

DeepSeek-V4.1-Flash-1017の性能解析・市場評価および知的財産業務への戦略的影響に関する調査報告

Gemini 4 Argon

2026年10月17日、中国のAI研究開発企業であるDeepSeek(杭州深度求索人工智能基礎技術研究有限公司)は、同社のマルチモーダル軽量推論モデル「DeepSeek-V4.1-Flash」の最新アップデート版となる「DeepSeek-V4.1-Flash-1017」(以下、1017版)を公開した¹。前版(2026年9月10日公開の0910版)のリリースからわずか37日後に投入されたこの新チェックポイントは、アーキテクチャの骨格を維持したマイナーアップデートと位置づけられているものの、自己フィードバック型強化学習(Reinforcement Learning with Self-Feedback)の導入と出力トークン長の抜本的な短縮設計により、ソフトウェア開発および自律エージェント領域において劇的なコストパフォーマンスの改善(パレート改善)を達成している¹。さらに同日、DeepSeekが独自構築した検索基盤を外部開放する「Search API」および「Fetch API」が全ユーザー向けに一般公開されたことで、モデルの推論能力と外部情報取得を直結させた自律型調査パイプラインの構築コストが歴史的な低水準へと引き下げられた²。本報告では、1017版のアーキテクチャ上の技術的特異性とベンチマーク性能を精査した上で、開発者コミュニティおよび第三者検証機関による実機評価・デグレード(性能後退)リスクを分析し、最後にこれらの技術的・法的特性が企業や特許事務所の知的財産業務(先行技術調査、図面解析、明細書作成、機密管理)に及ぼす構造的影響を包括的に論じる。

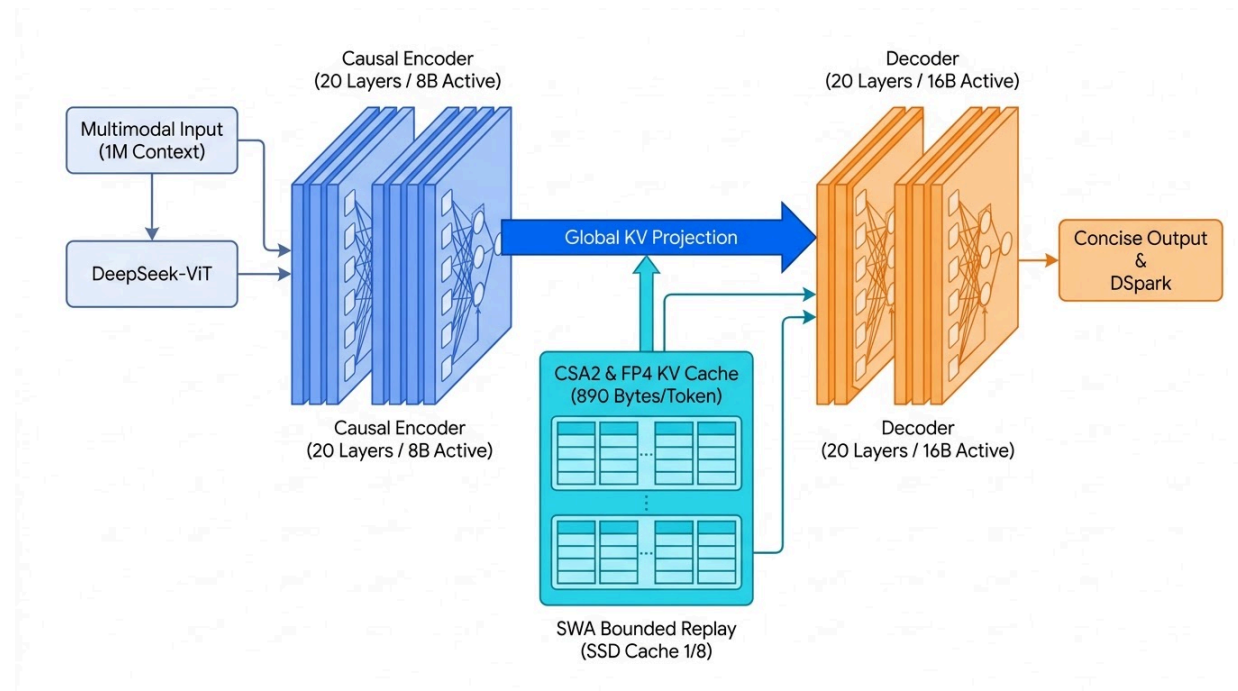
1. アーキテクチャの技術的特異性と1017版の性能プロファイル

非対称CED構造と超圧縮KVキャッシュによる長文脈処理の経済性

1017版が継承する「DeepSeek-V4.1-Flash」系列の最大の技術的特徴は、単なるパラメータ数の削減ではなく、長文脈エージェント運用において最大のボトルネックとなる物理メモリ(GPU高帯域幅メモリ:HBM)の保持コストを極限まで圧縮した点にある⁶。従来のデコーダ専用(Decoder-only)トランスフォーマーでは、入力コンテキストが長大化するにつれて全層でキー・バリュー(KV)キャッシュを保持する必要があり、100万トークン級の処理では演算能力以上にメモリ帯域と容量がコストを押し上げる要因となっていた⁶。

これに対し、本モデル群は40層のトランスフォーマーを20層の因果エンコーダ(Causal Encoder)と20層のデコーダ(Decoder)に分割した「Causal Encoder-Decoder(CED)」構造を採用している⁶。デコーダ側のグローバルKVキャッシュを各デコーダ層の隠れ状態から個別に生成するのではなく、エンコーダ最終層の隠れ状態から射影する設計とすることで、長大な入力を読み込むプリフィル(Prefill)段階の計算層を実質的に半減させた⁶。モデル本体の総パラメータ数は5,520億(552B)の疎な混合エキスパート(Sparse MoE)構成であり、これに1,960億(196B)パラメータの条件付きメモリ機構「Engram」および約140億(14B)の投機的デコーディング機構「DSpark」、約5億(0.5B)の画像処理モジュールが組み合わされている(Hugging Face上の表示サイズは計763B)²。特筆すべきは入出力の非対称性であり、1トークンあたりの活性パラメータ数は入力読み込み時(プリフィル)がわず

か80億(8B)、出力生成時(デコード)が160億(16B)に抑えられている²。この非対称設計により、大量の文書やコードベースを入力して短いアクションや回答を出力するエージェント型ワークロードにおいて、計算資源の消費が劇的に効率化されている⁶。



さらにメモリ圧縮の中核として「Compressed Sparse Attention 2 (CSA2)」と4ビット浮動小数点形式 (FP4 E2M1)による量子化保存が導入されている⁶。CSA2では各アテンション層にFull (KVデータと参照位置を新規計算)、Reindex (先行層のKVデータを再利用しつつ参照位置を再選択)、Reuse (KVデータと参照位置の双方を先行層から継承)の3つの固定動作モードを割り当て、層間での情報重複を排除している⁶。加えて、長時間稼働するエージェントがセッション状態をSSDやホストメモリへ退避・再開する際の永続キャッシュには「SWA Bounded Replay」を採用し、スライディングウィンドウのKV状態をすべて保存する代わりに再開時に直近ウィンドウのみを再計算して状態を近似再構築する⁶。これらの相乗効果により、HBM上のグローバルKVキャッシュは1トークンあたり890バイトへと縮小され、前世代のDeepSeek-V4-Flash (3,514バイト)の約4分の1、DeepSeek-V3.2 (48,068バイト)の約54分の1、初代DeepSeek-V1 (389,120バイト)との比較では実に約437分の1という驚異的な圧縮率を実現した²。永続KVキャッシュも前世代比で約8分の1に削減されており、同一ハードウェア上で同時処理可能なセッション数を飛躍的に拡大させている⁶。

