

DeepSeek-V4.1-Flash-1017

パフォーマンス解析と知財業務への戦略的影響

Gemini 4 Argon 調査報告・実装プレイブック

Target: IP Strategists, Patent Attorneys, CIO, CTO
Date: 2026.18
Clearance: STRATEGIC BRIEFING

Breakthroughs (技術的突破)

- 非対称CED構造と自己フィードバック強化学習
- 1Mトークンの長文脈を極小コストで処理する実務特化型エージェントの誕生

Risks & Traps (運用上の罠)

- サイレント置換による既存システム破壊
- ハルシネーション (SimpleQA : 42.3) の残存

Economics (コスト破壊)

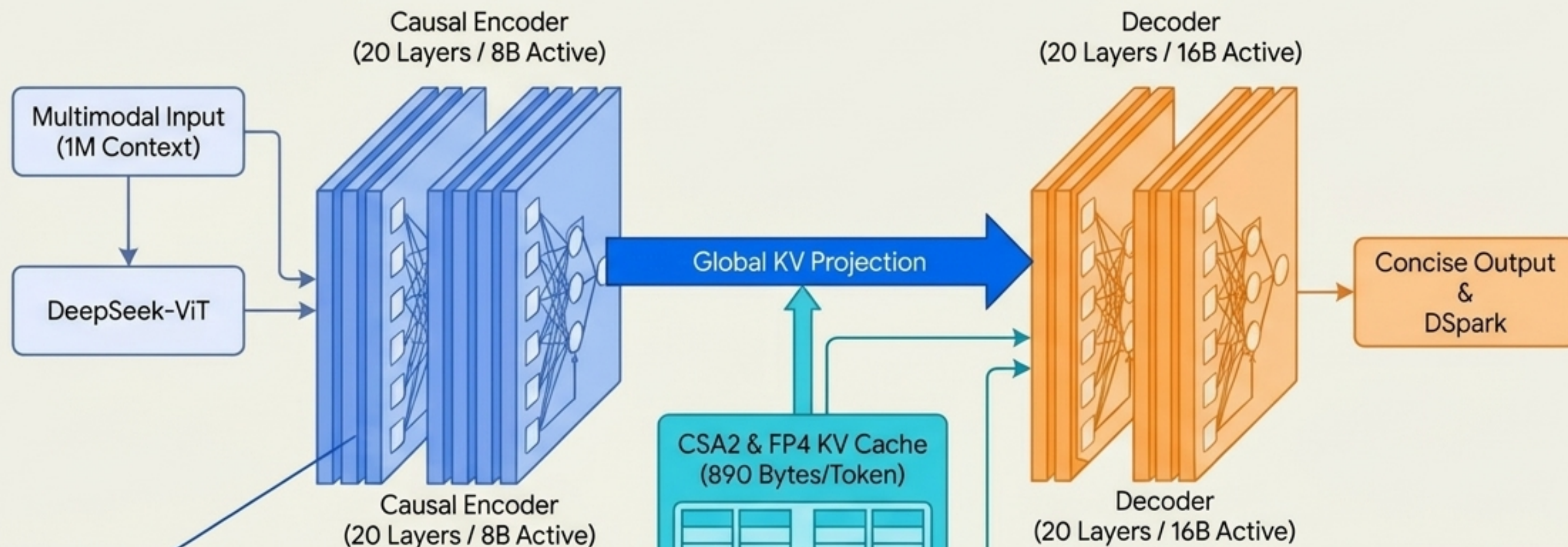
- パレート改善による劇的なコストダウン
- 出力短縮により1タスク実行コストが\$0.10～\$0.20へ暴落 (Claude Opus 5の数十分の一)

IP Fatal Flaw (知財の致命傷)

- 中国国家情報法と守秘義務の欠如
- 未公開発明の入力は「新規性喪失」と「営業秘密の喪失」を直撃する

The Playbook (戦略的解決策)

「DeepClaude」3層ハイブリッド・ルーティングアーキテクチャによる安全なデータ分離



パラメータの非対称性 (CED構造)

