Baekpica「DeepSeek-V4.1-Flash-1017-VQ-GGUF」 Hugging Face。
<https://huggingface.co/Baekpica/DeepSeek-V4.1-Flash-1017-VQ-GGUF> ※第三者による量子化版のモデルカ

ード

- [15] DeepSeek-AI 「DeepSeek-V4.1-Flash: Pushing the Limits of KV Cache Compression」 arXiv:2609.19969v1、2026 年。 <https://arxiv.org/pdf/2609.19969v1>
- [16] DeepSeek-AI 「DeepSeek-V4.1-Flash (モデルカード)」 Hugging Face、2026 年 9 月 10 日公開。
<https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash>
- [17] NVIDIA 「deepseek-v4.1-flash Model Card」 * NVIDIA Build。 <https://build.nvidia.com/deepseek-ai/deepseek-v4.1-flash/modelcard>
- [18] AI Release Tracker 「DeepSeek V4.1 Flash vs Gemini 3.8 Flash (比較ページ)」 *。
<https://aireleasetracker.com/compare/deepseek/deepseek-v4.1-flash/google/gemini-3.8-flash> ※第三者集計
- [19] DeepSeek-AI 「DeepSeek-V4.1-Flash-1017 Files and versions」 Hugging Face。
<https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash-1017/tree/main>
- [20] DeepSeek 「DeepSeek-V4 公開告知」 DeepSeek API Docs News、2026 年 4 月 24 日。 <https://api-docs.deepseek.com/news/news260424/>
- [21] DeepSeek 「DeepSeek-V4.1-Flash 公開告知」 DeepSeek API Docs News、2026 年 9 月 10 日。 <https://api-docs.deepseek.com/news/news260910/>
- [22] AI Weekly 「DeepSeek opens V4.1-Flash-1017 weights and \$0.75 Search API」 * 2026 年 10 月。
<https://aiweekly.co/alerts/deepseek-opens-v41-flash-1017-weights-and-075-search-api>
- [23] OpenRouter 「DeepSeek V4.1 Flash 1017 (モデルページ)」 。 <https://openrouter.ai/deepseek/deepseek-v4.1-flash-1017>
- [24] Artificial Analysis 「DeepSeek V4.1 Flash (モデル評価ページ)」 。 <https://artificialanalysis.ai/models/deepseek-v4-1-flash> ※掲載値は 9 月版 V4.1-Flash のもの
- [25] 什么值得买 (SMZDM) 「DeepSeek-V4.1-Flash-1017 に関する各投稿」 2026 年 10 月。
<https://post.smzdm.com/p/apqmm7e7/> / <https://post.smzdm.com/p/aqrqqq3k/> /
<https://post.smzdm.com/p/a03xxxdr/> / <https://post.smzdm.com/p/avgpp6wn/> ※中国メディア・コミュニティ投稿 (二次情報)
- [26] ZenMux 「DeepSeek V4.1 Flash vs Gemini 3.8 Flash: A Practical Routing Cost Comparison」 *。
<https://zenmux.ai/blog/deepseek-v41-flash-vs-gemini-38-flash-a-practical-routing-cost-comparison>
- [27] ofox.ai 「DeepSeek V4.1 Flash vs Gemini 3.8 Flash vs Qwen3.8 Flash」 *。 <https://ofox.ai/blog/deepseek-v4-1-flash-vs-gemini-3-8-flash-vs-qwen3-8-flash/>
- [28] DataCamp 「DeepSeek V4.1 Flash vs Gemini 3.8 Flash」 *。 <https://www.datacamp.com/blog/deepseek-v4-1-flash-vs-gemini-3-8-flash>
- [29] Tech Insider 「DeepSeek V4.1 Flash vs Claude Haiku 5.5 vs Gemini 3.8 Flash (2026)」 *。 <https://tech-insider.org/deepseek-v4-1-flash-vs-claude-haiku-5-5-vs-gemini-3-8-flash-2026/>
- [30] OrcaRouter 「DeepSeek V4.1 Flash 1017 Artificial Analysis Rank」 *。
<https://www.orcarouter.ai/blog/deepseek-v4-1-flash-1017-artificial-analysis-rank> ※予測記事
- [31] Respan 「DeepSeek V4.1 Flash (モデルページ)」 。 <https://www.respan.ai/models/deepseek-v4.1-flash>
※LMArena 順位の集計
- [32] OpenRouter 「Rankings」 。 <https://openrouter.ai/rankings>
- [33] NIST (米国立標準技術研究所) 「CAISI Evaluation of DeepSeek V4 Pro」 * 2026 年 5 月。