視覚理解(マルチモーダル)機能においても、既存の言語モデルに後から視覚エンコーダを接続する外付け方式ではなく、2次元回転位置埋め込み (2D-RoPE)と3×3ピクセルアンシャッフルによるダウンサンプリングを備えた32層の「DeepSeek-ViT」および2層MLPプロジェクタをスクラッチから構築し、45兆トークンのマルチモーダルコーパスを用いて事前学習の初期段階からテキストと画像を統合学習させている⁶。

仕様項目	DeepSeek-V4.1-Flash-1017 (10月17日)	DeepSeek-V4-Pro (0813 GA版・参考)	備考・実務上の影響
------	-----------------------------------	-------------------------------	-----------

	版)	比較)	
APIモデルID	deepseek-flash(変更なし・自動置換)	deepseek-v4-pro	旧名 deepseek-v4-flash 等も自動で1017版へ転送 ²
総パラメータ数	552B(MoE)+196B(Engram)※HF表示763B	1.6T(1.6兆・MoE)	1017版はDSpark(約14B)と視覚系(約0.5B)を包含 ²
活性パラメータ数	入力処理時 8B / 出力生成時 16B	49B	入力が長大なエージェント・文書処理で圧倒的に有利 ²
コンテキスト長 / 最大出力	100万(1M)トークン / 最大 384Kトークン	100万(1M)トークン / 最大 384Kトークン	既定最大出力は非思考時8K、思考時64K(max設定時128K) ²
グローバルKVキャッシュ	890 bytes / token(HBM常駐分)	非公開(V4-Flashは3,514 bytes)	永続KVキャッシュ(SSD/ホスト)も前世代比約1/8に圧縮 ²
マルチモーダル対応	テキスト・画像入力 / テキスト出力	テキスト入出力のみ(画像非対応)	Flash系のみがネイティブ画像理解(DeepSeek-ViT)を搭載 ²
API同時実行数上限	2,500	500	大規模バッチ処理時の並列スループットがProの5倍 ²
公開ライセンス	MIT License(商用利用・改変・再配布可)	MIT License	Hugging FaceおよびModelScopeにて重みを無償公開 ²

「許願式プログラミング」と出力短縮によるパレート改善

2026年9月10日に公開された前版(0910版)は、高い推論速度(Artificial Analysis計測で每秒214.4~217出力トークン)を記録した一方で、回答完了までに極めて冗長な思考プロセスと出力テキストを生成する「長文癖」が実務上の課題として浮上していた¹。第三者評価機関であるArtificial Analysisの計測によれば、0910版がIntelligence Indexの全課題を完走するまでに生成した出力トークン数

は2億5,000万トークンに達し、全モデル中央値(1億4,000万トークン)の約1.8倍に膨れ上がっていた¹⁵。大規模言語モデルのAPI課金体系において出力単価は入力単価の4倍(キャッシュミス比)から200倍(キャッシュヒット比)に設定されているため、この冗長性はそのまま利用企業の請求額の増大と実作業の待ち時間悪化に直結していた¹。

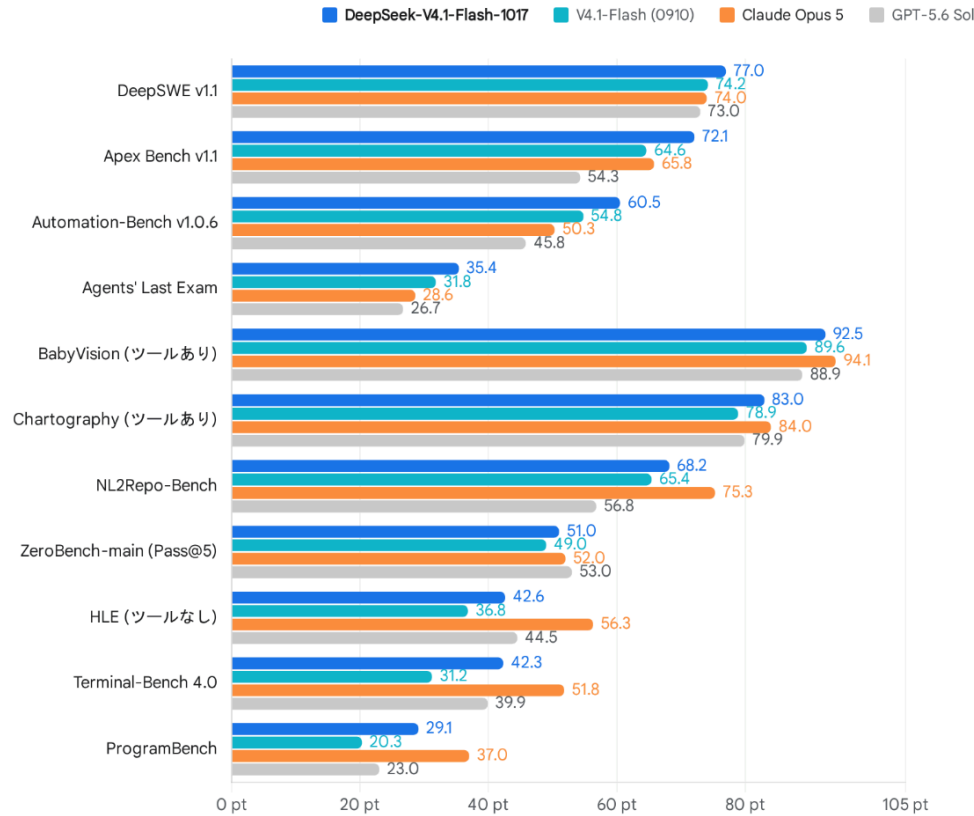
1017版は、この課題に対して「自己フィードバックを用いた強化学習(self-feedback RL)」を適用することで、推論の質を上げながら不要な長文出力を削ぎ落とす調整を断行した¹。公式発表のタイトルである「Wish-Driven Programming Made Real(中国語原題:許願式プログラミング=願いを言うだけで実現するプログラミング)」が示す通り、ユーザーが意図を伝えるだけで、紫禁城の3Dモデリング、ブラウザで動作するレースゲーム、Godotエンジンを用いた『Portal 2』の再現、人類の進化を辿る複雑な3Dシーン「Valley of Ages」といったプロジェクトを最初から最後まで一気通貫で構築するエンドツーエンドの実装力が強化されている²。

エージェント性能でOpus 5を凌駕し1タスク単価を劇的に圧縮

第1部：公式ベンチマーク11項目 スコア比較

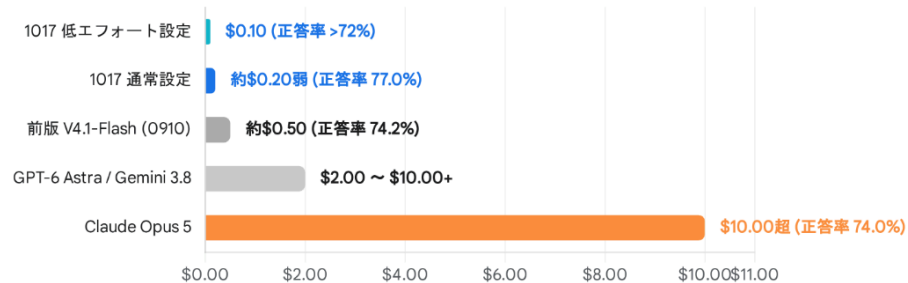
ワークフロー・自動化4項目で1017が先行、高度知識推論7項目でOpus 5が優位

1017優位 (4) Opus 5優位 (7)



第2部：DeepSWE v1.1 における1問あたりの実行コスト比較 (USD)