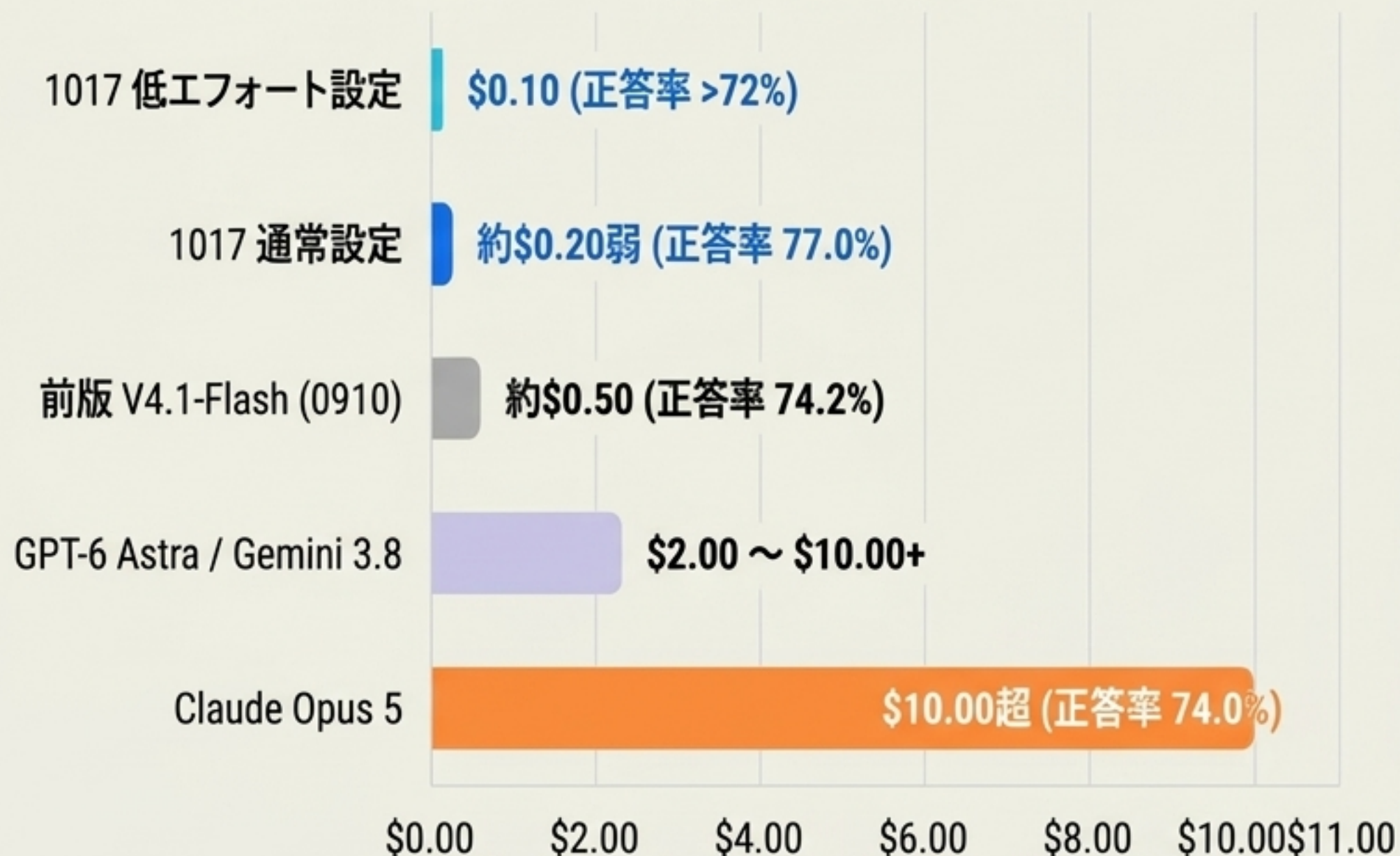
- 総パラメータ: 763B (MoE 552B + Engram 196B)
- 入力読込時 (Prefill): 8B 活性
- 出力生成時 (Decode): 16B 活性
- 大量の特許公報読み込みに圧倒的有利な非対称設計

