

DeepSeek-V4.1-Flash-1017 調査報告

調査基準日：2026年10月18日（日本時間）

対象：DeepSeek-V4.1-Flash-1017（2026年10月17日公開）

エグゼクティブサマリー

- **DeepSeek-V4.1-Flash-1017は、新アーキテクチャへの世代交代というより、9月10日公開のV4.1-Flashに対する大規模なポストトレーニング／強化学習アップデートと見るのが正確である。** DeepSeek自身も「minor update」と位置付けつつ、推論、コーディング、エージェント、視覚推論を改善し、同時に実行時間と出力長を短縮したと説明している。APIモデル名は引き続き `deepseek-flash` で、APIの呼び出し方法・価格体系は変更されていない。 ¹
- **性能改善は特にソフトウェア開発・エージェント業務で大きい。** 公式評価では、前版V4.1-Flash比で Terminal-Bench 4.0が31.2→42.3、ProgramBenchが20.3→29.1、Automation-Benchが54.8→60.5、DeepSWE v1.1が74.2→77.0へ上昇した。DeepSWEでは公式比較表上Claude Opus 5やGPT-5.6 Solを上回る一方、HLE、Terminal-Bench、NL2RepoなどではClaude Opus 5に依然及ばない。つまり「全面的な最高性能」ではなく、**コードエージェントの価格性能比に突出したモデル**という評価が妥当である。 ²
- **ただし、独立評価との乖離は無視できない。** OpenRouter経由で掲載されているArtificial Analysisの1017評価では、HLE 39.2%、Terminal-Bench 4.0 26.8%で、DeepSeek公式の42.6%、42.3%より低い。評価ハーンズ、推論設定、エージェント構成が異なるため直接比較はできないが、少なくとも「公式ベンチマーク値をそのまま実運用品質とみなすべきではない」ことを示す。 ³
- **知財部門にとって最大の価値は、1Mトークン級の長文処理、画像・OCR、検索／Fetch、コード・エージェント能力を一体化できる点である。** 特許明細書群、審査経過、先行技術文献、契約書、ソースコード、図面・チャートの横断分析を低コスト化できる可能性が高い。一方、**特許・法務ドメインについての独立したRecall/Precisionベンチマークは1017公開時点で未公開**であり、特許性判断やFTO、商標クリアランスの最終判断を任せられる根拠はない。 ⁴
- **知財・コンプライアンス面では、API利用と自社ホストを明確に分けるべきである。** 重みはMITライセンスだが、DeepSeekのホステッドサービスでは入力権利を利用者が保証し、出力についてDeepSeekが保有し得る権利を利用者へ譲渡する一方、非侵害性・正確性は保証されず、利用者責任が大きい。さらに現行プライバシーポリシーでは入力を収集しモデル改善に利用し得ること、中国所在のサーバーに個人情報を保存すること、モデル訓練利用のオプトアウト権が明記されている。未公開発明、営業秘密、第三者秘密情報を扱う日本企業では、**原則「公開情報はAPI、機密情報は自己ホスト／承認済み閉域環境」**という分離が合理的である。 ⁵

製品・技術基盤

製品の位置付けとバージョン差分

開発・提供主体は Hangzhou DeepSeek Artificial Intelligence Co., Ltd.（杭州DeepSeek／深度求索）。公式Change Logで1017への更新日は**2026年10月17日**と確認できる。Web、アプリ、APIが同時更新され、APIでは従来どおり `deepseek-flash` を指定する。 ⁶

名称について注意が必要である。ユーザー指定の「DeepSeek V4.0」は一般的な世代呼称としては理解できるが、DeepSeekの公式資料で比較可能な旧モデル名は主として **DeepSeek-V4-Flash**、**DeepSeek-V4-Pro**、**DeepSeek-V4.1-Flash** である。本報告では「V4.0相当の旧世代」を、公式数値が存在するV4-Flash、とりわけ0731版を基準として扱う。V4-Flash 0731のDeepSWE v1.1は54.4、NL2Repoは54.2、Automation-Benchは25.1だったのに対し、1017ではそれぞれ77.0、68.2、60.5まで上昇している。 ⁷

1017固有の変更点は、DeepSeekによれば、**自己フィードバックを用いた強化学習を通じたエンドツーエンドのプロジェクト開発、ゲーム作成、3Dモデリング等の改善**、推論・コーディング・エージェント・視覚能力の改善、実行時間および出力長の削減である。DeepSeek Harness、Tencent CodeBuddy、WorkBuddy等への最適化も告知された。 ⁸