出力トークン短縮により前版比約60%削減（\$0.50→約\$0.20弱）。フロンティア級最高精度を最安水準で実現



上段は公式ベンチマーク11項目におけるモデル別スコア比較。1017版は全11項目で前版（O910）を上回り、ワークフロー自動化やエージェント課題の4項目でClaude Opus 5をリードする一方、高度な知識推論（HLE）や大規模リボトリ構築（NL2Repo-Bench等）の7項目ではOpus 5が優位を維持している。下段はDeepSWE v1.1における1タスクあたりの実行コスト比較であり、出力短縮によって前版の約0.50ドルから約0.20ドル弱へと半減し、他社フロンティアモデル（2~10ドル超）に対して圧倒的な低コストを実現している。

Data sources: [AI Revolution](#), [note.com](#)

公開された全11項目のベンチマークにおいて、1017版は前版（0910版）を例外なく上回っている²。特にClaude Code環境で測定されたTerminal-Bench 4.0では31.2から42.3へと+11.1ポイントの急伸を見せたほか、ProgramBenchで+8.8ポイント（20.3→29.1）、Apex Bench v1.1で+7.5ポイント（64.6→72.1）、ツールなしの学術難問ベンチマークHLE（Humanity's Last Exam）で+5.8ポイント（36.8→42.6、テキスト限定問題では39.1→44.0）、Automation-Bench v1.0.6で+5.7ポイント（54.8→60.5）の改善を記録した²。また、競技プログラミングのCodeforcesレーティングでは3471に到達し、自社の最上位フラッグシップであるDeepSeek-V4-Pro（3348）すら凌駕している¹⁷。

ベンチ マーク 項目	V4.1-F lash-1 017	V4.1-F lash (0910)	前版比 差分	Claud e Opus 5	GPT-5 .6 Sol	Kimi K3	GLM- 5.3	測定環 境・領 域特性
Deep SWE v1.1	77.0	74.2	+2.8	74.0	73.0	67.5	66.9	Mini-S WE-V2 環境（ 1017首 位） ²
Auto matio n-Ben ch v1.0.6	60.5	54.8	+5.7	50.3	45.8	46.7	48.8	ワーク フロー 自動化 （1017 首位） ²
Agent s' Last Exam	35.4	31.8	+3.6	28.6	26.7	27.6	28.5	複合 エー ジェン ト課題 （1017 首位） ²
Apex Bench v1.1	72.1	64.6	+7.5	65.8	54.3	47.8	56.6	エー ジェン ト実行 力（ 1017首 位） ²
HLE （ツ－	42.6	36.8	+5.8	56.3	44.5	43.5	42.0*	専門学 術知識

ルなし)								推論(Opus 5首位) ²
Terminal-Bench 4.0	42.3	31.2	+11.1	51.8	39.9	12.6	37.9	Claude Code 環境(Opus 5首位) ²
ProgramBench	29.1	20.3	+8.8	37.0	23.0	17.5	19.0	大規模プログラム構築(Opus 5首位) ²
NL2Repo-Bench	68.2	65.4	+2.8	75.3	56.8	58.0	58.0	自然言語からのリポジトリ生成 ²
Chartography (ツール有)	83.0	78.9	+4.1	84.0	79.9	68.1	—	視覚グラフ・図表解析 ²
BabyVision (ツール有)	92.5	89.6	+2.9	94.1	88.9	85.7	—	基礎視覚認識・ツール連携 ²
ZeroBench-main	51.0	49.0	+2.0	52.0	53.0	41.0	—	Pass@5計測(

(ツール有)								GPT-5 .6 Sol 首位) ²
--------	--	--	--	--	--	--	--	-------------------------------

注: GLM-5.3のHLEスコアはテキスト問題のみのサブセット計測値²。
 しかし、メディア等で散見される「Claude Opus 5超え」という標語については冷静な因数分解が求められる²。公式比較表の11項目を精査すると、1017版がClaude Opus 5を上回っているのはDeepSWE v1.1(77.0対74.0)、Automation-Bench v1.0.6(60.5対50.3)、Agents' Last Exam(35.4対28.6)、Apex Bench v1.1(72.1対65.8)の4項目(主に自律エージェント動作およびワークフロー自動化)であり、残る7項目では依然として**Claude Opus 5**が優位に立っている²。とりわけ、外部ツールに頼らない純粋な専門知識と論理推論を測るHLE(42.6対56.3)や、ゼロから大規模なリポジトリ構造を設計・構築するNL2Repo-Bench(68.2対75.3)およびProgramBench(29.1対37.0)ではOpus 5に明確なアドバンテージがある²。すなわち、1017版は「あらゆる知能面で最上位モデルを代替するモデル」ではなく、「定型～中難易度のコーディング改修やツール呼び出しを伴うエージェント実行において、フロンティアモデルと同等以上の成果を桁違いの低コストで叩き出す実務特化型モデル」と定義するのが正確である²。

API料金体系とSearch / Fetch APIの同時開放によるエコシステム拡張

1017版のAPI単価は前版から据え置かれており、ピーク時間帯とオフピーク(閑散)時間帯で2倍の価格差を設けるダイナミックプライシングが継続されている²。ピーク時間帯は月曜から金曜のUTC 01:00～04:00および06:00～10:00(中国の祝日を除く)と定められており、日本時間(JST)に換算すると平日の**10:00～13:00**および**15:00～19:00**という日本のコア業務時間帯に直撃する²。裏を返せば、昼休み時間帯(13:00～15:00)、夜間(19:00～翌10:00)、週末、および中国の祝日にバッチ処理をスケジュールするだけで、すべてのトークン単価が自動的に半額となる²。

モデル・API 区分	時間帯区分 (日本時間基準)	入力: キャッシュヒット	入力: キャッシュミス	出力単価	課金・運用上の特記事項
deepseek-flash (V4.1-Flash-1017)	オフピーク (平日夜間・昼13-15時・土日・中国祝日)	\$0.003 / 1M tok	\$0.15 / 1M tok	\$0.60 / 1M tok	単価はOpus 5の十数分の一～四十分の一水準 ²
deepseek-flash (V4.1-Flash-1017)	ピーク(平日 10:00-13:00, 15:00-19:00 JST)	\$0.006 / 1M tok	\$0.30 / 1M tok	\$1.20 / 1M tok	出力短縮により1タスク実質コストは約0.10～0.20弱 ²

deepseek-v4-pro (V4-Pro-0813)	オフピーク ／ピーク	\$0.022 / \$0.044	\$0.66 / \$1.32	\$1.98 / \$3.96	100万入力 +10万出力 時、Flashは Proの約4分 の1の費用 ²
Search API (POST /v1/tools/search)	全時間帯共 通(成功リク エスト課金)	—	—	\$0.75 / 1,000 req	HTTP 200か つ非空結果 のみ課金。 上限5 QPS/ アカウント ⁴
Fetch API (POST /v1/tools/fetch)	全時間帯共 通(成功リク エスト課金)	—	—	\$0.75 / 1,000 req	Markdown 形式で本文 取得。2026 年11月1日 1:00 JSTま で一部無料 ²

この単価構造に1017版の「出力トークン短縮」が組み合わさったことで、タスク単位の実行経済性は劇的に向上した¹。DeepSWE v1.1における1タスクあたりの平均コストは、前版(0910版・正答率74.2%)の約0.50ドルから、1017版(正答率77.0%)では約**0.20ドル**弱へと**60%以上**削減されており、推論エフォートを抑えた設定(正答率72%超)では1タスク約**0.10ドル**で完結する²。競合するGPT-6 Astra、Gemini 3.8 Flash、Claude Opus 5が同一タスクに1件あたり2ドル～10ドル超を要することと比較すれば、10分の1から50分の1以下のコストで同等以上のコード改修タスクを完遂できる計算となる²。

