

DeepSeek-V4.1-Flash 徹底調査レポート

事実確認から知財・法務実務
への戦略的影響まで

[DATE]	2026.09.10
[AUTHOR]	Claude Fable 5.1

[EXECUTIVE SUMMARY] The Model

正体は「Flash」の名に反する 552B
の準フラッグシップ。CED新アーキ
テクチャによる構造的断絶。



[EXECUTIVE SUMMARY] Market Impact

API価格を60%削減（キャッシュヒット
入力 \$0.003 / 1M tokens）。
軽量マルチモーダルの低価格コモディテ
ィ化が競合を圧倒。



[EXECUTIVE SUMMARY] IP Strategy

日本の知財実務における最適解は「自己
ホストによる非機密文書の大量下読み」。
APIの画像圧縮と中国データ法制に警戒。

[WARNING]



情報の整理：リリースにまつわる「事実と誤認」

[MARKET CONFUSION]

「9月8日に正式リリースされた」

Sept 8

[MARKET CONFUSION]

「V4-Flashのマイナーチェンジ」
「視覚エンコーダは外付け」

Architecture

Sept 10 - 14

[CONFIRMED FACT]

コミュニティ限定の「48時間で失効するベータ版（1アカウント20同時実行制限）」の静かな開始。

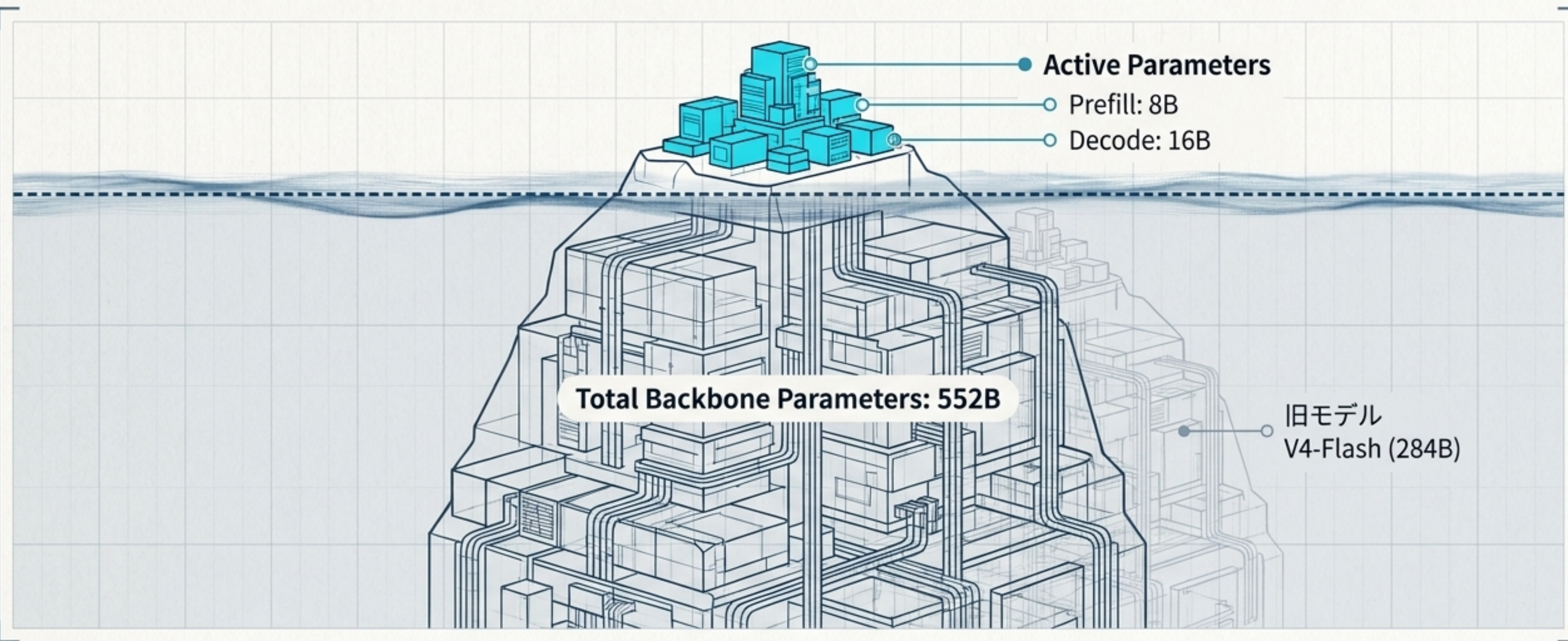
[CONFIRMED FACT]

