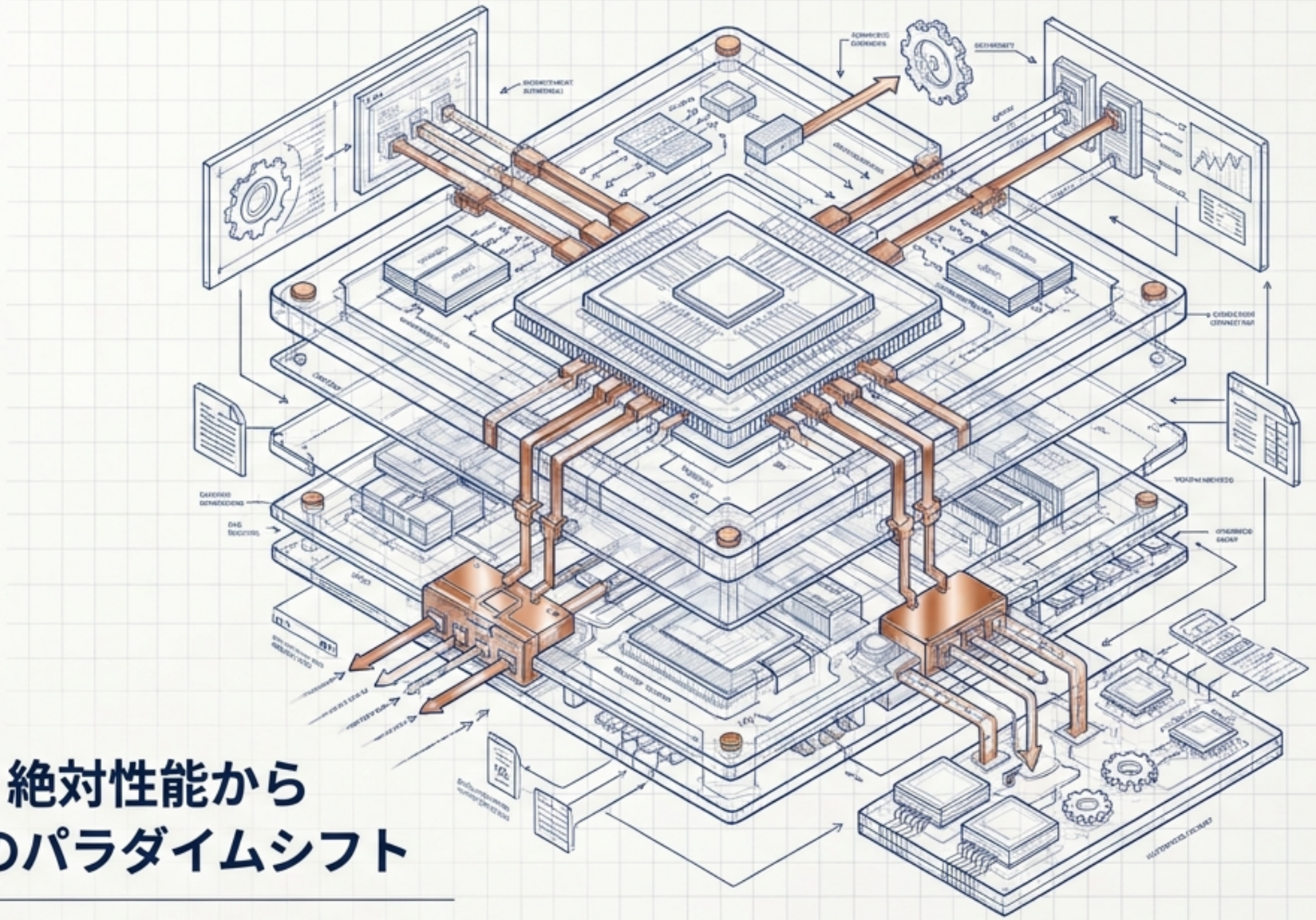


Qwen4-Max 徹底解剖：絶対性能から「エージェントROI」へのパラダイムシフト

日本企業のための戦略的評価レポート（2026年11月16日正式リリース版対応）

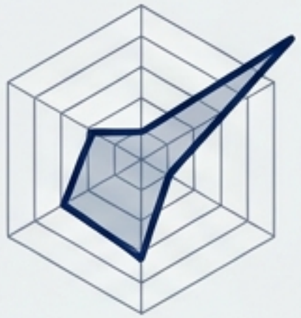
Gemini Notebook

The Bottom Line：Qwen4-Maxがもたらす3つのインサイト



削ぎ落とされた「エージェント専用機」

アクティブパラメータを**25%削減**（710億）。暗黙キャッシュ単価**$0.075**により、1タスクあたりの実行コストを前世代比で**約60%削減**。



「万能」ではなく「特化」

総合力（**AA指数43**）や自律コーディングではClaude Opus 5.5（59）に及ばない。一方、**専門文書**（法務・生物）と**ツール利用に極端な強み**（スパイク）を持つ。