さらに実務上の重大な転換点となるのが、1017版と同時に一般開放された**Search API**および**Fetch API**である²。Search APIはDeepSeekが自社開発し世界中で毎日数億件のリクエストを処理している独自検索エンジンを基盤としており、数千億ページのインデックスからSEOスパムや低品質なAI生成コンテンツを排除した構造化検索結果を返す²。クエリにはsite:、-site:、lang:、after:、before:、完全一致"..."などの高度な検索演算子を複数組み合わせることが可能であり、レスポンスには関連度・権威性・鮮度を統合したrerank_score(0～3、1.5以上が高品質)やドメインの権威性を示すauthority(0～4、3以上が高権威)、UI表示用のsnippet(100～200文字)、LLM読解用のsummary(600～800文字)、さらにはオプション(need_content: true)でページ全文を含めることができる⁴。一方のFetch APIは、指定したURLのウェブページ本文を抽出し、LLMが直接解釈しやすいMarkdownテキスト形式で返却する⁴。両APIとも成功したリクエスト(HTTP 200かつ結果・本文が非空の場合のみ)1,000件あたり0.75ドルという低価格で提供され、返却件数(count: 最大50件)を増やしても1リクエストあたりの料金は変動しない(なお、DeepSeek Harness、Claude Code、Codexのサーバー側ツール経由での呼び出しは2026年11月1日01:00 JSTまで無料キャンペーンが適用されている)²。これにより、企業は自前でクローラーやスクレイピング基盤を保守したり高価な外部検索APIを別途契約したりすることなく、単一のDeepSeek APIキーのみで高度なWeb検索・RAG(検索拡張生成)エージェントを構築できるようになった⁵。

2. 市場の評判・実機評価と運用上の構造的リスク

実機検証で判明した「挙動の入れ替わり(デグレード)」とハーネス依存性

公開直後から開発者コミュニティやAIプラットフォームでは1017版への移行が急速に進んでおり、OpenCode上ではV4.1 Flash系列が全体トークン消費量の第1位(過去2か月間の全トークンの25%)を占めるなど圧倒的な支持を集めている¹⁹。しかし、エンジニアによる詳細な実機検証からは、公式ベンチマークの平均値だけでは見えない「タスクごとの挙動の激しい入れ替わり」が報告されている¹⁸。

AIエージェントのCLIに対して25問の実践的な開発課題を与え、生成された成果物をブラウザ上で実際に動作させて採点した独立検証では、前版(0910版)と1017版ともに全25問を完走し総合平均スコアもほぼ同等であったにもかかわらず、個別の成果物では「前版で壊れていたものが劇的に動くようになった事例」と「前版で正常に動いていたものが動かなくなった事例(デグレード)」が同時に多数発生した¹⁸。具体的には、前版では敵キャラクターの生成処理でエラーが起き戦闘が開始されなかった3Dゲーム「ゴブリンFPS」や、初期描画でフリーズしていた「ピンボール」が、1017版では視点の傾きまでリアルに再現された遊べるゲームとして完成し、SVGアニメーション(「歩く人形」)の視覚的品質も別次元へ向上した¹⁸。ところがその反面、前版では図形の描画から変形まで正常に機能していたベクター図形エディタ「簡易Figma」が、1017版ではUI部品や内部設計こそ綺麗に整っているものの、マウス操作イベントがキャンバスに一切バインドされておらず、エディタとして全く機能しない状態へと劣化した¹⁸。また、ピンボールにおいても球がランプを無限に周回して抜け出せなくなるロジック上の欠陥が残存していた¹⁸。

この現象は、自己フィードバック強化学習と出力短縮の最適化プロセスにおいて、モデル内部の重み付けが「コード全体の構造的な見栄えやコンパクトな完結性」へ強く振られた結果、一部の細かいイベントハンドリングやエッジケースの論理結合が省略・変質したことを示唆している¹。さらに、モデルの性能がエージェントの実行基盤(ハーネス/スキャフォールド)に極度に依存する点も見逃せない⁷。同一のV4.1-Flashモデルであっても、ツールループやシステムプロンプト、コンテキスト管理を担うスキャフォールドを変更するだけで、DeepSWE v1.1のスコアは65.5から74.2へと8.7ポイント、Terminal-Bench 2.1では84.1から90.6へと6.5ポイントも変動することが確認されている⁷。1017版は自社の「DeepSeek Harness(v0.2.1-rc.1以降推奨)」やテンセントの「CodeBuddy」「WorkBuddy」向けに学習段階から最適化されており、Claude CodeをハーネスとするTerminal-Bench 4.0では依然としてClaude Opus 5に大差(42.3対51.8)をつけられている²。

旧モデルIDのサイレント置換問題とオンプレミス運用の技術的障壁

企業のシステム運用担当者にとって最も警戒すべき仕様は、API上のモデルID(deepseek-flash)が同一のまま、10月17日をもってバックエンドが1017版へ強制的に自動切り替えされた点である²。公式APIには前版(0910版)や旧世代モデル(すでに退役したdeepseek-v4-flashやdeepseek-v4-flash-vision-exp)を指定・固定するバージョンピンニング機能が用意されておらず、旧モデル名でリクエストを送信してもエラー(HTTP 4xx/5xx)は一切発生せず、すべて1017版が応答しFlash料金で課金される²。

APIがエラーで停止すれば監視アラートが発報されるが、ステータスコード200で正常応答し続けながら「出力文字数が大幅に短くなる」「特定の処理ロジックが入れ替わる」というサイレント変更は、出力長を前提としたUIレイアウトや正規表現・JSONパーサー、後続のワークフロー分岐を無言のうちに破壊するリスクを孕む²。さらに、APIでは既定で思考モード(thinking: enabled)が有効化されてい

るため、非思考モードを前提としていた旧システムからの移行時には想定外の思考トークン消費が生じる場合もある⁴。万が一、0910版の挙動を厳密に維持する必要がある場合は、Hugging Faceに残されている0910版のオープンウェイトを取得して自社インフラでホストする以外に回避策がない²。しかし、MITライセンスで公開されている重みを用いた自社運用（オンプレミス／プライベートクラウド展開）にも高い技術的ハードルが存在する²。活性パラメータこそ8B/16Bと軽量であるものの、モデル全体の重み（552B MoE＋196B Engram、表示上763B）をメモリ上に展開するためには、NVIDIA H100/H200/B200やAMD MI300Xといったデータセンター級GPUを複数基束ねた大規模クラスタが不可欠となる²。加えて、公開されているモデルリポジトリには標準的なJinja形式のチャットテンプレートが同梱されていないという実務上の落とし穴がある²。一般的な推論エンジンにそのままロードすると対話フォーマットが崩れて出力が破綻する恐れがあるため、公式が提供するPython参照実装（encoding/encoding.py）または本番用のRust製ライブラリ「deepseek-recipe」を組み込んでvLLMやSGLang上で稼働させる高度なMLOpsエンジニアリングが要求される²。