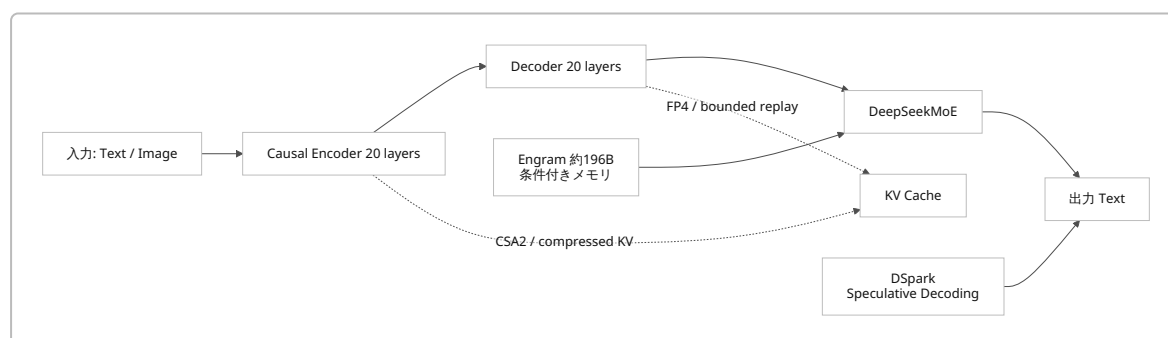
一方、1017専用の新しいネットワーク構造、専門家数、層数等は発表されていない。したがって、**1017でアーキテクチャ自体が変更されたとの公開根拠はなく、V4.1-Flashの同一モデルライフサイクル内のポストトレーニング更新とみるのが合理的**である。これは公式の「minor update」という表現、およびV4.1の訓練情報開示が「subsequent releases in the model lifecycle」にも適用されるとの記載からの推論である。 ⁹

アーキテクチャとモデル規模

ベースとなるV4.1-Flash技術報告では、**40層のCausal Encoder-Decoder (CED)** 構成で、20層のcausal encoderと20層のdecoderを持つ。言語バックボーンは**552Bパラメータ**で、推論時のアクティブパラメータはprefill約8B、decode約16B。DeepSeekMoE、CSA2、Engram、Single-Pass mHC、DSpark speculative decoding、FP4 KV cacheなどを組み合わせ、ネイティブな視覚エンコーダも備える。 ¹⁰

Hugging Face上の1017リポジトリ表示では総パラメータ数は**約763B**。技術報告上は552Bのバックボーンとは別に約196BのEngram条件付きメモリが記載されており、総リポジトリ規模と「実際に毎トークン計算するパラメータ数」が大きく異なるMoE系モデルである。したがって「763Bモデルだから毎トークン763Bを計算する」という理解は誤りである。 ¹¹

V4.1の大きな特徴がメモリ効率である。CSA2とFP4 KV cacheにより、技術報告は**global KV storageを約890 bytes/token**、**V4-Flashのおよそ4分の1**と説明する。さらにSliding-Window Attentionのbounded replayを組み合わせ、持続的KVストレージを旧V4-Flashのおよそ8分の1まで削減する設計を報告している。これは1Mトークン文脈を現実的な推論コストで扱うための中核技術である。 ¹²



この図はV4.1技術報告に基づく概念整理であり、1017固有の構造変更を示すものではない。 ¹⁰

学習データ

V4.1技術報告は**約45兆トークンのマルチモーダル事前学習**を記載している。テキストのみとマルチモーダルデータはおおむね7:1で構成され、コードについては新しいオープンソースリポジトリ、コミット、ライブラリ、新興フレームワーク等を追加したとされる。マルチモーダル側ではウェブの画像-テキスト対、インター

リープされたウェブページ／PDF、OCR、visual grounding、画像-コード、computer-use trajectories等を利用した。¹³

さらに2026年9月24日付の**DeepSeek-V4.1 Public Summary of Training Content**は、テキストが10兆トークン超、画像が10億枚超であると開示し、Common Crawl、Stack Exchange、多分野の公開テキスト・コード、公開画像、商用ライセンスデータ、第三者との協業で得た非公開データを含むとしている。¹⁴

重要なのは、同開示が**ユーザーとのAIインタラクションを訓練に使用することを「Yes」と明記している点**である。テキストと画像が対象で、匿名化・非識別化を行い、オプトアウトを提供するとしている。ポストトレーニングでは社内従業員・外部パートナーのフィードバックもエージェント学習環境の構築に使われる。¹⁵

なお、開示には注意点もある。「Latest date」欄では一部データについて概ね2025年5月までの記述がある一方、DeepseekBotのクロール期間は「approximately 2024-2026」とされている。両者の範囲関係は資料だけでは完全には明確でない。さらに**1017アップデートに追加されたRLデータの具体的件数、データセット名、権利処理内訳は未公開**である。¹⁵

DeepSeekはクローラがrobots.txt等の標準プロトコルを尊重し、CAPTCHA、paywall、パスワード保護を回避するよう設計されていないと説明し、TDMのrights reservation／opt-out signalも尊重するとしている。ただし、EU GPAI Code of Practiceへの署名者ではないことも同じ開示に記載されている。これはあくまで**事業者の自己申告による透明性資料**であり、個々の訓練コンテンツについて第三者権利処理を保証するものではない。¹⁶

ライセンスと利用規約

モデルリポジトリおよび重みは**MIT License**で公開されている。DeepSeek自身もモデルweights、parameters、inference tool codeをMITで提供すると明記している。したがって、自己ホスト、改変、商用組込みについてライセンス上の自由度は非常に高い。¹⁷

ただし、**MITライセンスとDeepSeekのホステッドサービス利用規約は別物**である。2026年3月27日更新のTerms of Useでは、入力について必要な権利・ライセンス・許可を利用者自身が有することを保証する義務があり、出力についてはDeepSeekが有する「if any」の権利・権原・利益を利用者に譲渡するとしている。出力が非固有で、他ユーザーに類似出力が提供され得ることも明示されている。¹⁸

さらにホステッドサービスについては、出力の正確性、最新性、信頼性、**非侵害性を保証しない**。利用規約上の総責任額も、原則として直前12か月に当該サービスについて支払った金額または100米ドルのいずれか高い方に制限され、準拠法は中国本土法、紛争管轄はDeepSeek登録所在地の裁判所とされている。日本企業の知財用途では重要な契約リスクである。¹⁹