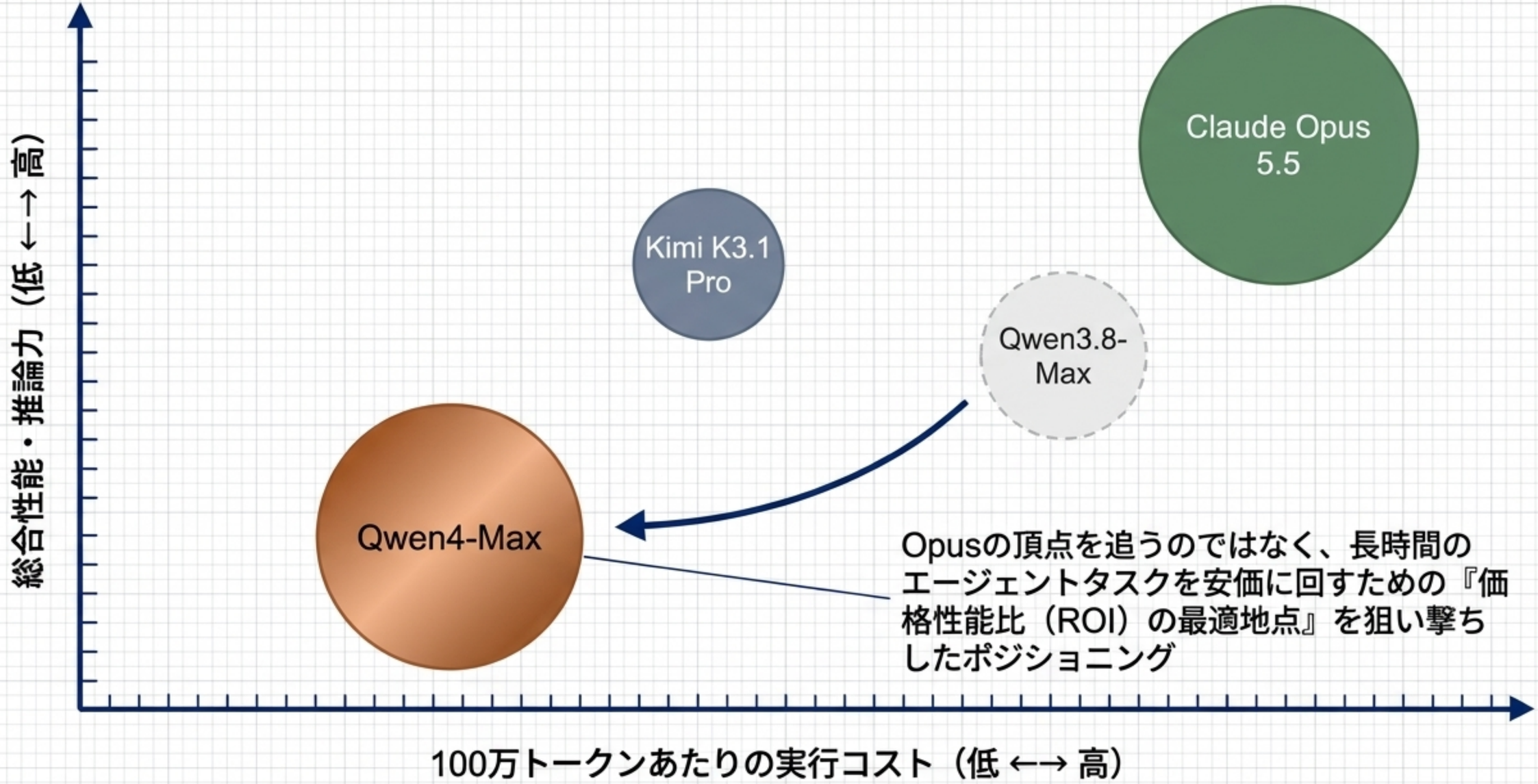


導入の壁と戦略的待機

「**東京リージョン**」でもデプロイ範囲は「**Global**（国外推論の可能性）」。機密業務には、今後登場する**小型オープン版**（27B/Flash）のローカル環境での活用が現実解。