ハルシネーション傾向・長文脈の近似劣化・日本語処理の特性

基礎的な言語・知識性能の観点からは、事実知識QAにおけるハルシネーション（虚偽生成）リスク、長文脈における収容力と精読力の乖離、および日本語性能の特有の偏りという3つの構造的弱点が確認されている⁷。まず事実性の側面では、エージェント系ベンチマークが飛躍的に向上した一方で、モデルの純粋な事実記憶と正確性を測るSimpleQA-VerifiedにおいてV4.1-Flashは42.3にとどまり、旧世代の上位モデルであるV4-Pro（55.2）から12.9ポイントも低下している⁷。Artificial Analysisの包括評価（Intelligence Index 350モデル中30位）においても、AA-Omniscience Non-Hallucination Rate（非ハルシネーション率）が3.5%と低迷している点が浮き彫りになっており、パラメトリック知識（モデル内部の記憶）のみに頼った事実確認では虚偽情報を自信ありげに出力するリスクが高い⁷。次に長文処理の側面では、アーキテクチャ上こそ100万トークンの入力に対応しているものの、長文理解力を測るLongBench-V2のスコアは45.2であり、V4-Pro（51.5）を下回っている⁷。技術報告書でも言及されている通り、CSA2における層間の参照位置共有は特定のトークン位置への選択バイアスを生む可能性があり、SWA Bounded Replayによる近似的な状態再構築も「遠い過去の文脈に埋もれたごく僅かな限定事実」を正確に掘り起こす処理（Needle in a Haystack）において精度劣化を引き起こす懸念がある⁶。すなわち、100万トークンをメモリに収容できることと、その細部を漏れなく精密に読解できることは同義ではない⁷。

さらに日本語処理の特性について、Qualitegの「日本語対応LLMランキング2026」等の評価では、DeepSeek V4系は実務利用に耐える自然な日本語能力を備え、とりわけPro系よりFlash系の方が日本語スコアが高いという特徴を持つ（ただしClaude系の最上位層とは0.05ポイント前後の差が存在する）¹。より詳細な日本語ベンチマーク群（tengu_bench、elyza、ja-mt-bench、rakuda）の傾向分析からは、DeepSeek系モデルは版を重ねても一貫して「指示追従性（elyza、tengu）で全体首位級の強さを発揮する一方、日本固有の知識（rakuda）や自然な対話品質（ja-mt-bench）が弱い」という明確な性格を維持していることが判明している²⁰。1017版ではこれに出力を短く切り詰める調整が加わったため、日本語での長大なレポートや論文の執筆には不向きであり、要約、データ抽出、構造化変換、コード生成といった用途に高い適性を持つ¹。

3. 知的財産業務へ与える構造的影響と導入戦略

1017版が備える「100万トークンの超長文脈」「ネイティブ画像理解」「桁違いの低価格と高い同時実行数（2,500）」「Search/Fetch APIの統合」という4つの武器と、その裏返しである「中国の国家情報

法リスク・守秘義務条項の欠如」「ハルシネーションと長文近似劣化」「出力短縮の癖」は、企業知財部および特許事務所の業務プロセスに対して極めて非対称なインパクトをもたらす¹。

知財業務プロセス別の1017版適用適性と推奨ルーティング構成

最大コンテキスト長 1,000,000 tokens 最大特許公報入こと連携可能	最先API準拠 (オフピーク) \$0.003 /MToK Cache Hit時 (平日夜間・週末)	タスク実行コスト目安 <\$0.20 /請求 知財特許照会・マークアップ等	公式API中核業務準拠 0条項 (なし) 非公開特許のAPI連携保証
---	---	---	---

すべて表示 (5条項)	公開情報ワークフロー (3)	特許・実務ワークフロー (2)	公式API準拠 (緑)	ハイブリッド準拠 (オレンジ)	公式API準拠 (赤)
-------------	----------------	-----------------	-------------	-----------------	-------------

公開情報領域・高ROI

1. 公開特許の大量一次スクリーニング・IPランドスケープ

対象データ

既発行の特許公報・公開公報

速度/費用効率: 5.0 / 5.0 (緑)

OT: peak: \$0.003/M Cache Hit

技術的留意点・品質リスク

1017版の出力短縮調整により分幅フラグ生成速度が最大化。ただし、TMコンテキスト対応の一方で長文理解 (LongBench-V2 45.2) やCSA2 追加に伴う複雑な実装・企業クレームの認識見落としに注意。一次スクリーニング向け。

法的・ガバナンス評価

対象公報はすべて公開情報のため、公式API利用による営業秘密喪失や新規性喪失のリスクは極小。

推奨システム構成

公式API [夜間・週末バッチ]

平日中ピーク時を避け、夜間や週末のオフピーク時 (\$0.10/M) は適宜入力。\$0.003/Mの低コストで大量公報を効率的にバッチ処理。最大50%のAPIコスト削減を要約。

公開情報領域・自働エージェント

2. 非特許文献(NPL)・技術動向の自律Web調査

対象データ

公開論文・学術論文・Web記事

速度/費用効率: 4.5 / 5.0 (黄)

SearchFetch: \$0.75/calls

技術的留意点・品質リスク

1017版で追加された公式SearchFetch APIとエージェント連携 (Automation-Bench 60.5) が高効率。一方で事実QA (SimpleQA 42.3) のスコア低下によるハルシネーションリスクがあるため、抽出URLの誤検知が不可避。

法的・ガバナンス評価

特許・企業文書のサーバー側検索ポリシーが適用中。社内の非公開契約テーマやコードネームを検索キーワードに含めることは厳禁。

推奨システム構成

公式SearchFetch + 自働エージェント

DeepSeek IRISense等の特許向け「公開技術キーワード」を用いた自律スクレイピングを実行。取得Markdownの要約を1017に接続し、エビデンスリンク付きレポートを自動生成。

公開情報領域・マルチモーダル

3. 図面・化学構造式・表を含む公報のマルチモーダル読解

対象データ

公報の機械読解・フロー・表画像

速度/費用効率: 4.0 / 5.0 (黄)

V4-Flash: 複雑な利用法で定期的

技術的留意点・品質リスク

API2PDF直接入力を受けサポート。また図面は内部的に約800-8000画素 (上: 384x4トークン) へ縮小処理されるため、複雑な化学構造式や特許図面の細小引き出し・細線が読まれるリスクに厳格な注意が必要。

法的・ガバナンス評価

発行済みの公開特許図面を対象とする限り、機密保持上の低リスクは生じない。

推奨システム構成

クライアント側事前ラスタライズ + 1017 API

特許PDFをクライアント側で高解像度PNG/JPEGへ変換し、APIパラメータdetails: "url_image" で送信。複雑図面のOCR処理前後で解像度調整を推奨。

特許・要約保証領域

4. 構成要件対比表(クレームチャート)・拒絶理由通知の要約

対象データ

審査記録・クレーム・拒絶理由

速度/費用効率: 4.0 / 5.0 (黄)

対比要件あたり \$0.10-\$0.20

技術的留意点・品質リスク

思考モード (reasoning) と構造的出力が対比表の要素分類に貢献。ただし、審査官の膨大な論議の機軸や先行技術との差違性判断は高度推論 (HLE 42.6 vs Opus 5 36.3) で明確な差があり、1017単体での完結は困難。

法的・ガバナンス評価

自社係争案件や公開前補正案を含む場合、公式APIに守秘義務規定がないため情報漏洩リスクが生じる。マスキング・匿名化の前処理が不可欠。

推奨システム構成

DeepClaude ハイブリッド構成

安価な1017版で公開記録の構造化抽出・要点抽出 (一次下駄) を行い、最終的な要約・差違性判断の高精度をClaude Opus 5 (または人間担当) が担当する2段階ハイブリッド。

高度機密・非公開発明領域

5. 出願前の発明提案書処理・明細書全文執筆・侵害予防調査(FTO)

対象データ

未公開発明・社内機密・FTO調査

速度/費用効率: 1.5 / 5.0 (赤)

自社IP/オンプレ運用が必須

技術的留意点・品質リスク

1017版は中核的な強みに特化調整されたため、特許情報に必要とされる高度かつ複雑な実装・多層的なクレーム生成には不向き。また専断OA標準 (42.3) の成否はFTOクリアランスで数値的リスク。

法的・ガバナンス評価 (最重要)