V4系を完全に置換・統合する新アーキテクチャ。スクラッチ学習による「真のネイティブ・マルチモーダル」。

[CONFIRMED FACT]

9月10日正式公開（MITライセンス）。9月14日以降、既存V4 Proは強制的にV4.1-Flashヘルパーティング、実質引退へ。

モデルの正体：「Flash」を冠する巨大な”準フラッグシップ”



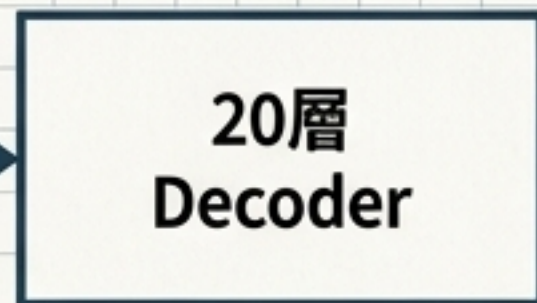
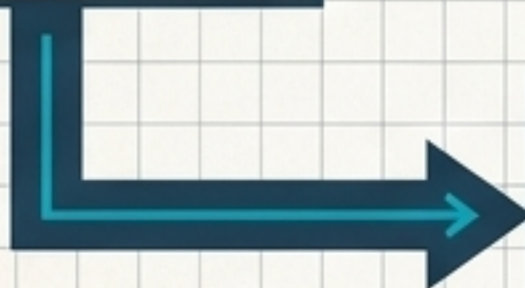
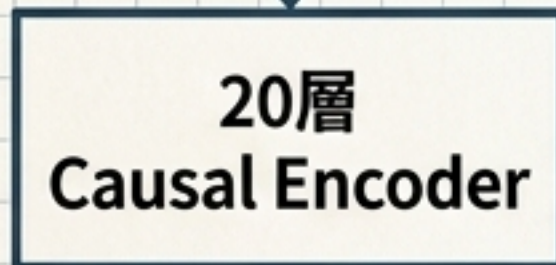
[ANALYSIS] The Quasi-Flagship Identity

V4.1-Flashは、今後登場が予告されている巨大モデル「V4.1-Pro」へのスケールアップを前提として設計された、新アーキテクチャ・ファミリーの「最小モデル (Base Unit)」に過ぎない。

メモリ革命：CEDとCSA2による限界突破

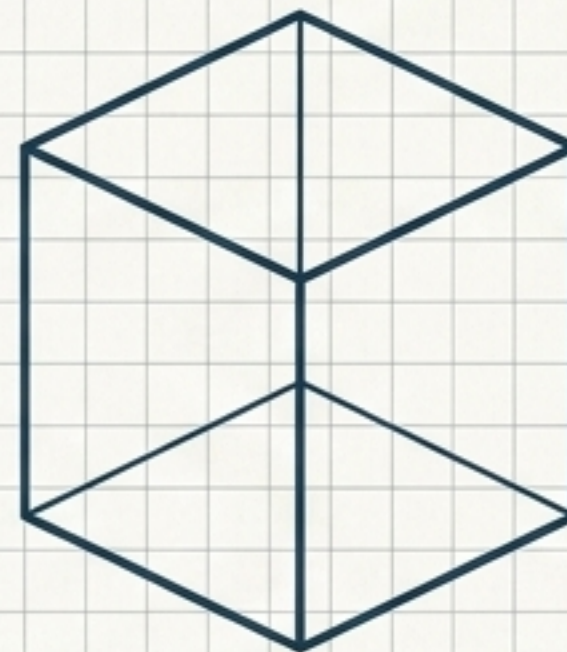
[CONFIRMED FACT] Causal Encoder-Decoder (CED)