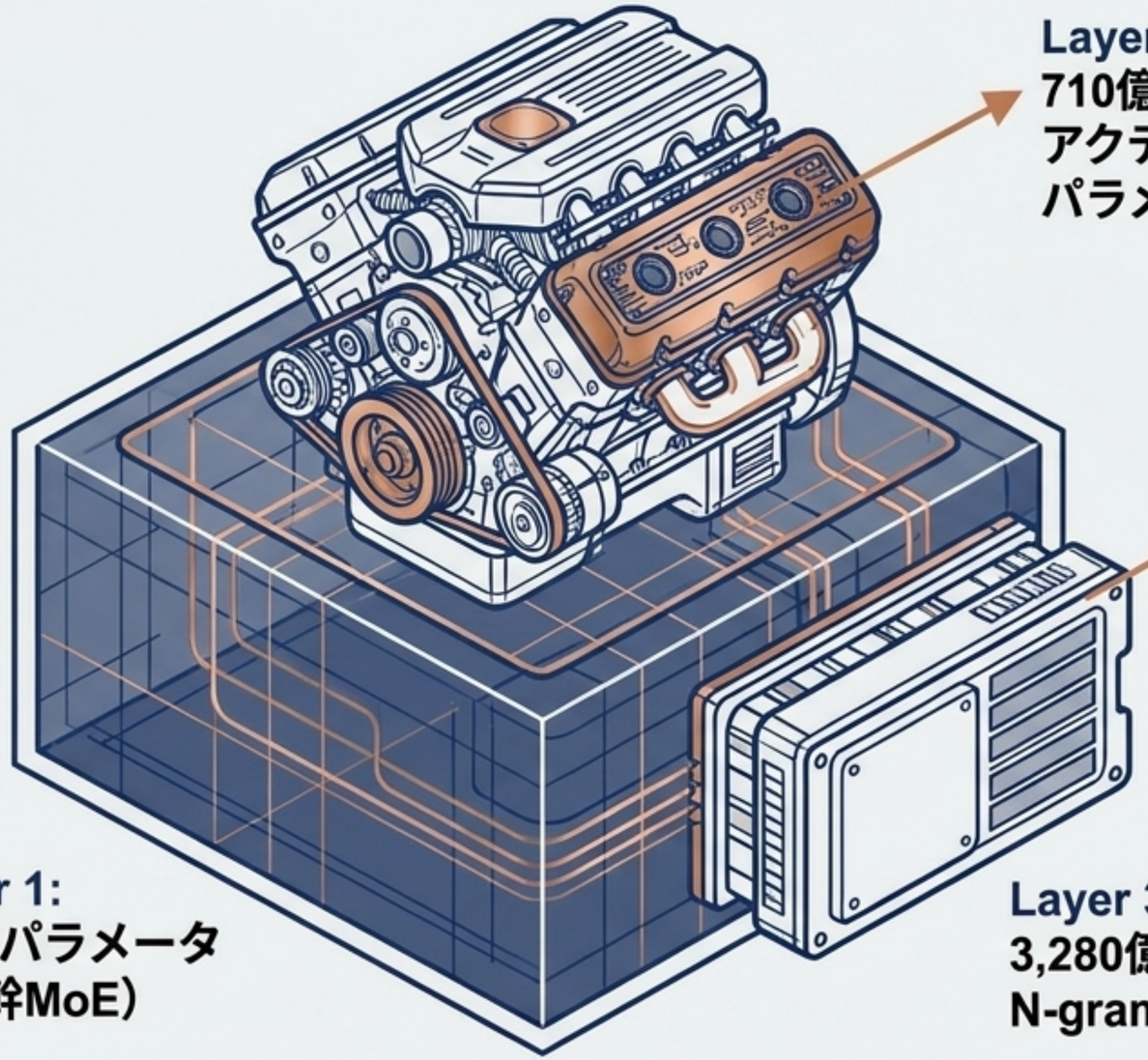
Gemini Notebook

競争軸の転換：絶対性能の追求から、エージェント業務の量産体制へ



Gemini Notebook

アーキテクチャの解剖図：低単価を生み出す「3層の精密設計」



Layer 2:
710億
アクティブ
パラメータ

GPUメモリ制約下でも効率的に知識を拡張する設計。

Layer 1:
2.4兆パラメータ
（主幹MoE）

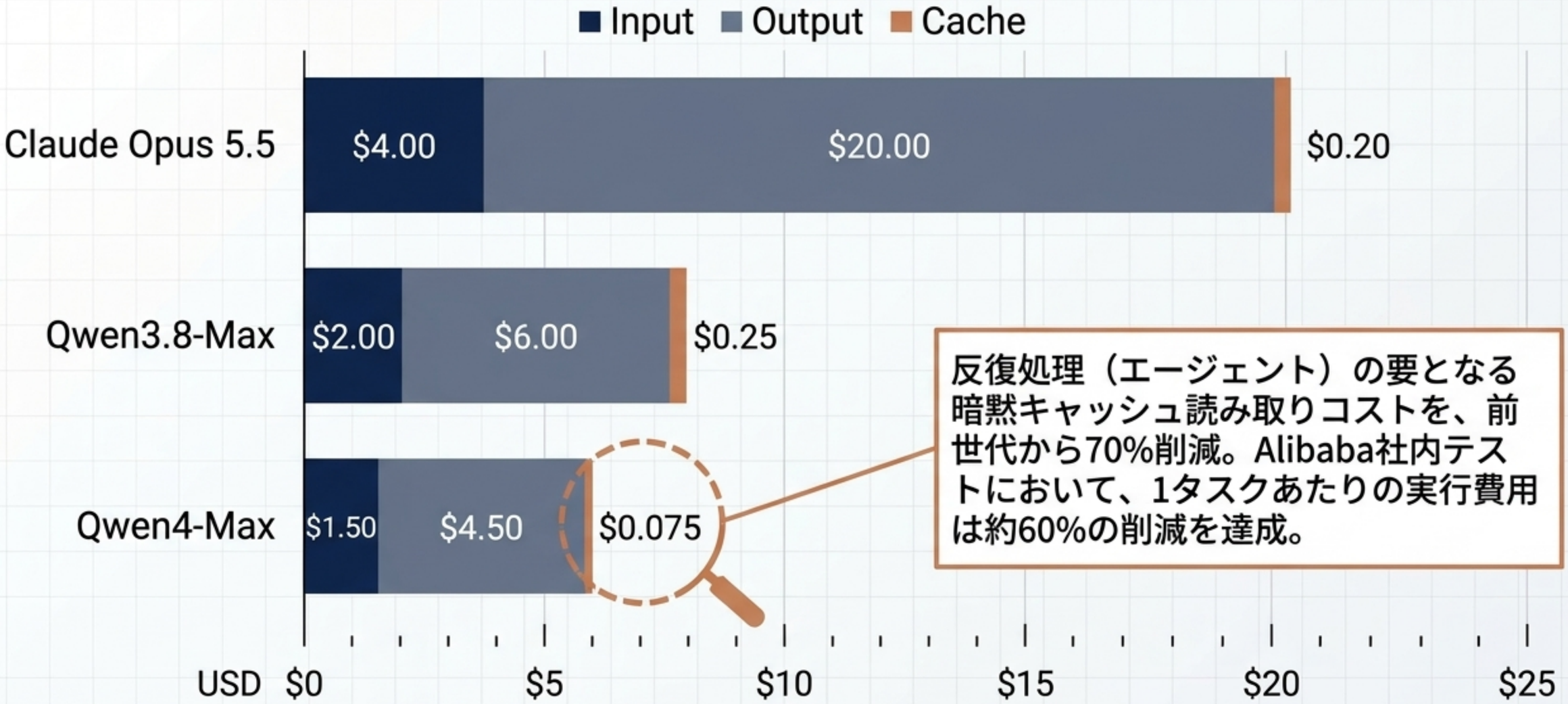