KVキャッシュの極限圧縮

- 1トークンあたり HBM常駐分 890 Bytes
- CSA2とFP4量子化による極限圧縮
- 初代V1比「1/437」、前世代比1/4
- SWA Bounded Replayにより永続キャッシュも1/8へ縮小

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



パレート改善とコスト破壊のメカニズム

自己フィードバック強化学習 (許願式プログラミング)

願いを言うだけで実現するエンドツーエンドの実装力



出力の劇的な短縮化

前版の長文癖 (中央値の1.8倍消費) を是正し、最短ルートでタスクを完了



ROIの完全逆転

DeepSWE v1.1タスクコストが\$0.50から約\$0.20 (低エフォート時\$0.10) へ暴落。Opus 5 (\$10超) を圧倒。

**結論: エージェント動作において、
フロンティア最高精度を最安水準で実現**

動的価格設定ヒートマップ

(日本時間)	MON	TUE	WED	THU	FRI	SAT	SUN
00:00							
01:00							
02:00							
03:00							
04:00							
05:00	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)
06:00						Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)
07:00							
08:00							
09:00							
10:00							
11:00	Peak (\$0.15/\$1.20)	Peak (\$0.15/\$1.20)	Peak (\$0.15/\$1.20)	Peak (\$0.15/\$1.20)	Peak (\$0.15/\$1.20)		
12:00							
13:00							
14:00	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)
15:00							
16:00							
17:00	Peak (\$0.15/\$1.20)	Peak (\$0.15/\$1.20)	Peak (\$0.15/\$1.20)	Peak (\$0.15/\$1.20)	Peak (\$0.15/\$1.20)		
18:00							
19:00							
20:00							
21:00	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)	Off-Peak (\$0.003/\$0.60)		
22:00							
23:00							

24時間 in JST (日本時間)

ピークの半額以下。キャッシュヒット時は極小コストへ。

Search & Fetch API パイプライン

価格：成功リクエスト1,000件あたり \$0.75



Search API

rerank_score / authority 付き高度検索



Fetch API

対象ページのMarkdown本文抽出



知財インパクト

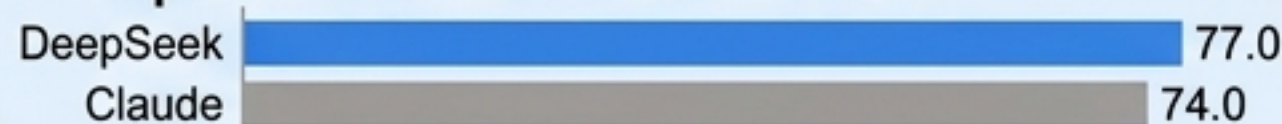
外部スクレイピング基盤不要。
出願日前の非特許文献 (NPL)
無効資料調査がAPIキー1つで自律化。