公式契約に守秘義務規定が存在せず、送信データは高度機密 (秘密管理性) を要する法的リスクが高い。また中国近接特許法の適用リスクもあり、未公開発明の送信は特許新規性喪失の致命的危険を伴む。

推奨システム構成

公式API遮断 / オンプレ or 他社NDA環境

MITライセンスに基づく組み込み利用し、ネットワーク隔離された自社GPU環境 (vLLM/SQLang等) に高度実装。あるいは正式な秘密保持契約の締結済みの商用クラウドサービス (Claude Opus 5 Enterpriseプライズ等) ヘルパーリング。

知財実務の5大領域におけるDeepSeek-V4.1-Flash-1017の経済的メリット、技術的制約、法的リスク、および推奨されるシステム構成の対比。公開情報に対象とする大量スクリーニングやクレーム対比表の一次生成では公式API (オフピーク適用) が圧倒的な投資対効果を生む一方、出願前の発明提案書やFTO調査などの機密業務では、公式クラウドAPIの直接利用は営業秘密喪失・新規性喪失リスクを伴うため、オープンウェイトの自社環境展開または他社フロンティアモデルとのハイブリッド構成が不可欠となる。

先行技術調査（公知文献スクリーニング）とIPランドスケープの経済性革命

すでに公開された特許公報や学術論文といった「公知情報（Public Domain / Published Data）」を大量に処理する業務において、1017版は既存の知財調査のコスト構造を根底から覆す破壊力を持つ¹³。IPランドスケープ分析における数千～数万件規模の特許公報の一次スクリーニング（SDI調査、テーマ別クラスタリング、課題と解決手段のマトリクス化）では、APIの同時実行上限が2,500（V4-Proの5倍）に設定されているため、大規模な並列バッチ処理を高速に回すことができる²。さらに、夜間（19:00以降）や週末のオフピーク時間帯にジョブを実行し、共通のシステムプロンプトや分類定義書にKVキャッシュ（入力100万トークンあたり0.003ドル＝約0.47円）を効かせることで、100万文字単位の公報データを数円～数十円レベルの極小コストで処理できる¹。1017版の「余計な前置きや冗長な解説を省いて端的に出力する」特性は、JSON形式での技術属性タグ付けや関連度スコアリングにおいて出力トークン課金を大幅に圧縮し、処理完了までのリードタイムを短縮する直接的なメリットとなる¹。

加えて、同時開放されたSearch APIとFetch API（各1,000成功リクエストあたり0.75ドル）は、特許文献だけでなく技術ブログ、標準化団体の仕様書、オープンソースリポジトリ、製品マニュアルといった非特許文献（NPL: Non-Patent Literature）の自律調査エージェントを極めて安価に実現する⁴。たとえば、site:github.comやbefore:2025-01-01（出願日前の公知日指定）といった検索演算子をSearch APIに渡し、抽出された高権威ページ（authority >= 3かつrerank_score >= 1.5）の全文をFetch APIでMarkdownとして取得し、1017版の思考モードで請求項の構成要件と対比させる一連の無効資料調査パイプラインを、外部クローラーなしで構築できる⁴。ただし、公式ガイドにはSearch/Fetch APIに送信した検索クエリや取得URLの保存・学習利用に関する明確な非保持規定が存在しないため、未公開発明の核心キーワードや自社の開発コードネームを含む検索クエリは絶対に送信してはならず、すでに公知となった競合他社の特許番号や汎用技術用語を用いた調査（無効資料調査等）に限定する運用規律が必須である²。

特許図面・化学構造式・表データのマルチモーダル解析とAPI制約

これまでDeepSeekの上位モデルであるdeepseek-v4-proはテキスト専用であり画像を読み込むことができなかったが、V4.1-Flash系はネイティブ・マルチモーダル機構（DeepSeek-ViT）を搭載しているため、特許公報に不可欠な「図面（ブロック図、機械構造図、フローチャート、UI画面）」「化学構造式」「複雑な実験データ表」と明細書本文を単一の100万トークンコンテキスト内で統合理解させることが可能となった⁸。視覚ツールを用いるChartography（83.0）やBabyVision（92.5）でもClaude Opus 5（84.0 / 94.1）に肉薄するスコアを達成しており、図表を含む技術文書の読解において極めて高いコスト優位性を発揮する²。

もっとも、日本の知財実務で特許公報PDFを処理する際には、いくつかの技術的・仕様上の制約をシステム設計に織り込む必要がある¹³。第一に、公式APIはPDFファイルの直接解析をサポートしておらず、対応フォーマットはJPEG、PNG、GIF、WebPの4種類に限定されているため、特許公報PDFはクライアント側で事前にページ単位または図面単位の画像へラスタライズ（画像変換）してからAPIへ送信する前処理パイプラインが必須となる⁴。第二に、API上では1リクエストあたり最大600枚の画像を添付でき、最大辺は8,192ピクセル（15枚以上の画像を同時添付する場合は4,096ピクセル）まで受け付ける仕様となっているものの、モデル内部での処理時には1画像あたり最大384～1,024視覚トークン（解像度換算でおおむね800×800～1300×1300ピクセル相当、detail="low"指定時は512×512ピクセル）へと自動的に縮小（ダウンサンプリング）される⁴。日本の特許図面に特有の細かい引き出し線（リード線）、密集した参照符号（10a, 10b等）、複雑な回路図や微小な化学式の添字

は、ページ全体を1枚の画像として入力すると縮小処理によって潰れて誤読されるリスクが高いため、公報ページ全体をそのまま投げるのではなく図面領域のみを高解像度でクロップ(切り出し)して分割入力する工夫が求められる¹³。第三に、公式発表ではDocVQAで95.6という高スコアが報告されているものの、これは一般的な標準指標(ANLSスコア)ではなくLLM-Judge(LLMによる採点)を用いた独自評価であり、他社モデルとの厳密な横並び比較ができない⁷。さらにOCRBenchやChartQAなどの標準的な文書読解ベンチマークも非開示であるため、日本語の縦書き・横書き混在公報や古いスキャン公報に対する文字認識精度は、導入前に自社の実データで検証する必要がある²。

明細書ドラフティング・クレーム解釈・中間処理(OA対応)における適性と死角
特許明細書の作成や拒絶理由通知(Office Action: OA)への対応業務においては、1017版の「思考モード(Thinking Mode)」と「出力短縮設計」が明確な二面性を見せる¹。構成要件の充足性対比(クレームチャート作成)やOA要約の場面では、思考モード(reasoning_effort="high"または"max")を有効にした1017版が出力前に内部で仮説構築・検証・自己修正(Self-correction)の思考プロセスを経るため、本願発明の請求項(クレーム)を構成要件(分説A、B、C...)に分解して引用文献の該当箇所と論理的にマッピングするタスクや、審査官の拒絶理由の論理的瑕疵を抽出する作業において高い精度を発揮する¹。日本語の指示追従性(elyza等で首位級)が高く、出力が端的にまとまる特性も相まって、長大な公報から指定したJSONスキーマやMarkdownテーブル形式でクレームチャートを自動生成させる用途には極めて相性が良い¹。

その一方で、特許明細書の「発明を実施するための形態(実施形態・実施例)」の全文ドラフティングにおいては、1017版の「短文癖」が弊害となり得る¹。知財実務における明細書の執筆では、出願後に補正の根拠を確保し権利範囲の解釈を広げるために、あえて上位概念から下位概念、変形例に至るまで網羅的かつ冗長に書き込むことが求められる。1017版は強化学習によって不要な長文出力を削り最短距離でタスクを完了させるよう調整されているため、無調整のプロンプトで明細書の執筆を指示すると、記載が過度に簡略化され、実施可能要件やサポート要件を満たす十分な肉付けが不足する傾向がある¹。長文の明細書ドラフティングに用いる場合は、セクションごとに分割して生成させるか、システムプロンプトで最小文字数や記載すべき変形例の粒度を厳格に指定する必要がある¹。