Layer 3:
3,280億
N-gram埋め込み

前世代の950億から約25%削減。これが消費トークンと単価劇的低下の源泉。Gated DeltaNetとQSAのハイブリッド構成。

絶対的なサイズ（2.4T）を維持しつつ、一度の処理で動く部分（710B）を極限まで絞り込むことで、高速・低コストなエージェント連続稼働を実現。

Gemini Notebook

エージェント業務の量産を可能にする破壊的コスト構造



反復処理（エージェント）の要となる暗黙キャッシュ読み取りコストを、前世代から70%削減。Alibaba社内テストにおいて、1タスクあたりの実行費用は約60%の削減を達成。

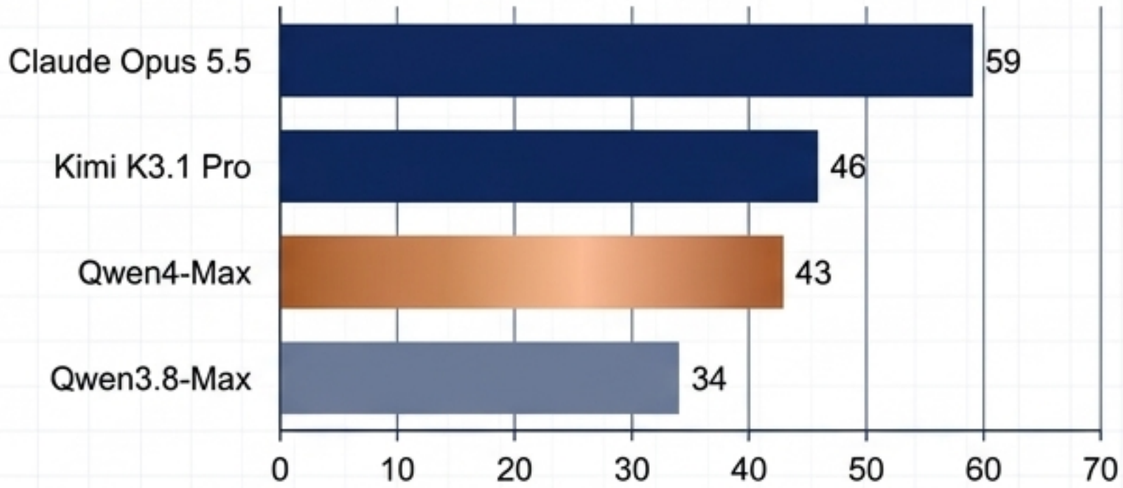
Gemini Notebook

公式発表 vs 第三者評価：総合力における現在地

公式の主張

- 独自の16パネル比較において、「DeepSWE」など6項目でClaude Opus 5.5を凌駕。
- Agents' Last Exam（46.7）、Work360（52.69）で大幅な進歩をアピール。