また、ホステッドサービスから生成物を公表・頒布する場合、利用規約は事実確認とともに**AI生成であることを明示すること**を利用者に求めている。自己ホストしたMITモデルについて同一のサービス規約が当然に適用されるとは限らず、導入形態ごとの契約整理が必要である。²⁰

性能と競合比較

公式ベンチマーク

以下はDeepSeek公式の1017モデルカード／Change Logに基づく。同一表中の競合値もDeepSeek掲載値であるため、**独立第三者が統一条件で再測定したランキングではない点**に注意が必要である。¹

ベンチマーク	V4.1 Flash 前 版	V4.1- Flash-1017	改善 幅	Kimi K3	GLM-5.3	GPT-5.6 Sol	Claude Opus 5
HLE・toolsなし	36.8	42.6	+5.8	43.5	42.0*	44.5	56.3
Terminal-Bench 4.0	31.2	42.3	+11.1	12.6	37.9	39.9	51.8
DeepSWE v1.1	74.2	77.0	+2.8	67.5	66.9	73.0	74.0
ProgramBench	20.3	29.1	+8.8	17.5	19.0	23.0	37.0
NL2Repo	65.4	68.2	+2.8	58.0	58.0	56.8	75.3
Automation- Bench	54.8	60.5	+5.7	46.7	48.8	45.8	50.3
Agents' Last Exam	31.8	35.4	+3.6	27.6	28.5	26.7	28.6
Apex Bench v1.1	64.6	72.1	+7.5	47.8	56.6	54.3	65.8
Chartography + tools	78.9	83.0	+4.1	68.1	未掲載	79.9	84.0
BabyVision	89.6	92.5	+2.9	85.7	未掲載	88.9	94.1
ZeroBench-main Pass@5	49.0	51.0	+2.0	41.0	未掲載	53.0	52.0

* GLM-5.3のHLE値は公式表のtext-only条件。全数値はDeepSeek公式モデルカード掲載値。 ²¹

相対的な改善率で見ると、特にProgramBenchが約43%、Terminal-Bench 4.0が約36%、HLEが約16%、Apexが約12%改善している。逆にDeepSWE、NL2Repo、BabyVisionなどは数ポイントの上積みであり、1017は「全能力を均等に上げた」というより、**複雑なエージェント実行・プロジェクト遂行能力を重点的に改善した更新**と読む方がデータに整合する。 ²²

旧V4-Flashとの世代差はより大きい。V4-Flash 0731ではDeepSWE 54.4、NL2Repo 54.2、Automation-Bench 25.1、Agents' Last Exam 25.2だったため、1017はそれぞれ77.0、68.2、60.5、35.4となっている。特にAutomation-BenchはV4世代から2倍以上となる。 ⁷

独立ベンチマーク

公開翌日時点で確認できる最も有用な1017固有の独立評価の一つが、OpenRouter上に掲載されるArtificial Analysisデータである。Artificial Analysisは、実際のserverless endpointを利用した顧客視点の性能測定を方法論に含む。 ³

独立指標	1017結 果	読み方
Artificial Analysis Intelligence Index	39.5	総合知能指標。公式ベンチとは別体系
HLE	39.2%	公式42.6%より低い

独立指標	1017結果	読み方
AA-LCR	84.0%	長文・文脈能力系
GDPval-AA	55.0%	実務タスク系
SciCode	51.9%	科学コーディング
Terminal-Bench 4.0	26.8%	公式42.3%と大きな差
AA Omniscience Accuracy	46.4%	知識・回答精度系
AA Omniscience Non-hallucination Rate	3.5%	テスト固有指標。一般的な「96.5%幻覚率」という意味ではない

出典：OpenRouter掲載Artificial Analysis評価。 ²³

特にTerminal-Benchの**42.3対26.8**という差は重要である。DeepSeek公式ではClaude Code／Mini-SWE系など指定ハーネスを使い、1017自体も特定ハーネス向けに最適化されている。したがって、エージェントモデルは「基盤モデル単体の能力」以上に、system prompt、tool schema、reasoning effort、timeout、コンテキスト圧縮、再試行方式などでスコアが変動する。知財部門がPoCを行う場合も、一般ベンチではなく**自社のエージェント・検索環境そのもので測ることが不可欠**である。 ²⁴

推論速度、メモリ、実運用レイテンシ

1017固有の公式tokens/sec、time-to-first-token、P50/P95 latency、同時リクエスト別throughputの数値は、2026年10月18日時点で未公開である。技術報告はEPD disaggregationや多数の最適化kernelによるhigh-throughput／low-latencyを説明するが、1017での本番API実測数値を公表してはいない。 ¹⁰

そのため、公開数値の取り扱いには次のように分けるべきである。

項目	公開情報	信頼度・留意点
KVキャッシュ効率	約890 bytes/token、V4 Flash比約1/4	V4.1技術報告による公式アーキテクチャ値。 ¹²
永続KVストレージ	旧V4 Flash比およそ1/8	bounded replayを含む設計値。 ¹⁰
コンテキスト	推奨1M tokens	公式モデルカード。 ²¹
1017公式tokens/sec	未公開	ベンダー別・hardware別に検証が必要。 ⁸
OpenRouter 1017 latency/throughput	プロバイダー依存	OpenRouterには複数providerがあり、単一のモデル固有値とはならない。 ²³
V4.1 Flash実運用参考値	4×H200自己ホスト試験でピーク約213 output tokens/s	9月版V4.1であり1017ではない 。実ワークロードではcache reread、起動安定性等がボトルネックになった。 ²⁵