現実チェック: 1017版の真の得意・不得意

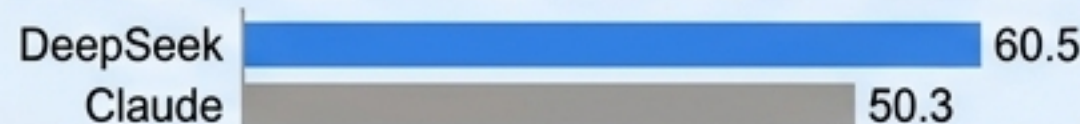
DeepSeek 1017の勝利領域

実務自動化とエージェント

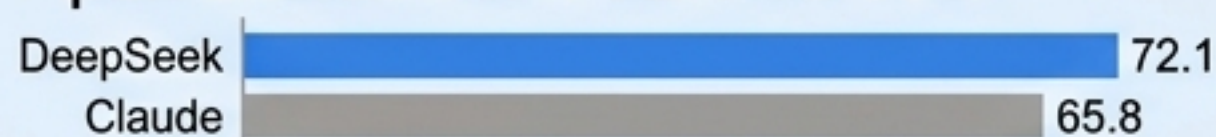
- DeepSWE v1.1: 77.0 vs 74.0



- Automation-Bench: 60.5 vs 50.3



- Apex Bench v1.1: 72.1 vs 65.8

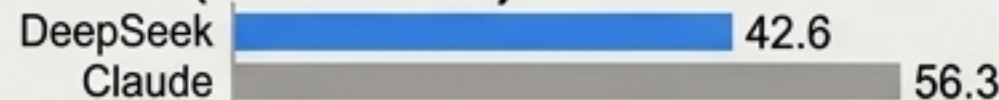


強み: ツール呼び出し、ワークフロー自動化、定型コード改修

Claude Opus 5の勝利領域

高度推論とゼロからの構築

- HLE (学術難問): 42.6 vs 56.3



- NL2Repo-Bench: 68.2 vs 75.3



- ProgramBench: 29.1 vs 37.0



- Terminal-Bench 4.0: 42.3 vs 51.8

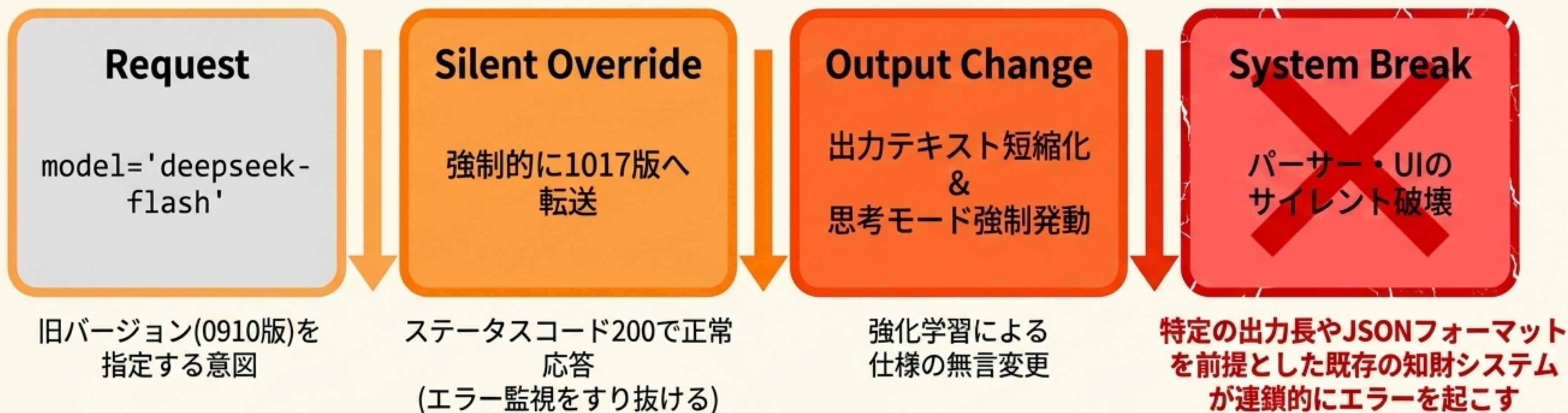


強み: 純粋な論理推論、大規模リポジトリ構築

⚠️ ボトムライン: 1017版は「汎用AGI」ではない。
「実務特化のタスク・スペシャリスト」として扱うこと。



システム崩壊の罠: サイレント置換プロセス



対策 (Mitigation)

公式APIでのバージョン固定 (Pinning) は不可能。
旧版の維持にはHugging Faceオープンウェイトのオンプレミス展開が必須。

技術的盲点: 長文脈と最適化の副作用

記憶の劣化 (Amnesia)



指標: LongBench-V2 45.2
(旧V4-Proを下回る)