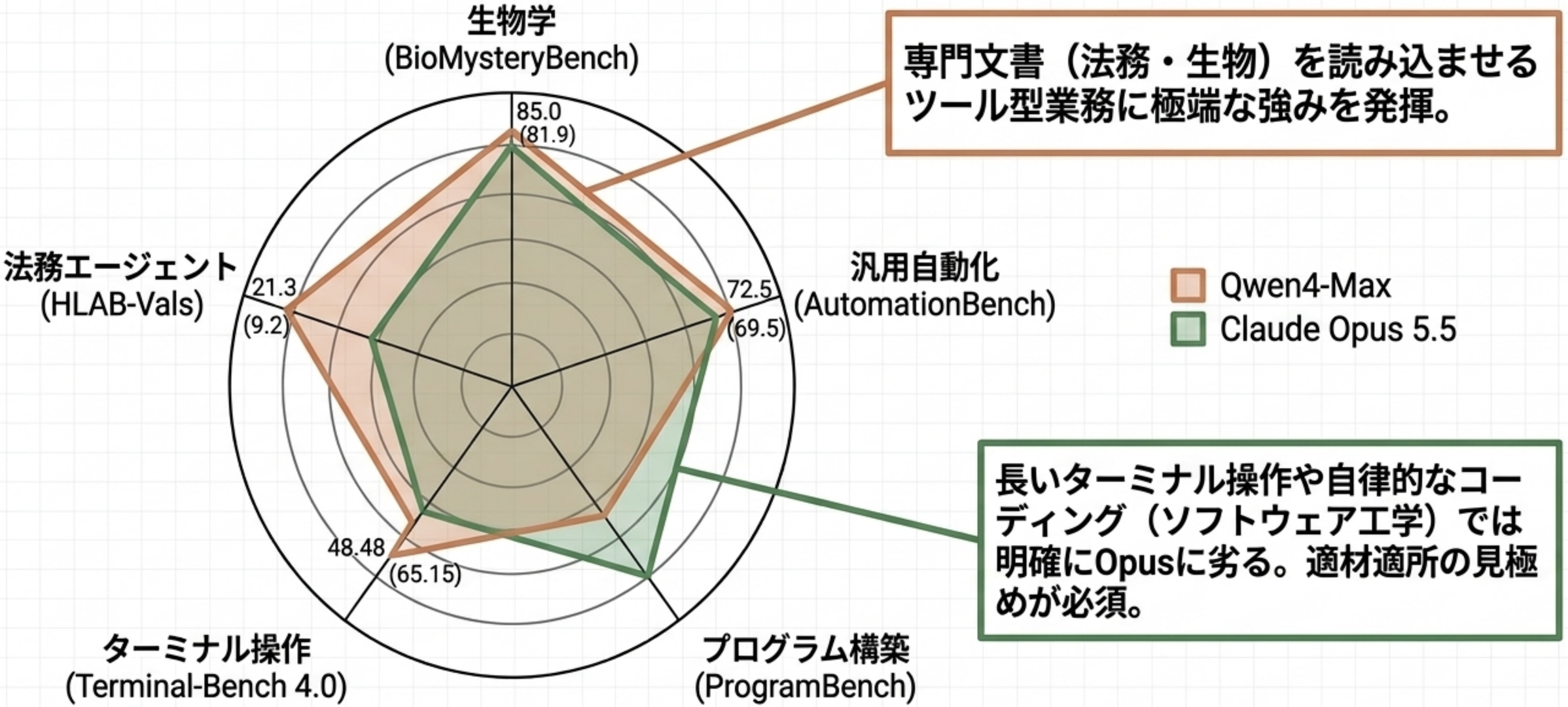
第三者評価（Artificial Analysis 指数）



Vals Index等では中国モデルトップの評価を獲得。しかし、総合力では依然としてClaude OpusやKimi K3.1の域には達しておらず、『同価格帯の中央値（41点）をわずかに上回る優等等生』というのが客観的立ち位置である。