- <https://www.nist.gov/news-events/news/2026/05/caisi-evaluation-deepseek-v4-pro>
- [34] Yahoo Tech「CAISI による DeepSeek V4 Pro 評価に関する報道」 2026 年 5 月。
<https://tech.yahoo.com/ai/articles/us-government-says-chinas-best-185845088.html> ※標題未確認
- [35] 知乎 (Zhihu)「DeepSeek-V4.1-Flash-1017 に関する質問ページ」 2026 年 10 月。
<https://www.zhihu.com/question/2094825372521178603>
- [36] PC Watch (インプレス)「DeepSeek-V4.1-Flash-1017 に関するニュース記事」 2026 年 10 月。
<https://pc.watch.impress.co.jp/docs/news/2149045.html> ※標題未確認
- [37] AI Revolution「DeepSeek-V4.1-Flash-1017 に関する解説記事」 2026 年 10 月。<https://ai-revolution.co.jp/media/deepseek-v4-1-flash-1017/> ※標題未確認
- [38] AI Business「NIST Report Pinpoints Risks of DeepSeek Models」* 2025 年。
<https://aibusiness.com/foundation-models/nist-report-pinpoints-risks-deepseek-models> ※NIST/CAISI 報告書 (2025 年 9 月 30 日) の本文は未確認
- [39] ITmedia NEWS「個人情報保護委員会の DeepSeek に関する情報提供および官房長官会見の報道」 2025 年 2 月。 ※URL 未確認
- [40] 日本経済新聞「政府機関等への DeepSeek 等生成 AI の業務利用に関する注意喚起の報道」 2025 年 2 月。
<https://www.nikkei.com/article/DGXZQOUA069NC0W5A200C2000000/> ※標題未確認
- [41] PYMNTS「WSJ 報道を引用した各国の DeepSeek 利用禁止に関する記事」 2025 年 2 月。 ※URL 未確認
- [42] ITmedia「台湾の公的機関での DeepSeek 使用禁止に関する報道」 2025 年 2 月。 ※URL 未確認
- [43] VDF AI「Is DeepSeek Safe?」*。<https://vdf.ai/blog/is-deepseek-safe> ※各国・各州の規制状況の整理 (二次情報)
- [44] Reuters「米商務省の政府支給端末での DeepSeek 利用禁止に関する報道」 2025 年 3 月。 ※URL 未確認
- [45] Tech Wire Asia「DeepSeek V4 and Huawei AI chips」* 2026 年 4 月。
<https://techwireasia.com/2026/04/deepseek-v4-huawei-ai-chips/>
- [46] TrendForce「DeepSeek Reportedly Expands Huawei Chips to AI Training; New Supply Could Arrive in 4Q26」* 2026 年 9 月 23 日。<https://www.trendforce.com/news/2026/09/23/news-deepseek-reportedly-expands-huawei-chips-to-ai-training-new-supply-could-arrive-in-4q26/> ※The Information 報道の転載 (二次情報)
- [47] Fortune「DeepSeek V4: AI model price-performance, China, open source」* 2026 年 4 月 24 日。
<https://fortune.com/2026/04/24/deepseek-v4-ai-model-price-performance-china-open-source/>
- [48] The China Academy「Why DeepSeek V4 Hasn't Fully Cut Ties with Nvidia」*。
<https://thechinaacademy.org/why-deepseek-v4-hasnt-fully-cut-ties-with-nvidia/>
- [49] The Next Web「China AI export controls draft」* 2026 年 10 月。<https://thenextweb.com/news/china-ai-export-controls-draft>
- [50] The Decoder「China plans to tighten export controls on open-weight models and 23 key AI technologies」* 2026 年 10 月。<https://the-decoder.com/china-plans-to-tighten-export-controls-on-open-weight-models-and-23-key-ai-technologies/>
- [51] LibertAIDAI「DeepSeek-V4.1-Flash-1017-EXL3-2.9bpw」 Hugging Face。
<https://huggingface.co/LibertAIDAI/DeepSeek-V4.1-Flash-1017-EXL3-2.9bpw> ※第三者による量子化版のモデルカード

補足：調査過程で参照したが本文で個別に引用していない資料

- (a) Yahoo!ニュース「DeepSeek 関連の国内報道」。
<https://news.yahoo.co.jp/articles/422fb793472a598dfdab2753ef7bd9212ab3e964> ※標題未確認
- (b) npaka (note)「DeepSeek 関連の解説記事」。<https://note.com/npaka/n/nebbfa829304b> ※標題未確認
- (c) SmartScope「DeepSeek V4.1 Flash 1017: release, benchmarks, price」*。
<https://smartscope.blog/blog/deepseek-v4-1-flash-1017-release-benchmarks-price-2026/>
- (d) Kingy AI「DeepSeek V4.1 Flash 1017: specs, benchmarks, pricing」*。<https://kingy.ai/blog/deepseek-v4-1-flash-1017-specs-benchmarks-pricing/>
- (e) MorphLLM「DeepSeek API」*。<https://www.morphllm.com/deepseek-api>
- (f) BenchLM「DeepSeek API pricing」*。<https://benchlm.ai/deepseek/api-pricing>
- (g) chat-deep.ai「Pricing」*。<https://chat-deep.ai/pricing/>
- (h) Requesty「DeepSeek V4.1 Flash (モデルページ)」。<https://www.requesty.ai/models/deepseek/deepseek-v4.1-flash>
- (i) OpenRouter「Models」。<https://openrouter.ai/models>