原因: CSA2層間共有と状態近似

知財リスク: 長大な特許包袋
の中の「ごく僅かな限定記述
(Needle-in-a-Haystack)」を
見落とす危険性。

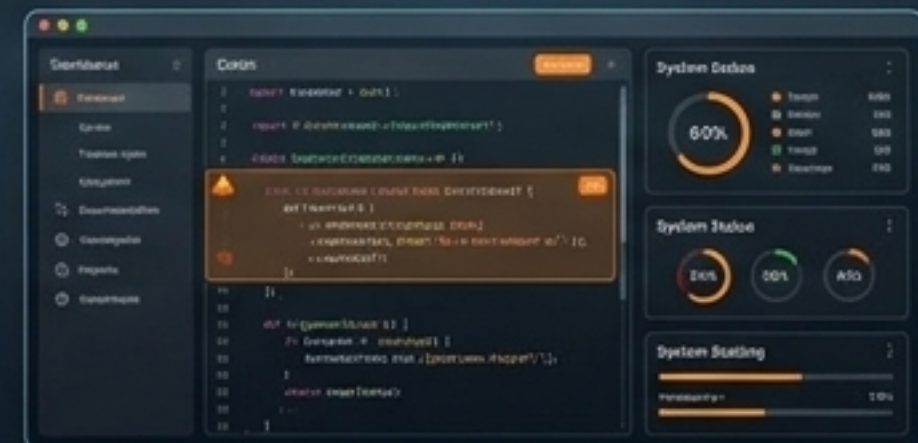
ハルシネーション



指標: SimpleQA 42.3 /
非虚偽率 3.5%

知財リスク: 内部記憶のみで
「判例」や「周知技術」を語
らせると虚偽情報を捏造す
る。原文テキストの厳密な
グラウンディングが必須。

最適化によるデグレード



事例: 簡易Figmaエディタの
機能不全

知財リスク: コードの「見栄え」
に最適化され、細かい論理結合
やイベントが省略される癖が
ある。0910版で稼働していた
プロンプトの再検証が必須。



知財における致命的欠陥: 法的リスクと新規性喪失 (Fatal Flaw in IP: Legal Risks and Loss of Novelty)

Probability (発生確率)

High
(高)

Low
(低)



Low (低)

High (高)

Impact on IP (知財への打撃)

法的根拠 (Legal Facts)

- 利用規約に「守秘義務条項 (Confidentiality Clause)」が存在しない
- デフォルトでAIの学習データとして利用される (オプトアウト保証欠如)
- 中国国内サーバーでの処理と国家情報法の適用リスク



発生する事象 (The Action)

未公開発明 (発明提案書、ソースコード等) を公式APIやWeb画面に入力する行為



法的帰結 (The Consequence)

- 「守秘義務のない第三者への開示」とみなされる
1. 不正競争防止法上の「営業秘密の喪失」
 2. 特許法上の「新規性喪失 (公然知られた発明)」
- 結論: 自社特許の無効化リスクへ直結する。

知財タスク適性マトリクス: 1017版の安全な活用領域



Perfect Fit (導入推奨 - パブリックデータ領域)

タスク

大量一次スクリーニング・IPランドスケープ / NPL無効資料調査

理由

公開済みの特許・文献を使用するため法的リスクゼロ。2,500並列のバッチ処理と極小コストが最大限に活きる。



Conditional Conditional (条件付き - プロンプト制御必須)

タスク

クレームチャート作成・OA (拒絶理由通知) 要約

理由

思考モードの論理展開は優秀だが、審査基準等の厳密なプロンプト入力 (グラウンディング) を行わないとハルシネーションを起こす。



Danger (不適合 - リスク極大)