4基のH200を借りた企業試験では、9月版V4.1 Flashでピーク約213 tokens/sが報告された一方、エージェントが会話履歴を大量に再送するため計算の多くが再読込に使われ、オンデマンドGPUレンタルはDeepSeek APIより高かった。この事例は、「**オープンウェイト＝自己ホストすれば必ず最安**」ではないことを示す好例である。ただし1017の直接評価ではないため参考値にとどめるべきである。 ²⁵

評判・専門家・コミュニティ評価

現時点の評価は「有望だが未成熟」

1017は本調査日の**前日公開**である。このため、長期間の本番運用、査読論文、再現研究、障害率、企業導入後のTCOといった本来重要な情報はまだほぼ存在しない。1017固有の査読済み学術論文・独立学術検証は確認できず、学術的な一次資料は主としてV4.1のDeepSeek自身によるarXiv技術報告である。同報告は2026年9月17日に投稿された開発者論文であって、1017を独立検証した査読論文ではない。 ²⁶

日本語メディアではPC Watchが10月18日、DeepSWEで77.0を記録し公式比較上Opus 5を上回ったこと、Terminal-Bench 4.0が31.2→42.3へ改善したことを紹介する一方、HLEやTerminal-BenchではOpus 5が上である点も報じている。この報道姿勢は「コード系の一部で最上位、全般では閉源フロンティアモデルに未到達」という今回のデータと整合する。 ²⁷

9月のV4.1正式公開時にもReutersは、DeepSeekがV4.1を新アーキテクチャ系統の小型／高速側モデルとして位置付け、高速推論、高throughput、将来の大型モデルへのscalingを狙っていると報じた。1017はこのV4.1 Flashの「速度・コストを維持しながら能力を引き上げる」方向を強化した更新とみられる。 ²⁸

Reddit、X、日本語圏

コミュニティ評価は、現時点では**熱狂的な第一印象と慎重論が混在**している。日本語Xでは1017についてDeepSWEの性能と低コストを理由に「正直やばい」と評価する投稿が確認できるが、これは個人のリリース直後の反応であり、独立した実運用試験とは区別すべきである。 ²⁹

Reddit / LocalLLaMAではV4 Flash/V4.1系列について、大規模コード変更時のコンテキスト管理、複数ツール呼び出し、100回超のtool callの安定性を高く評価するユーザー報告がある一方、「token generationが遅い」「計画・実行に数分考える」とする指摘もある。これは1017以前の系列モデルの体験であるため、1017の短縮されたruntimeでどこまで解消されたかは**未検証**である。 ³⁰

1017固有のReddit、X、日本語フォーラムについては、公開後24時間程度しか経過していないため、母数のある定量的sentiment分析を行える状態ではない。現段階で「ユーザー評価は高い／低い」と統計的に断定するのは不適切である。

長所と短所の総合評価

観点	評価
コーディング／エージェント	非常に強い 。DeepSWE、Automation、ALE、Apexで公式比較上トップクラス。 ²¹
長文	強い設計 。1M contextと圧縮KVが知財文書群に適する。ただし1M全域の日本語特許検索精度は未検証。 ³¹
画像・図面	ネイティブmultimodal、OCR・chart系の改善あり。図面・チャートレビューに有望。 ³²
コスト	非常に低いAPI／オープンウェイトが強み。自己ホストはhardware運用次第で必ずしも安価ではない。 ³³
メモリ効率	CSA2/FP4等により旧Flashから大幅改善。 ²⁶
一般推論	強いが、HLE等ではClaude Opus 5など上位閉源モデルに差が残る。 ²¹

観点	評価
再現性	注意。 Terminal-Benchで公式と独立評価の差が大きい。 ³⁴
ハルシネーション	DeepSeek自身も非事実情報を生成し得るとして保証していない。独立評価でもcalibration面に注意を要する。 ³⁵
日本語知財	専用ベンチ未公開。 日本語特許、判例、J-PlatPat検索について1017固有の精度データなし。
成熟度	公開翌日のため、長期運用品質・学術再現・企業事例が不足。

知財業務への影響と法的リスク

生成物の著作権

日本では、AI生成物の著作物性について文化庁が**人間の「創作意図」と「創作的寄与」**を中心に検討する枠組みを示している。したがって、「DeepSeekの利用規約が出力権利をユーザーへ譲渡する」とこと、「その出力に日本法上そもそも著作権が成立する」ことは別問題である。DeepSeekが譲渡できるのは規約自身がいのように「if any」の権利であり、契約だけで著作物性を創設することはできない、という整理になる。 ³⁶

実務上、AIに「この特許の要約を書け」の一指示だけでほぼ自動生成された文章より、人間が構成、表現、段落、技術的主張を繰り返し選択・修正した文書の方が、人間の創作的寄与を立証しやすい。このため、重要な知財成果物については**プロンプト、採否、編集履歴、人間の修正内容を保存することが有効**である。これは文化庁の創作的寄与の枠組みから導く実務的提案である。 ³⁷