さらに、審査経過書類(包袋)一式や複数の先行技術文献を100万トークンのコンテキストに一括投入して包袋禁反言(ファイルラッパー・エストッペル)リスクを洗い出すような高度な法務・知財タスクでは、前述したLongBench-V2のスコア低下(45.2 vs V4-Proの51.5)およびCSA2/SWA Bounded Replayの近似再構築による「長文中の局所的な限定記述の見落とし」が致命傷となり得る⁶。また、SimpleQA-Verified(42.3)が示す通り、モデル自身の内部記憶に頼って「周知技術」や「判例・審査基準」を語らせるとハルシネーションを起こしやすいため、必ず審査基準や引用文献の原文テキストをコンテキスト内に明示的に与えた上で参照箇所(段落番号等)を引用させるグラウンディング設計が不可欠である⁷。加えて、0910版で構築した自動プロンプトをそのまま使用している場合、1017版への自動切り替わりによってイベント結合やロジックの一部が欠落するデグレード現象が起き得るため、既存の知財ワークフローについては直ちに再評価(リグレッションテスト)を実施しなければならない²。

機密保持・新規性喪失リスクと「DeepClaude」型ハイブリッド・アーキテクチャの構築

知財業務における生成AI活用で最も厳格な境界線が引かれるのが、「出願前の未公開発明(発明

提案書・研究開発データ)」および「自社の事業戦略に関わる機密情報（FTO調査の対象製品仕様、M&A時の知財デューデリジェンス等）」の取り扱いである¹³。DeepSeekの公式プライバシーポリシーおよび利用規約（Terms of Use / Open Platform Terms of Service）を法務的観点から精査すると、中国国内サーバーでのデータ保存と国家情報法の適用、および守秘義務条項（Confidentiality Clause）の欠如という2つの重大なガバナンスリスクが確認される²²。

DeepSeekはユーザーから収集した入力データ（プロンプト、アップロードファイル、画像等）を中華人民共和国国内のサーバーで直接収集・処理・保存することを明記しており、中国の法令および政府機関・法執行機関からの開示要請に従う規定が存在する¹³。日本のデジタル庁も各省庁に対してDeepSeek等の業務利用に関する注意喚起を行っており、機密情報を扱う組織には厳格な管理が求められている²³。さらに、専門法律事務所の規約解析でも指摘されている通り、DeepSeekの利用規約等には運営側がユーザーの入力情報の秘密を保持する旨の守秘義務条項が存在しない²²。プライバシーポリシー上はユーザーの入力内容をAIモデルの学習およびサービス改善に利用することがデフォルトで認められており、無料版・有料版を問わず製品画面上で確実に学習を拒否（オプトアウト）する標準的な保証メカニズムが欠如している²⁴。

この法的現実、知財実務において決定的な意味を持つ。もし出願前の発明提案書や未公開のソースコードをDeepSeekの公式Web画面、公式アプリ、または公式APIへ直接入力した場合、守秘義務のない第三者（および潜在的にはモデルの学習データセット）へ発明内容を開示したとみなされ、不正競争防止法上の「営業秘密」の要件である非公知性・秘密管理性を喪失するリスクが生じるだけでなく、最悪の場合には特許法上の「新規性喪失（公然知られた発明）」に該当し、自社で特許を取得できなくなる法的リスクを招く²²。したがって、知財部門および特許事務所が1017版の卓越した性能とコスト優位性を安全に享受するためには、データの機密性に応じた3段階のアーキテクチャ分離と「DeepClaude」型ハイブリッド・ルーティングを確立することが最善の戦略となる¹³。

運用アーキテクチャ 階層	対象となる知財データ・業務範囲	採用すべきインフラ・ モデル構成	ガバナンスおよび技術的根拠
第1階層：公式クラウドAPI（オフピーク・ バッチ運用）	公知情報のみ（公開特許公報の大量一次スクリーニング、IP ランドスケープ分類、 公知製品を対象とする無効資料のWeb/ 文献検索）	deepseek-flash（公式API・夜間/週末実行）+ Search / Fetch API	すでに公開された公報や汎用技術用語のみを入力するため新規性喪失リスクがない。オフピーク単価（キャッシュヒット\$0.003/M）と並列2,500枠により調査コストを極小化 ² 。
第2階層：自社環境（オンプレミス／ゼロ データ保持クラウド）	機密・未公開情報を含む大量処理（社内発明提案書の一次分類、未公開ソースコードの解析、社内	MITライセンスの公開重み（DeepSeek-V4.1-Flash-1017）を自社管理GPUクラスター（vLLM /	重みファイルを自社環境へ閉じて運用することで、外部（中国サーバー）への通信を完全に遮断し、国

	技術文書の構造化・図面読解)	SGLang + deepseek-recipe) またはゼロデータ保持保証のOllama Cloud等に展開	家情報法や営業秘密喪失リスクをゼロにする ¹ 。
第3階層:「 DeepClaude 」ハイブリッド・ルーティング	高度な法的推論・長文精読・最終成果物作成(クレームドラフティング、包袋禁反言の精査、FTOの法的当てはめ、長文明細書の執筆)	バックエンド(下読み・公開文献抽出): 1017版 フロントエンド(最終論理推論・機密起案): Claude Opus 5 (ゼロデータ保持のエンタープライズ契約環境)	1017版が苦手とする高度な専門知識推論(HLE)、長文精密理解(LongBench-V2)、および冗長性が求められる明細書執筆を Opus 5で補完し、全体のAPI費用を劇的に圧縮する ² 。

なお、サードパーティの中継サービス(例: mydeepseekapi.com等)の中には独自の「ゼロデータ保持」を謳うものも存在するが、公式基盤ではない非公式プロキシに機密知財データを預けることはサプライチェーン上のセキュリティリスクが極めて高いため、機密データを扱う際は必ずHugging FaceまたはModelScopeから取得したMITライセンスの公式ウェイトを自社の管理下にあるインフラ(または法的にゼロデータ保持と学習非利用が契約保証された国内・欧米の推論基盤)にデプロイすべきである¹。

また、よりマクロな知財戦略の視点として、DeepSeek社自身の知的財産動向にも注視が必要である³。同社はモデル重みをMITライセンスで無償開放してエコシステムの覇権を握る一方、その基盤となるMoEルーティングやKVキャッシュ圧縮(CSA2等)のコアアルゴリズムおよびシステムアーキテクチャについては独自の特許網構築を急速に進めている³。同時に、米国政府機関(NSA・CISA・FBI)からは中国企業によるフロンティアモデルからの「知識蒸留(Distillation)」に対する警戒勧告が出されており、1017版の学習プロセスでも40以上の教師モデルからの蒸留や大規模自動合成が活用されていることが指摘されている¹³。自社プロダクトにDeepSeekのモデルを組み込んで商用展開する企業においては、MITライセンスが「無保証(AS IS)」であり第三者権利侵害に対する補償を含まないことを踏まえ、利用規約および各国のAI規制・輸出管理法制の動向を継続的に監視する体制が求められる²³。

4. 結論と実務導入に向けた推奨アクション

2026年10月17日に公開された「DeepSeek-V4.1-Flash-1017」は、非対称CED構造とKVキャッシュ超圧縮(890 bytes/token)という強固なインフラ効率の上に、「自己フィードバック強化学習による実装力の向上」と「出力トークン短縮による実質コストの半減(DeepSWE v1.1で1タスク約0.20ドル弱)」を実現した、エージェントおよび実務自動化領域におけるゲームチェンジャーである¹。同時に開放されたSearch APIおよびFetch APIとの連携により、外部情報の収集から構造化・コード生成までを単一プラットフォーム上で完結できる経済的優位性は他社を圧倒している⁴。