タスク

明細書 (実施形態) の全文ドラフティング

理由

モデルの「短文癖」により記載が簡略化され、特許に必要な冗長な肉付け (サポート要件) が不足する。未公開發明 の入力リスクとも衝突。

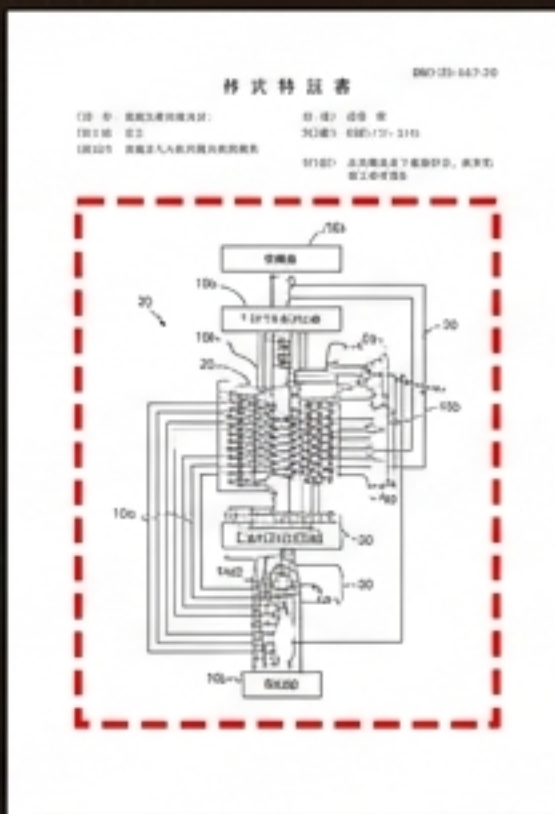
ネイティブ・マルチモーダル特許解析パイプライン

公報
PDF

クライアント側
ラスタライズ

APIはPDF非対応。
JPEG/PNGへ画像変換。

精密クロップ
(切り出し)