一方、AI出力による既存著作物の侵害については、通常の著作権侵害と同様、既存著作物との**類似性・依拠性**が中心論点になる。文化庁資料はAI学習・生成についてもこの枠組みを整理しており、「作風」の共通のみでは直ちに著作権侵害にならない一方、具体的な創作的表現を直接感得できる生成物では侵害となり得るとしている。 ³⁸

学習データの権利問題

日本の著作権法30条の4について、文化庁はAI学習を含む情報解析が、著作物の思想・感情を「享受」することを目的としない場合には権利制限の対象となり得ると整理する。ただし、享受目的が併存する場合や、権利者の利益を不当に害する場合などには同条の適用が問題になる。特定作品の創作的表現を意図的に出力させる過学習などは特に慎重な領域である。 ³⁹

DeepSeek-V4.1については、公開ウェブ、Common Crawl、Stack Exchange、ソースコード、科学技術資料、10億枚超の画像、ライセンスデータ、第三者協業データ、ユーザーインタラクション等が訓練情報として開示されている。ただし**個々の作品・コードリポジトリごとの権利状態、ライセンス遵守状況、1017強化学習用データの全リストは未公開**である。したがって、企業利用者が「学習データは全て権利クリア済み」と表明・保証できるだけの公開証拠はない。 ⁴⁰

商用利用と契約リスク

特に注意すべきなのは、MITライセンスによる**モデル利用権の広さ**と、**生成コンテンツの非侵害保証の有無**を混同しないことである。MITで重みを商用利用できても、そのモデルが生成した文章・コード・画像解釈結果が第三者権利を侵害しないことまではMIT Licenseは保証しない。ホステッドサービス規約でもDeepSeekは出力の非侵害性を保証しない。 ⁴¹

これはコード生成で特に重要である。ソースコードを知財資産として製品へ組み込む場合、生成コードについてOSS license scan、copyright similarity、dependency/SBOM確認を別工程にする必要がある。「モデルがMITだから生成コードもMIT」と考えるべきではない。 42

特許と営業秘密

日本の特許制度について特許庁資料は、AI利用発明でも従来どおり**自然人の発明者を認定する**考え方を示している。したがって、DeepSeekが技術案を生成した場合でも、出願時には、誰が課題設定、モデル・データの選択、プロンプト設計、結果の評価・選択、技術的具体化に実質的に寄与したかを記録しておくことが重要になる。 43

さらに、未出願発明を外部APIへ投入する問題は著作権よりも**秘密管理・新規性・契約**の観点で厳しく見るべきである。DeepSeekの現行ポリシーは入力を収集し、サービス改善・機械学習技術の訓練・改善に利用し得ること、中国所在サーバーで情報を保存し得ることを明示している。訓練利用はオプトアウト可能だが、「オプトアウト＝入力が一切保存・処理されない」ことを意味しない。 44

したがって、出願前発明、研究ノート、発明届、未公開クレーム、共同研究先の秘密情報を標準APIへそのまま投入するのは、企業の秘密管理規程上は高リスクと評価すべきである。**送信しただけで直ちに特許新規性が失われる、と一律に断定することはできないが**、秘密保持義務・アクセス範囲・データ保持・二次利用を契約上確認できない環境へ出願前情報を渡すことは避けるのが安全である。 45

商標・ブランド

モデルにブランド名候補、商品名、ロゴ文言を生成させること自体は便利だが、**AIが商標クリアランスを代替するものではない**。利用規約も第三者権利非侵害を保証しないため、生成候補はJ-PlatPat等による同一・類似商標調査、指定商品・役務、著名表示、不正競争法まで別途確認すべきである。なお「DeepSeek」等のDeepSeekブランドを許諾なく使用・登録すること自体も利用規約で制限されている。 46

情報管理と越境データ

現行日本語プライバシーポリシーでは、プロンプト、アップロードファイル、写真、チャット履歴等を収集し得る。また検索機能については、第三者APIへ**入力キーワードを共有する場合があります**と明記している。さらに個人情報を中国所在の安全なサーバーに保存するとしている。検索を伴う知財業務では、「DeepSeek本体に何を送るか」だけでなく、**検索クエリが第三者へ流れることも脅威モデルに含めるべきである**。 47

リスクと対応策

リスク	知財部門への影響	リスク水準	推奨対策
未公開発明の外部送信	営業秘密・出願戦略・契約秘密の漏えい	高	API投入禁止または企業契約／自己ホスト。データ分類を先行。 44
第三者著作物を含む入力	利用者が入力権利を保証する契約責任	高	入力元ライセンス確認、RAG corpus台帳化。 46
AI生成コード	著作権・OSS license contamination	高	SCA、license scan、similarity scan、人間review。 46

リスク	知財部門への影響	リスク水準	推奨対策
AI生成文章・図表	既存著作物との類似・依拠	中～高	similarity check、引用元確認、AI利用履歴保存。 ⁴⁸
ハルシネーション	架空特許・判例・引用・法令	高	原典URL/公報番号を必須化し、一次資料で検証。DeepSeek自身も正確性を保証しない。 ⁴⁹
特許発明者認定	AI寄与と人間寄与の混同	中～高	human contribution log、発明者ヒアリング。 ⁴³
商標案生成	先行商標との衝突	中	正式商標検索を別工程化。 ⁴⁶
モデル訓練への入力利用	秘密情報の二次利用懸念	高	“Improve the model for everyone”無効化だけに依存せず、契約でno-trainingを取得。 ⁴⁵
中国データ保存	個人情報・顧客契約・越境管理	高	API用途を低機密情報に限定、法務・個情委対応確認。 ⁵⁰
ベンダー規約	中国法準拠、責任上限、非侵害保証なし	高	reseller/enterprise契約で修正交渉、自己ホストも検討。 ¹⁹
モデル更新	deepseek-flash の中身が更新され得る	中	version pinning、回帰テスト、変更管理。1017も同一API名への更新だった。 ⁷