Gemini Notebook

能力の「スパイク（尖り）」：ドキュメント特化型としての生存戦略



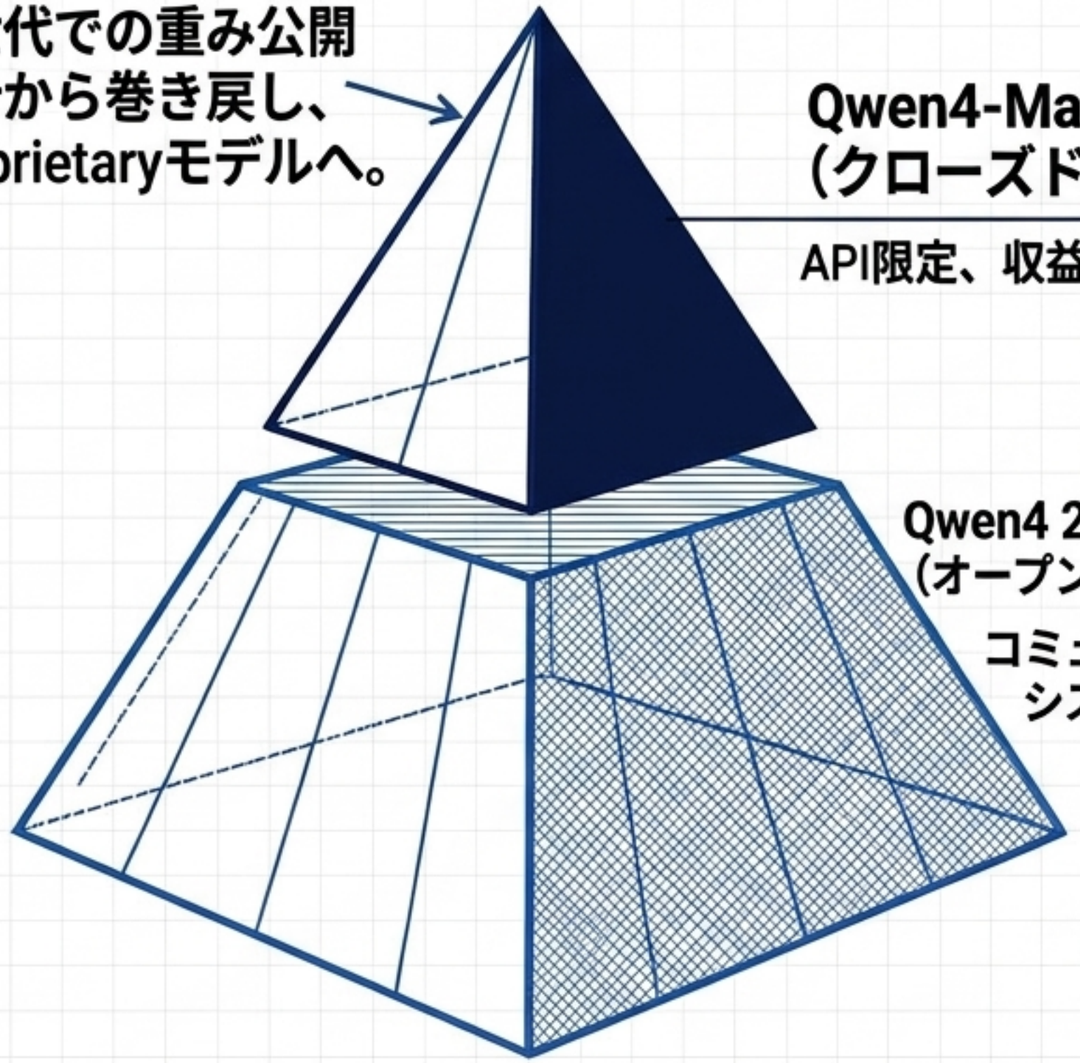
専門文書（法務・生物）を読み込ませるツール型業務に極端な強みを発揮。

長いターミナル操作や自律的なコーディング（ソフトウェア工学）では明確にOpusに劣る。適材適所の見極めが必須。

Gemini Notebook

クローズドへの回帰と二本柱のエコシステム戦略

前世代での重み公開方針から巻き戻し、Proprietaryモデルへ。



Qwen4-Max（クローズド）

API限定、収益化の要

Qwen4 27B / Flash等（オープンウェイト予定）

コミュニティとエコシステムの維持

戦略的背景（GPU制約の示唆）

- N-gram埋め込みや活性パラメータ削減のアーキテクチャは、「限られたメモリでの効率化」を意図。
- Alibabaの設備投資は前年比75%増だが、GPU不足によるキャパシティリスクが指摘されている。