Prompt Tokens

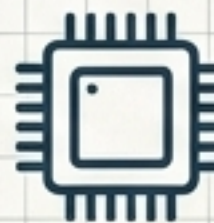


プロンプトトークンがエンコーダで止まり、デコーダへ直接射影される（YOCO着想）。Prefill時の計算量をほぼ半減。

[CONFIRMED FACT] Compressed Sparse Attention 2 (CSA2)



旧V4-Flash
KVキャッシュ



HBM要件
約1/4に削減



V4.1-Flash
(FP4 KV)



SSD要件
約1/8に削減

長文・マルチモーダル処理における圧倒的コスト優位の源泉。

パフォーマンス・マトリクス：勝てる領域、劣後する領域

[FRONTIER LEVEL] 優位・肉薄



Terminal-Bench **90.6**

※ 米国トップデル凌駕

DeepSWE



エージェント、コーディング、実務系長文処理においては、米国のトップモデル（Opus 5, GPT-5.6）に肉薄、あるいは凌駕。

[THE LIMIT] 劣後・限界点



Humanity's
Last Exam (HLE)

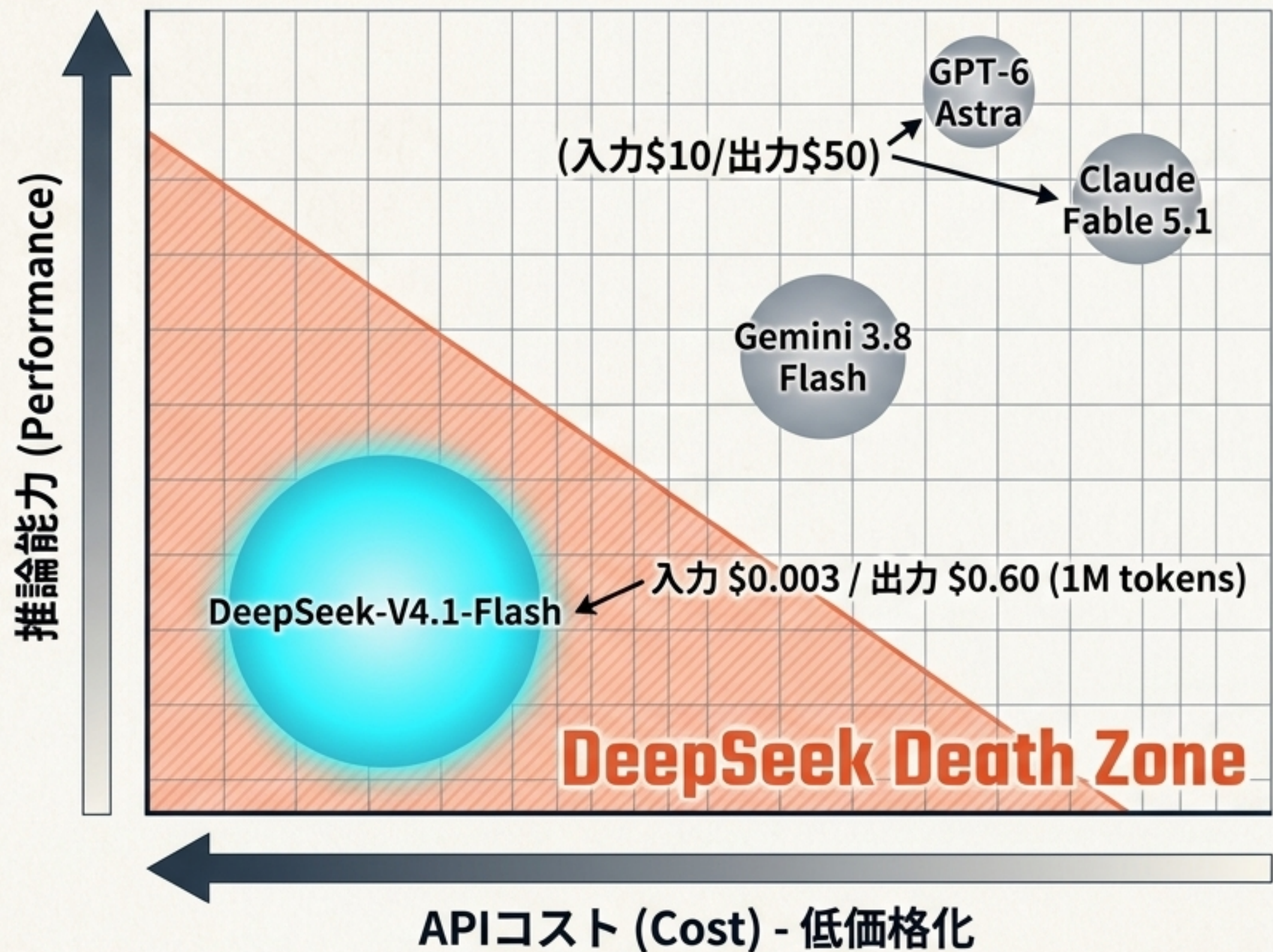
36.8 (vs Opus 5.0: 56.3)