知財部門への実務的示唆

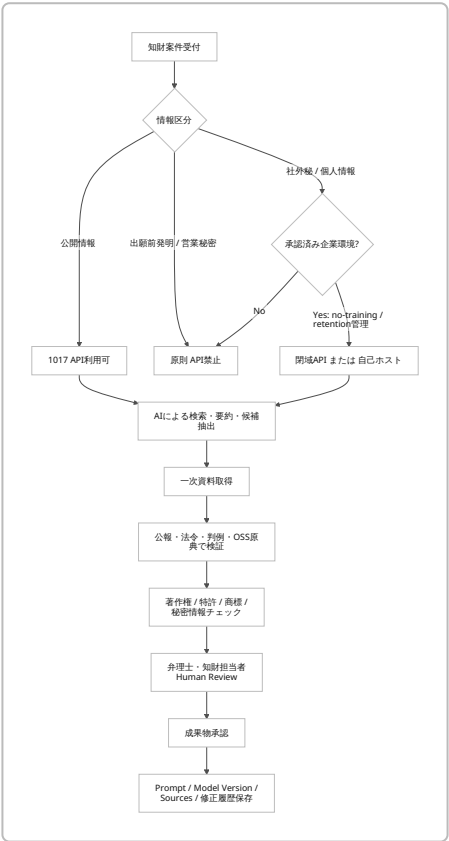
まず採用すべき用途、まだ任せるべきでない用途

現段階で最も合理的なのは、DeepSeek-V4.1-Flash-1017を「**知財専門家を置き換える判断エンジン**」ではなく、「**大量資料を読む・候補を出す・整理するエージェント**」として導入することである。長文・画像・コード・検索・エージェント能力というモデル特性と、未成熟な独立検証・ハルシネーションリスクの双方を踏まえた結論である。 ⁵¹

適性が比較的高いのは、公開特許の要約・分類、ファミリー情報の整理、技術分類案、公開先行文献の候補抽出、審査経過の時系列整理、契約条項の比較、OSS inventoryの一次整理、図面・チャートの説明、発明者ヒアリング用質問案、公開情報を用いたランドスケープ作成である。1M contextと画像理解、agentic benchmarkの強さがこれらと整合する。 ³²

逆に、1017単独での「新規性あり／なし」「進歩性あり／なし」「FTO問題なし」「この商標は登録可能」「この契約は適法」「この特許は無効」といった結論には、現時点で十分な根拠がない。特許検索に重要な **recall、family normalization、legal-status accuracy**、日本語FI/F-term/CPC処理についての1017独立ベンチマークは未公開である。

推奨ワークフロー



この運用は、DeepSeek自身が法的・専門的判断への依存を避けるよう求め、人に重大な影響を与える用途ではhuman reviewを要求していること、日本政府のAI事業者ガイドラインが2026年3月に第1.2版へ更新されていることとも整合する。 52

導入チェックリスト

知財部門として最低限、以下を導入前ゲートにするとよい。

確認項目	合格条件
モデル識別	DeepSeek-V4.1-Flash-1017 を記録。APIのエイリアス名だけを保存しない。 7
データ分類	公開／社外秘／営業秘密／個人情報／第三者秘密を分類
訓練利用	APIではtraining opt-out確認。高機密ではopt-outだけを安全根拠にしない。 45
保存場所	中国サーバー保存可否を情報管理規程・顧客契約と照合。 50
Search利用	検索keywordの第三者API共有を許容できるか確認。 53
入力権利	第三者文献、契約書、ソースコードを処理する権限確認。 46
出力著作権	人間の創作的寄与を記録。 37
出力侵害	類似性、OSS license、商標等を別検査
事実確認	特許番号・判例番号・引用・URLを原典照合

確認項目	合格条件
発明者	AIではなく人間の具体的創作寄与を記録。 43
公開表示	ホステッドサービスから公開する場合のAI生成表示要件を確認。 20
変更管理	モデル更新時にgolden setを再実行。1017も同一APIモデル名で更新された。 7

社内利用ルール案

生成AI知財利用ルール案

DeepSeekその他の外部生成AIサービスには、出願前発明、未公開クレーム、営業秘密、顧客秘密、アクセス制限された第三者資料、個人情報、情報管理責任者の承認なく入力してはならない。AIが提示した特許番号、文献、判例、法令、ライセンス、商標その他の法的情報は、必ず一次資料で検証する。AI出力を出願書類、鑑定、FTO意見、契約書または対外的な知財判断に使用するときは、資格・権限を有する担当者によるレビューを必須とし、使用モデル、バージョン、主要プロンプト、参照資料、人間による修正内容を案件記録に保存する。

このルール案は、DeepSeekの入力権利保証、非侵害・正確性の非保証、モデル訓練へのデータ利用可能性、文化庁のAI著作権整理を踏まえたリスクベースの提案である。 54