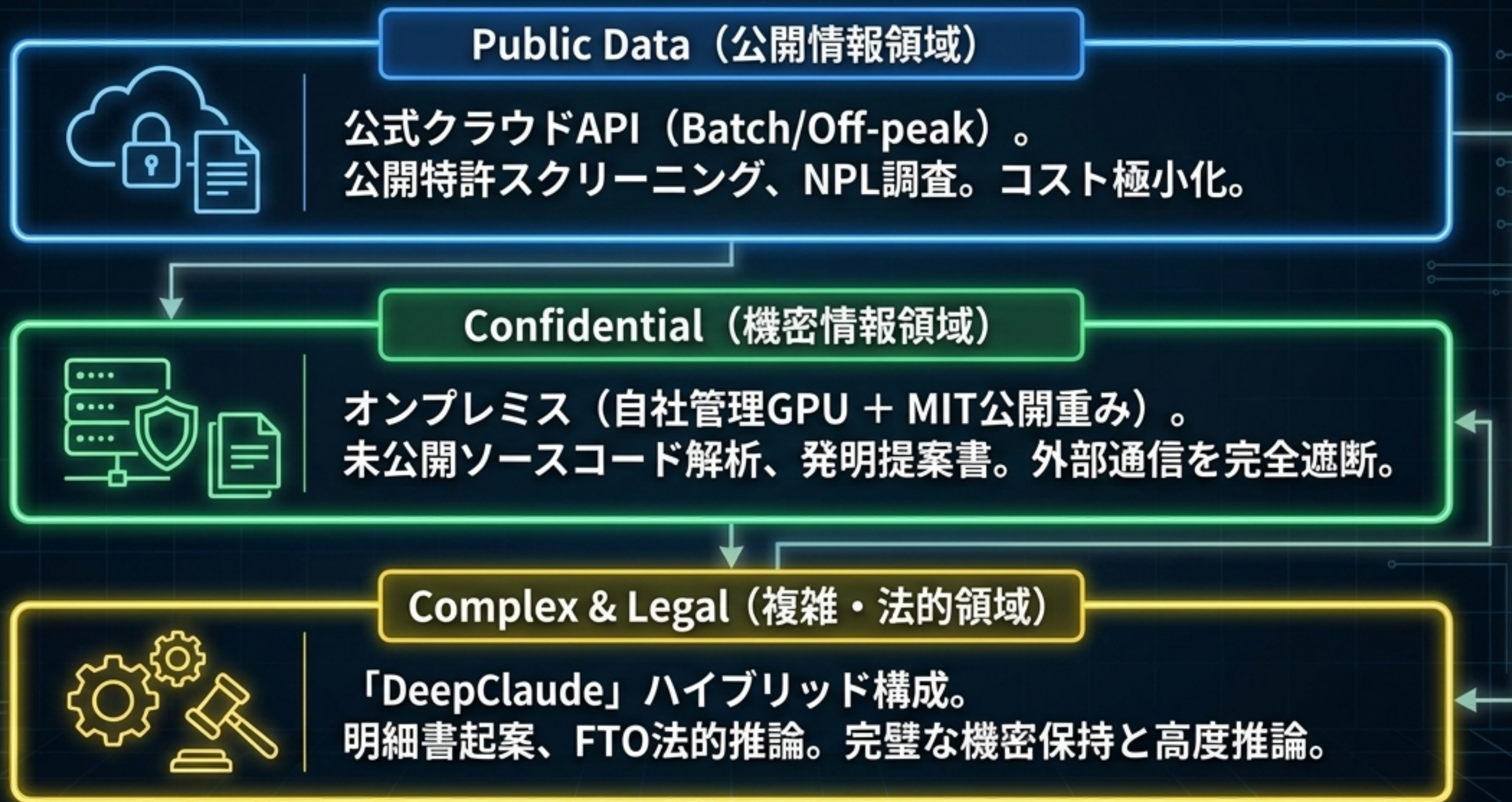


最重要工程。ページ全体を投げると自動ダウンサンプリング (最大384視覚トークン) により、密集した参照符号が潰れてOCRエラーを起こす。図面領域のみを高解像度で分割入力せよ。

DeepSeek-ViT
API

1Mトークン内で特許ブロック図・化学構造式・表をテキストと統合理解。

THE PLAYBOOK: 3層データルーティング・アーキテクチャ



究極の解決策: DeepClaude ハイブリッド実行モデル



Backend:
DeepSeek-V4.1-Flash-1017

役割: Heavy Lifting (重労働)

- 1Mトークンの長大な公開先行技術文献を超低コストで読み込み、関連箇所を抽出・構造化。

Frontend:
Claude Opus 5

役割: Final Thinking (最終推論)

- 抽出されたデータを受け取り、高度な法的推論、ゼロデータ保持契約下での機密情報の当てはめ、及び冗長な特許クレーム起案を行う。

メリット: DeepSeekの圧倒的なコストパフォーマンスで大量文書を捌き、Claudeの完璧な機密保持と推論力で最終成果物を仕上げる。単一モデルの弱点を相互補完する最強のパイプライン。

Tier 2 オンプレミス運用の技術的現実

The Hardware Reality (物理要件)



- ☐ 活性パラメータは軽量(8B/16B)だが、モデル全体の重み(763B)の展開が必須。
- ☐ NVIDIA H100/MI300X等を複数基束ねたデータセンター級GPUクラスタが要求される。
- ☐ 物理インフラの莫大な初期投資。

The Software Reality (MLOps要件)



- ☐ MITライセンス(無料)だが、標準的な対話テンプレートが同梱されていない。
- ☐ そのままロードすると出力フォーマットが破綻する。
- ☐ vLLM / SGLang上に、公式Python参照実装またはRust製ライブラリ「deepseek-recipe」を組み込む高度なMLOpsエンジニアリングが不可欠。

Immediate Action: 知財・ITリーダーのための即時行動計画



1. AUDIT (即時監査)

既存のAIパイプラインをテストせよ。APIのサイレント置換 (1017版) により、JSONパーサーや既存プロンプトが劣化 (デグレード) していないか実データでリグレーションテストを実施。



2. GOVERN (ガバナンス徹底)

社内通達を発行せよ。公式APIおよびWeb画面への「未公開発明・社内機密コード」の入力を厳格に禁止し、特許の新規性喪失リスクを物理的・制度的に遮断する。



3. BUILD (ハイブリッド構築)

夜間/週末バッチによる公開知財の超低コストスクリーニングを開始し、並行して「DeepClaude」ハイブリッドパイプラインのPoC (概念実証) に着手せよ。

AIの力は、そのリスクを完全に制御できるインフラの上にのみ成立する。