最難関推論においては依然として米国フロンティアモデルに明確に劣後。

[ANALYSIS]

「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、米中AI開発競争における性能混在の現実。

価格破壊と「DeepSeek Death Zone」の出現



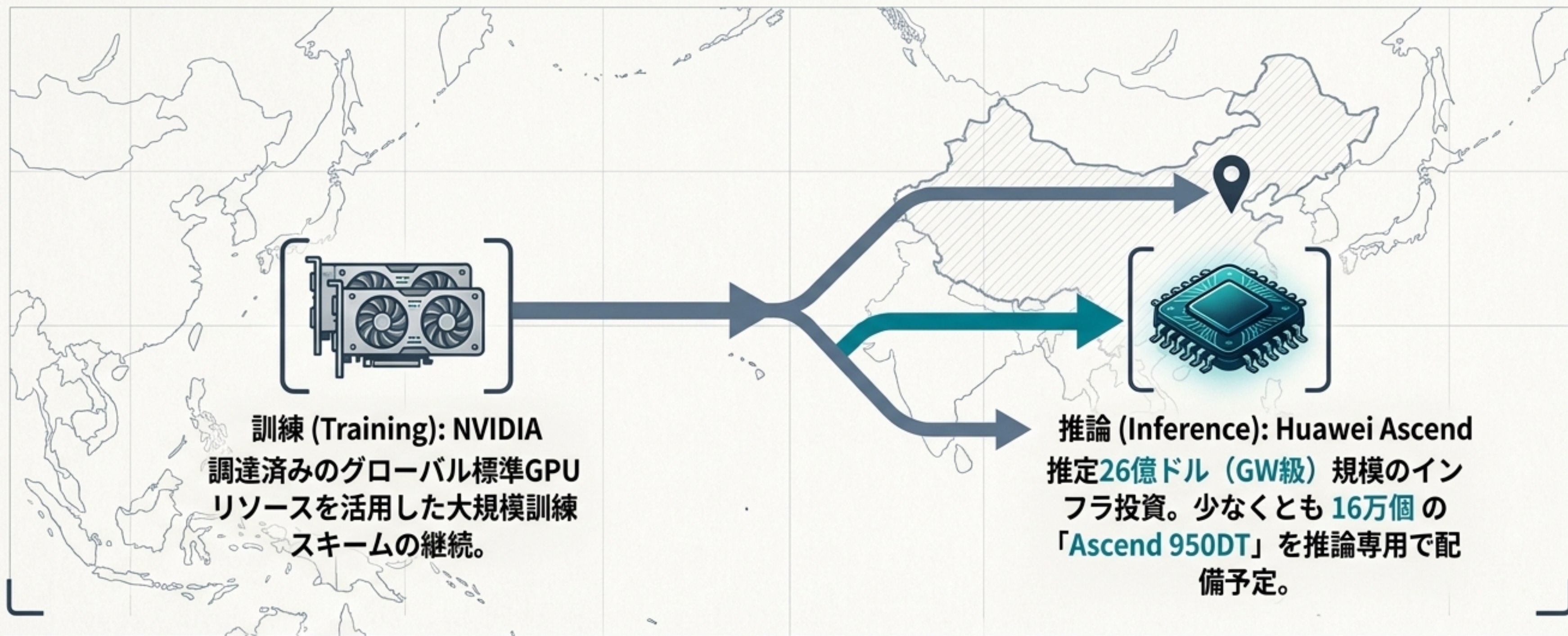
[WARNING] Market Impact

従来比約**60%**の価格破壊。

「十分に良い性能を**超低価格**で提供する」**コモディティ化戦略**の脅威。

香港市場でZ.aiやMiniMaxなど競合AI企業の**株価が8%超急落**。