しかし、その導入にあたっては「ベンチマーク平均値の向上」と「実務上の安全性・互換性」を切り分

けて判断しなければならない⁷。すでにdeepseek-flash(または旧deepseek-v4-flash)を業務パイプラインに組み込んでいる組織は、10月17日をもって中身が1017版へ強制置換されているため、出力の短縮化やイベント処理の変質(デグレード)によって後続のパーサーやUI、自動分類ロジックに異常が生じていないか直ちに実データで検証を行う必要がある²。その上で、公式Web、アプリ、および公式API(Search/Fetch APIを含む)への入力、公開特許公報や公知文献のスクリーニング・無効資料調査に厳格に限定し、出願前の発明提案書や未公開の技術仕様を公式クラウドへ送信することは営業秘密の喪失と新規性喪失を招くため社内規程で明確に禁止すべきである²。

実務運用の最適化に向けては、先行技術調査の一次スクリーニングやIPランドスケープのクラスタリングにおいて、日本時間の平日日中(10:00~13:00、15:00~19:00)を避け、夜間や週末のオフピーク時間帯に並列上限(2,500)とKVキャッシュ(\$0.003/MTok)を最大限に活用するバッチ設計へ移行することで調査コストを極小化できる²。公報図面を入力する際はPDFを事前ラスタライズし、微細な参照符号が潰れないよう図面領域をクロップして入力する前処理を実装することが望ましい¹³。そして最終的なシステムアーキテクチャとしては、定型的なエージェント処理や大量の公報下読みを1017版に割り振る一方、高度な専門知識推論(HLE領域)、100万トークン中の微細な限定事項の精査(包袋解析)、および意図的な冗長性が求められる明細書全文の執筆はエンタープライズ契約下のClaude Opus 5等へルーティングする二段構えの体制を敷き、機密データを1017版で直接処理したい場合はMITライセンスの公開重みをdeepseek-recipeとともに自社管理のGPU環境へデプロイして外部通信を遮断したクローズド環境で運用することが強く推奨される²。

引用文献

1. DeepSeek-V4.1-Flash-1017とは？APIの使い方や料金 - AI PICKS, <https://aipicks.jp/mag/deepseek-v4-1-flash-1017-what-is-2026>
2. DeepSeek V4.1 Flash 1017とは？変更点・料金・Search API・旧, <https://ai-revolution.co.jp/media/deepseek-v4-1-flash-1017/>
3. DeepSeek V4 最新モデル調査 Gemini, <https://yorozuipsc.com/uploads/1/3/2/5/132566344/e394b23e5f3219a618aa.pdf>
4. <https://api-docs.deepseek.com/news/news261017/>
5. DeepSeek-V4.1-Flash-1017とは？性能・料金・Search APIを解説, <https://dot-ai.myuuu.co.jp/times/articles/1620>
6. DeepSeek-V4.1-Flash、KVキャッシュを約4分の1に圧縮 長文脈AIの, <https://www.zaikei.co.jp/article/20260921/870800.html>
7. DeepSeek-V4.1-Flash 深掘り分析レポート, <https://yorozuipsc.com/uploads/1/3/2/5/132566344/655129536e3b6ec97a98.pdf>
8. DeepSeek-V4.1-Flash Analysis Gemini, <https://yorozuipsc.com/uploads/1/3/2/5/132566344/6d89e3bfb9b5e5cd733d.pdf>
9. DeepSeek V4.1 Flash - API Pricing & Benchmarks | OpenRouter, <https://openrouter.ai/deepseek/deepseek-v4.1-flash>
10. DeepSeek V4.1-Flashとは | 料金と旧モデルの扱い | AI活用顧問 - Fyve, <https://fyve.co.jp/ai-advisor/articles/deepseek-v41-flash-model-routing>
11. deepseek-v4.1-flash - Ollama, <https://ollama.com/library/deepseek-v4.1-flash>
12. DeepSeek-V4.1-Flash: Pushing the Limits of KV Cache Compression, <https://arxiv.org/abs/2609.19969>
13. DeepSeek-V4.1-Flash 徹底調査レポート,

- <https://yorozuipsc.com/uploads/1/3/2/5/132566344/f7662ca80969d79df181.pdf>
14. DeepSeek-V4-Pro-0813の全貌,
https://yorozuipsc.com/uploads/1/3/2/5/132566344/deepseek_v4_pro_0813_strategic_analysis.pdf
 15. DeepSeek V4.1 Flash ベンチマーク: 74.2 が本当に証明するもの,
<https://www.orcarouter.ai/ja/blog/deepseek-v4-1-flash-benchmark>
 16. DeepSeek V4.1とは？Flash版の料金・性能を解説 - AI革命株式会社,
<https://ai-revolution.co.jp/media/what-is-deepseek-v4-1/>
 17. 【DeepSeek-V4.1-Flash-1017】「軽量・爆速・低コスト」で ... - note,
https://note.com/humble_bobcat51/n/n760dd174f5bd
 18. DeepSeek「V4.1 Flash-1017」が登場したので試してみた！Opus 5,
https://note.com/masa_wunder/n/n0770dbbe6720
 19. DeepSeek V4.1 Flashの使用量、コスト、ランク | OpenCodeデータ,
<https://opencode.ai/ja/data/deepseek/deepseek-v4-1-flash>
 20. DeepSeek-V4-Flash、GLM-5.2、MiniMax-M3の日本語性能 ... - Qiita,
<https://qiita.com/nabe2030/items/7ae4a739bf45ecae3236>
 21. DeepSeek V4.1 Flash 1017 - API Pricing & Benchmarks | OpenRouter,
<https://openrouter.ai/deepseek/deepseek-v4.1-flash-1017>
 22. DeepSeek利用上の法的な注意点, <https://storialaw.jp/blog/11668>
 23. DeepSeekとは何か？性能・危険性・商用利用と業務での使い方を解説,
<https://liber-craft.co.jp/column/deepseek-explained>
 24. DeepSeek Privacy Policy,
<https://cdn.deepseek.com/policies/en-US/deepseek-privacy-policy-2025-02-14.html>
 25. DeepSeek Privacy Policy,
<https://cdn.deepseek.com/policies/en-US/deepseek-privacy-policy.html>
 26. DeepSeek Terms of Use,
<https://cdn.deepseek.com/policies/en-US/deepseek-terms-of-use.html>
 27. DeepSeek Open Platform Terms of Service,
<https://cdn.deepseek.com/policies/en-US/deepseek-open-platform-terms-of-service.html>
 28. DeepSeek Privacy and Training Data: What Happens to Your,
<https://ai-chat-importer.com/blog/deepseek-privacy-training-data>
 29. DeepSeek Privacy Policy 2026: Data Retention, Storage & AI Training,
<https://isogarindia.com/deepseek-privacy-policy-a-critical-analysis-of-risks-and-implications/>
 30. I just found that even if you use their paid API, they still train ... - Reddit,
https://www.reddit.com/r/DeepSeek/comments/1vcoi6s/i_just_found_that_even_if_you_use_their_paid_api/
 31. MyDeepSeekAPI Privacy Policy – Data Use & Retention,
<https://mydeepseekapi.com/privacy>
 32. DeepSeek V4,
https://yorozuipsc.com/uploads/1/3/2/5/132566344/deepseek_v4_secure_ip_strategy.pdf
 33. DeepSeek-V4.1-Flash-1017公開、検索APIも全ユーザーに開放,

<https://innovatopia.jp/ai/ai-news/119008/>