契約条項例

企業向けAPI、Sler、国内リセラー等との個別契約では、標準DeepSeek規約をそのまま受け入れる前に、少なくとも次のような特約を検討する価値がある。標準規約が中国法準拠、出力非侵害を非保証、責任額を限定しているためである。 19

入力データ非学習条項

「提供者は、顧客入力、添付ファイル、プロンプト、出力およびこれらから得られる情報を、基盤モデルその他のAIモデルの学習、微調整、評価又は改善に使用しないものとする。」

保持・削除条項

「顧客データは本サービス提供に必要な最小限の期間のみ保持し、契約終了又は顧客の求めに応じて、法令上保存が必要な場合を除き削除する。保持期間、保存国、subprocessorを別紙に明示する。」

知的財産条項

「顧客は入力に係る権利を保持し、提供者が出力について取得する一切の譲渡可能な権利を顧客に譲渡する。提供者は第三者から知的財産権侵害の申立てを受けた場合、合理的な範囲で調査および防御に協力する。」

モデル変更条項

「基盤モデル、重み、推論設定、データ保持条件又は利用規約に、出力品質・情報管理・知的財産リスクへ重大な影響を与える変更を行う場合、提供者は事前通知する。」

DeepSeek標準規約にはすでに一定の出力権利譲渡が存在するが、**第三者権利の非侵害保証・十分なIP indemnity・企業向けno-training保証を標準規約から当然に得られるわけではないため、重要案件では個別契約が望ましい。** 55

推奨する社内PoC

1017を知財業務へ採用するかは、一般ベンチではなく、社内の正解データを使って測るべきである。特に公式と独立Terminal-Benchの差が大きいことから、モデルランキングだけで導入判断するのは危険である。

56

推奨評価セットは、例えば既知の正解を持つ過去案件100～300件について、**先行技術検索Recall@20／Precision@20、重要請求項の要素抽出F1、審査経過の事実誤り率、架空公報番号率、引用文献の存在確認率、日本語→英語特許翻訳の用語一致率、契約条項リスク検出率、P50/P95レイテンシ、1案件当たり総コスト、人手削減時間**を測定する方式である。この部分は公開1017ベンチではなく、知財実務に合わせた本報告の推奨評価設計である。

特に重要なのは、「**正答率**」だけでなく「**見逃し率**」を測ることである。特許先行技術調査やFTOでは、もっともらしい回答を生成する能力より、関連文献を漏らさない能力が重要だからである。1017についてこの種の特許IRベンチは現時点で未公開であり、ここが知財利用可否を決める最大の未検証点である。

総合判断と主要出典

知財部門から見た評価

2026年10月18日時点の総合評価は、「**技術的には極めて有望、コスト競争力は高い。しかし知財の最終判断用途にはまだ証拠不足**」である。

性能面では、V4系からV4.1-1017への進歩は明白で、とりわけagentic coding、automation、長文・画像処理、KVキャッシュ効率等は知財業務の自動化と相性がよい。公式DeepSWE 77.0など、一部では高価な閉源モデルを上回る結果もある。 57

一方、公式42.3に対し独立Terminal-Benchが26.8という差、1017専用の実運用latencyデータ不足、特許検索ベンチ不存在、公開翌日ゆえのレビュー母数不足から、性能を過大評価してはならない。 23

知財リスクでは、MITライセンスは大きな利点だが、「**MITだから企業知財にも安全**」という結論にはならない。ホステッドサービスでは利用者が入力権利を保証し、DeepSeekは出力の非侵害性を保証せず、データはモデル改善に利用され得て、中国サーバーへの保存も開示されている。未公開発明・営業秘密を扱う部門では、APIを低機密用途に限定するか、オープンウェイトの利点を生かした自己ホスト／閉域運用を評価する価値が高い。 58

したがって、現時点の導入判断は次のように整理できる。

公開情報を使った特許・技術調査支援：条件付きで積極採用候補。

契約・公報・コードの大量一次レビュー：Human-in-the-loop前提で有力。

未公開発明・営業秘密を標準APIに投入：原則避ける。

特許性、FTO、無効性、商標登録可否の最終判断：不可。

自己ホスト：セキュリティ上有力だが763B級の運用TCOを必ず実測。 9月版V4.1では4×H200環境でも自己ホストがAPIより割高になった公開事例がある。 25

公開情報上の主な不明点

不明点	2026年10月18日時点
1017固有の新アーキテクチャ変更	未公開／発表なし。 V4.1ライフサイクル内更新とみられる。 9
1017固有のRL訓練データ件数・全出典	未公開
1017の追加学習compute/GPU時間	未公開
1017公式tokens/sec	未公開
TTFT、P50/P95/P99 latency	未公開
GPU別必要VRAMの1017公式表	未公開
日本語特許検索のRecall/Precision	未公開
日本語法務・知財専用benchmark	未公開
1017固有の査読論文	確認できず。 公開翌日であり、V4.1技術報告が主な学術一次資料。 26
企業の長期本番運用事例	1017固有では確認困難
学習コンテンツの個別権利者・全URL一覧	未公開。 カテゴリレベルの透明性開示のみ。 59

主要出典

DeepSeek公式リリース／Change Log — 2026年10月17日の1017公開、公式benchmark、旧版との差分、API互換性を確認する一次資料。
22