「インフラと地政学：米中分業と国産推論網」



【ANALYSIS】 米国の輸出規制下においても、推論フェーズを国産チップのみでスケールさせるエコシステムが実証段階に突入。STAR MarketでのIPO準備（評価額750億ドル）がこの巨大推論網の構築を支える。

日本の知財・文書実務における「可能性と罠」

魅力 (Opportunity)

1Mコンテキスト × ネイティブ画像理解 × 圧倒的
低コスト

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（一次スクリーニング）に極めて高いポテンシャル。

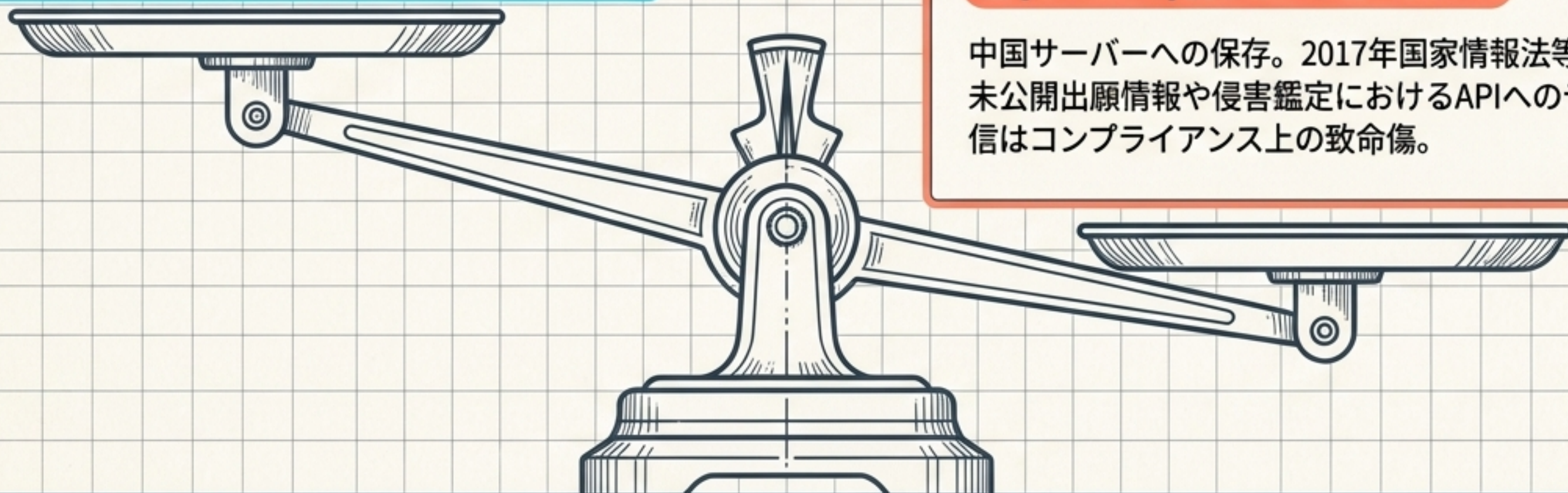
リスク (The Traps)