結論：軽量モデルのオープン化で裾野を広げつつ、最上位は限られた計算資源で収益を最大化する現実路線。

Gemini Notebook

【比較マトリクス】日本企業のための自社環境モデル使い分け戦略

	コスト効率（ROI）	適した主要タスク	データガバナンス リスク	実行環境の柔軟性
Qwen4-Max (API)	◎ (最高)	専門文書要約・抽出エージェント	⚠ (高)	△ (クラウドAPI限定)
Claude Opus 5.5	△ (高額)	自律的コーディング・複雑な推論	○ (エンタープライズ契約時)	△ (クラウドAPI主体)
Qwen4 小型 (27B等/将来)	○ (インフラ依存)	機密性の高い社内業務	◎ (最小)	◎ (オンプレ・ローカル可)

すべてを1つのモデルで解決するフェーズは終わった。タスクの『複雑性』と『データの機密性』に基づくマルチモデル・ルーティングがROI最大化の鍵となる。

Gemini Notebook

データ・主権リスクの評価：「東京リージョン」の死角



公式ドキュメント上の事実：
「東京を含む4リージョンのデプロイ範囲は『Global』と明記されている。」Token Planの推論もグローバル実行。

法務・知財担当者向けチェックリスト

- ☐ 東京リージョンのAPI利用でも、国外で推論処理される可能性を前提とする。
- ☐ 自社環境へのデプロイやファインチューニングは現在非対応。
- ☐ 中国企業のサービスであることに伴う経済安全保障・データ保持ポリシーの個別審査が必須。

Gemini Notebook

Deep Dive：知財・法務領域への応用検証と限界

ポテンシャル (100万トークンの活用)



広大なコンテキストウィンドウを活用し、複数の特許明細書や引用文献を一度に読み込ませ、要約や差分抽出を低コストで一括処理する用途に極めて適している。（HLAB-Valsスコアの突出）



限界とリスク (判断力の壁と機密性)



Legal Research BenchではOpus 5.5（50.48%）を下回る41.35%。法的な最終判断や深い調査を単独で任せる水準にはない。また、出願前の発明情報（未公開の機密情報）を「Global」デプロイのAPIに入力することは致命的なリスクを伴う。

結論：『大量文書の下準備・整理』に特化させ、最終判断と機密情報入力は切り離すワークフロー設計が必要。

Gemini Notebook

5つの戦略的提言（Actionable Recommendations）

本番移行前のROI再評価
トークン単価ではなく、「採用できた成果1件あたりの総費用」でOpus等と比較する。

エージェント型コーディングの限定的利用
Terminal-Bench等の弱さを考慮し、リポジトリ規模の開発は米国旗艦モデルを優先。Qwenは補助に留める。

データ所在の契約前確認
機密データ投入前に、東京リージョンにおけるデータ保持と学習利用の有無を書面で確認する。

Preview版からの即時移行
11月25日終了のプレビュー版APIから正式版への切り替えと、出力上限変更（393K→262K）へのシステム対応を行う。

機密タスクの「戦略的待機」
知財・機密業務については、将来公開されるオープン版（27B/Flash）のローカル環境での評価にリソースを集中させる。

Gemini Notebook

留意事項とソースコンテキスト（Disclaimers）

情報の鮮度と揺らぎ
本評価は2026年11月16日の正式リリース当日のデータに基づきます。Arenaのスコア等はプレビュー版のデータを含む暫定値であり、費用削減率（58%〜62%）などの数値には通貨や計算方法による揺らぎが存在します。

情報源の透明性
Qwen公式ブログ本文など、一部の一次ソースにはアクセス制限が確認されています。本レポートでは、信頼性の低い暗号資産系ニュースやAI生成サイトの未検証データは排除し、複数の独立した報道で裏付けられた事実のみを認定しています。

非公開状況に関する優先基準
前世代モデルの重み公開に関する情報源Cの相違については、Alibabaが米国SECに提出した公式文書（Form 6-K）の記述を一次ソースとして優先しています。

Gemini Notebook