DeepSeek-V4.1-Flash-1017 Hugging Face Model Card — モデル規模、MIT License、推奨設定、競合比較、1017の位置付けを確認する主要一次資料。
21

DeepSeek-V4.1-Flash Technical Report (arXiv) — CED、MoE、CSA2、FP4 KV cache、Engram、45T-token training等の技術的根拠。1017専用論文ではなくV4.1ベースの開発者技術報告である。
60

Public Summary of Training Content for DeepSeek-V4.1 — 2026年9月24日更新。10兆超text tokens、10億超images、Common Crawl、Stack Exchange、licensed data、crawler、user interaction、TDM opt-out等を開示。
61

DeepSeek Terms of Use — 2026年3月27日更新。入力権利保証、出力権利譲渡、非侵害・正確性の非保証、AI表示義務、責任制限、中国法準拠を確認。
62

DeepSeek Privacy Policy 日本語版 — 2026年2月10日更新。入力収集、モデル改善、training opt-out、中国サーバー保存、検索keywordの第三者API共有等を確認。
47

Artificial Analysis / OpenRouter — 1017固有の主要な独立benchmark。公式値との差を確認するうえで重要。
3

文化庁「AIと著作権に関する考え方について」 — 日本におけるAI学習と著作権法30条の4、類似性・依拠性、AI生成物の創作意図・創作的寄与に関する中心的な公的資料。 63

特許庁「AI技術の発達を踏まえた特許制度上の適切な対応」 — AI利用発明でも自然人の発明者を認定する日本の制度整理に関する一次資料。 43

経済産業省／AI事業者ガイドライン第1.2版 — 2026年3月に更新された日本企業のAIガバナンス上の基準資料。 64

PC Watch（2026年10月18日） — 1017の日本語圏での初期技術報道。公式ベンチの改善と競合との差を日本語で確認できる。 27

Tom's Hardware掲載のV4.1自己ホスト企業試験 — 1017そのものではないが、4×H200環境での実運用 throughput、cache reread、自己ホストTCOという、公式benchmarkでは見えない運用上の論点を示す。 25

1 9 11 17 21 41 51 57 deepseek-ai/DeepSeek-V4.1-Flash-1017 · Hugging Face
<https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash-1017>

2 6 7 22 56 Change Log | DeepSeek API Docs
<https://api-docs.deepseek.com/updates/>

3 23 33 34 DeepSeek V4.1 Flash 1017 - API Pricing & Benchmarks | OpenRouter
<https://openrouter.ai/deepseek/deepseek-v4.1-flash-1017>

4 10 13 32 DeepSeek-V4.1-Flash: Pushing the Limits of KV Cache Compression
<https://arxiv.org/html/2609.19969v1>

5 18 19 20 36 42 45 46 52 54 55 58 62 DeepSeek Terms of Use
<https://cdn.deepseek.com/policies/en-US/deepseek-terms-of-use.html>

8 24 DeepSeek-V4.1-Flash-1017: Wish-Driven Programming Made Real | DeepSeek API Docs
<https://api-docs.deepseek.com/news/news261017/>

12 26 31 60 [2609.19969] DeepSeek-V4.1-Flash: Pushing the Limits of KV Cache Compression
<https://arxiv.org/abs/2609.19969>

14 15 16 40 59 61 cdn.deepseek.com
<https://cdn.deepseek.com/policies/DeepSeek-V4.1-public-summary-of-training-content.pdf>

25 Firm rents four Nvidia H200s to test '80x cheaper' DeepSeek claim
https://www.tomshardware.com/tech-industry/artificial-intelligence/firm-rents-four-nvidia-h200s-to-test-80x-cheaper-deepseek-claim-usd13-200-monthly-gpu-rental-doubles-claude-bill-while-security-flaws-keep-code-offline?utm_source=chatgpt.com

27 コード系でOpus 5超え「DeepSeek-V4.1-Flash-1017」無償公開
https://pc.watch.impress.co.jp/docs/news/2149045.html?utm_source=chatgpt.com

28 China's DeepSeek launches V4.1-Flash model
https://www.reuters.com/world/asia-pacific/chinas-deepseek-launches-v41-flash-model-2026-09-10/?utm_source=chatgpt.com

29 まさお@AI駆動開発
https://x.com/AI_masaou/status/2111406145311285471?utm_source=chatgpt.com

30 Deepseek v4 flashを、大きめのコード変更の評価でテストした。 ...

https://www.reddit.com/r/LocalLLaMA/comments/1suhdki/tested_deepseek_v4_flash_with_some_large_code/?tl=ja&utm_source=chatgpt.com

35 49 Model Mechanism and Training Methods of DeepSeek

<https://cdn.deepseek.com/policies/en-US/model-algorithm-disclosure.html>

37 38 39 48 bunka.go.jp

https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/pdf/94037901_01.pdf

43 AI技術の発達を踏まえた特許制度上の適切な対応

https://www.jpo.go.jp/resources/shingikai/sangyo-kouzou/shousai/tokkyo_shoi/document/51-shiryuu/02.pdf?utm_source=chatgpt.com

44 47 50 53 DeepSeekプライバシーポリシー

<https://cdn.deepseek.com/policies/ja-JP/deepseek-privacy-policy.html>

63 AIと著作権について

https://www.bunka.go.jp/seisaku/chosakuken/aiandcopyright.html?utm_source=chatgpt.com

64 AI事業者ガイドライン

https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20260331_report.html?utm_source=chatgpt.com