⚠ [WARNING] 罠① APIの画像制限

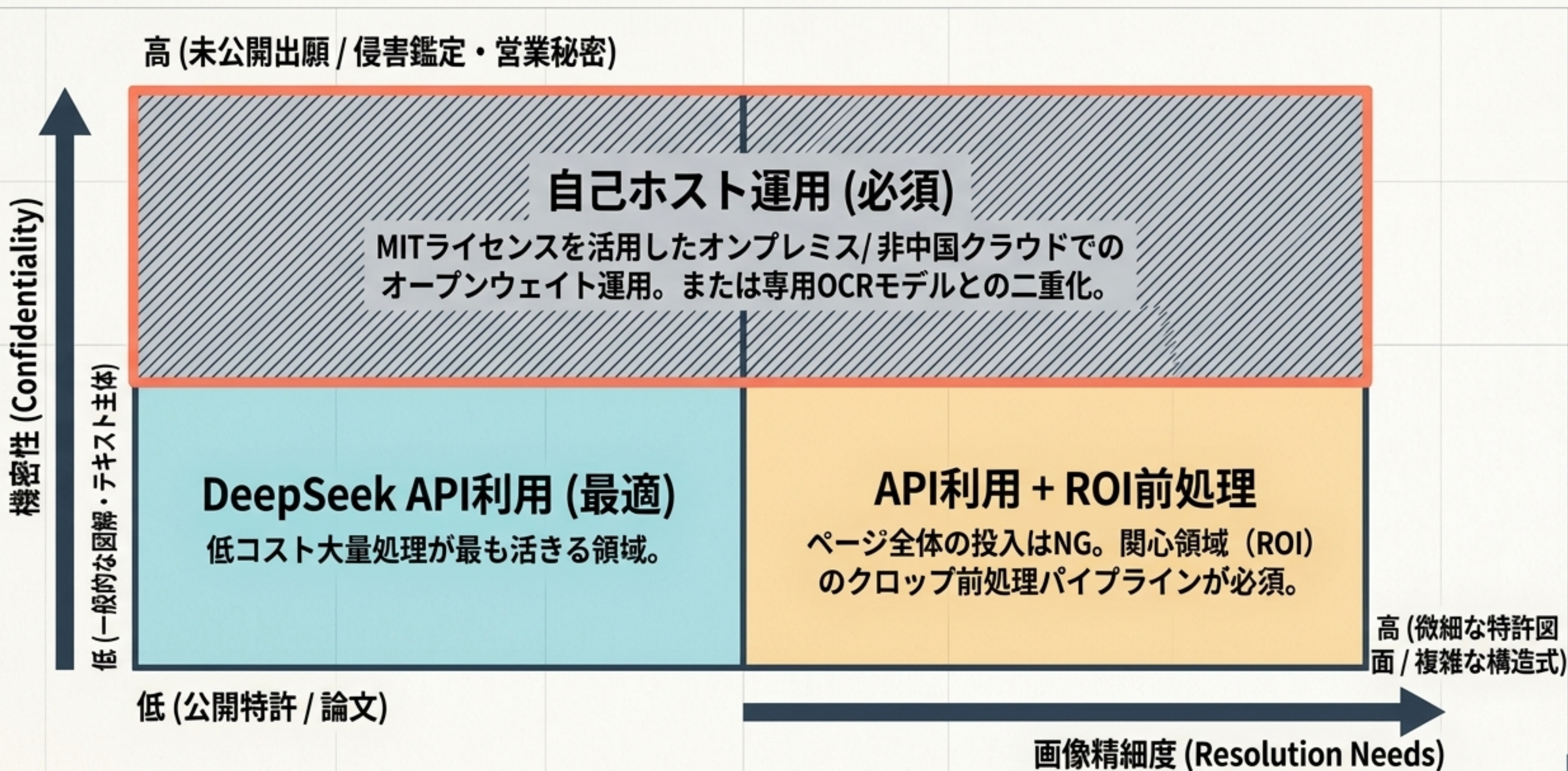
API経由の画像は実質「800×800相当・384トークン上限」に強制圧縮。特許図面の微細な符号や複雑な化学構造式では致命的な情報欠落の恐れ。

⚠ [WARNING] 罠② データガバナンス

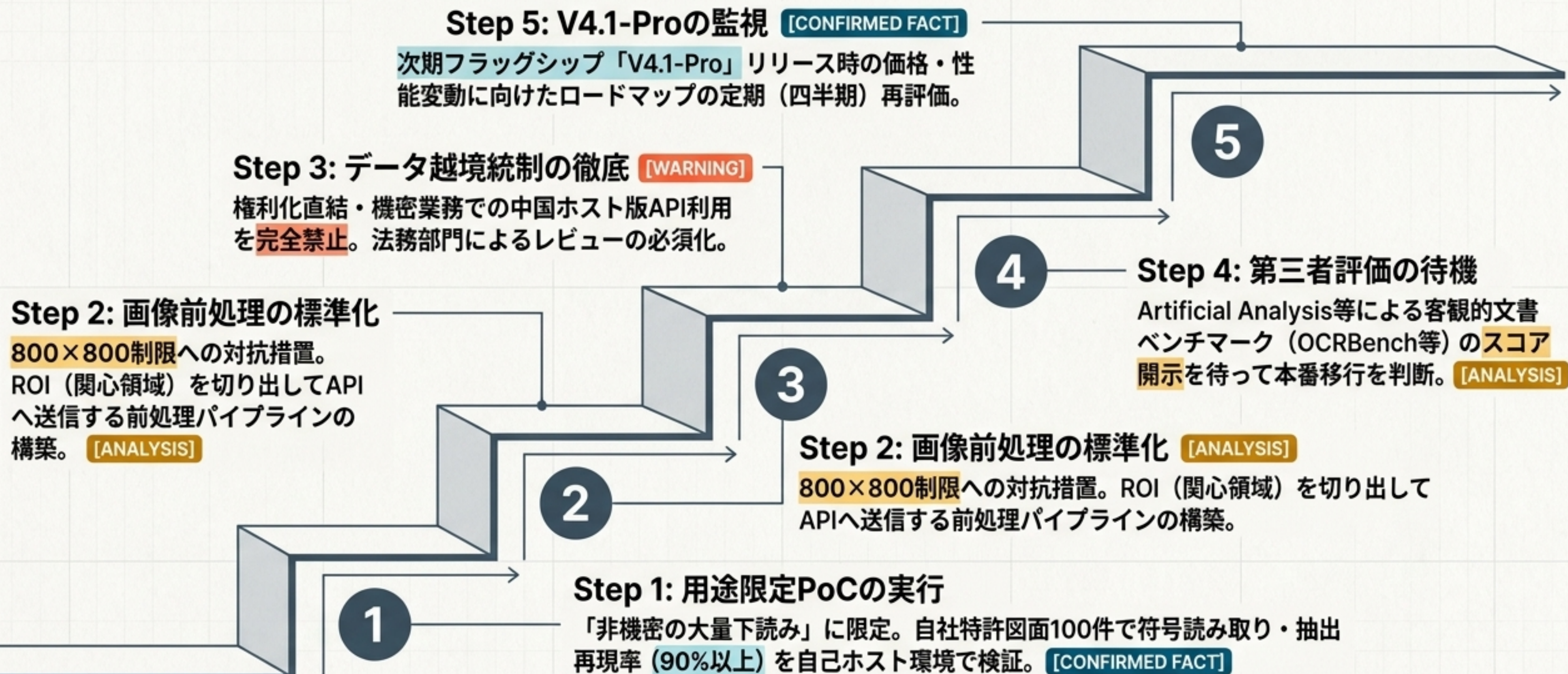
中国サーバーへの保存。2017年国家情報法等の適用。未公開出願情報や侵害鑑定におけるAPIへのデータ送信はコンプライアンス上の致命傷。



「戦略的利用ガイドライン (Strategic Usage Matrix)」



企業向け 推奨アクション5ヶ条 (5-Step Action Plan)



本レポートは2026年9月10日時点の情報に基づく。独立検証未了のベンダー公表値およびコミュニティ実測値を含む。

標準的な文書/視覚ベンチマーク（OCRBench, ChartQA, MathVista等）は公式非開示であり、実務精度の客観的評価は継続的な検証を要する。

IPO評価額（750億ドル）やAscend調達規模（16万個）等のインフラ動向は報道ベースであり、監査済みの開示情報ではない。

価格やAPI制約は短期間で変更される実績があるため、実装時には公式ドキュメントの再確認を推奨する。

一部技術レポート本文は権限エラーのため要約に基づく。

