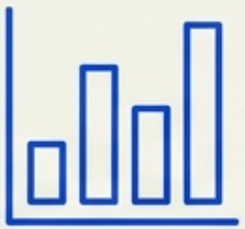
MiMo V2.6 Pro: 性能評価とエンタープライズ導入診断

メーカー公表値と第三者評価に基づく、実務適用の現実と日本市場における潜在的リスク

調査基準日: 2026年9月22日
対象モデル: MiMo-V2.6-Pro / Flash / 9B
分析フレームワーク: GPT-6 Astra 調査レポート準拠

Executive Snapshot



オープンウェイト最高峰

Artificial Analysis総合指数「46」を獲得。上位の非公開モデルに肉薄する基本性能。

エージェント作業の飛躍

Opus 5を一部凌駕するツール呼び出し能力と業務自動化スコア。

キャッシュによる価格破壊

コンテキストキャッシュ利用時、 入力API単価が約99%下落（$0.0036/1Mトークン）。

日本語実務と幻覚リスク

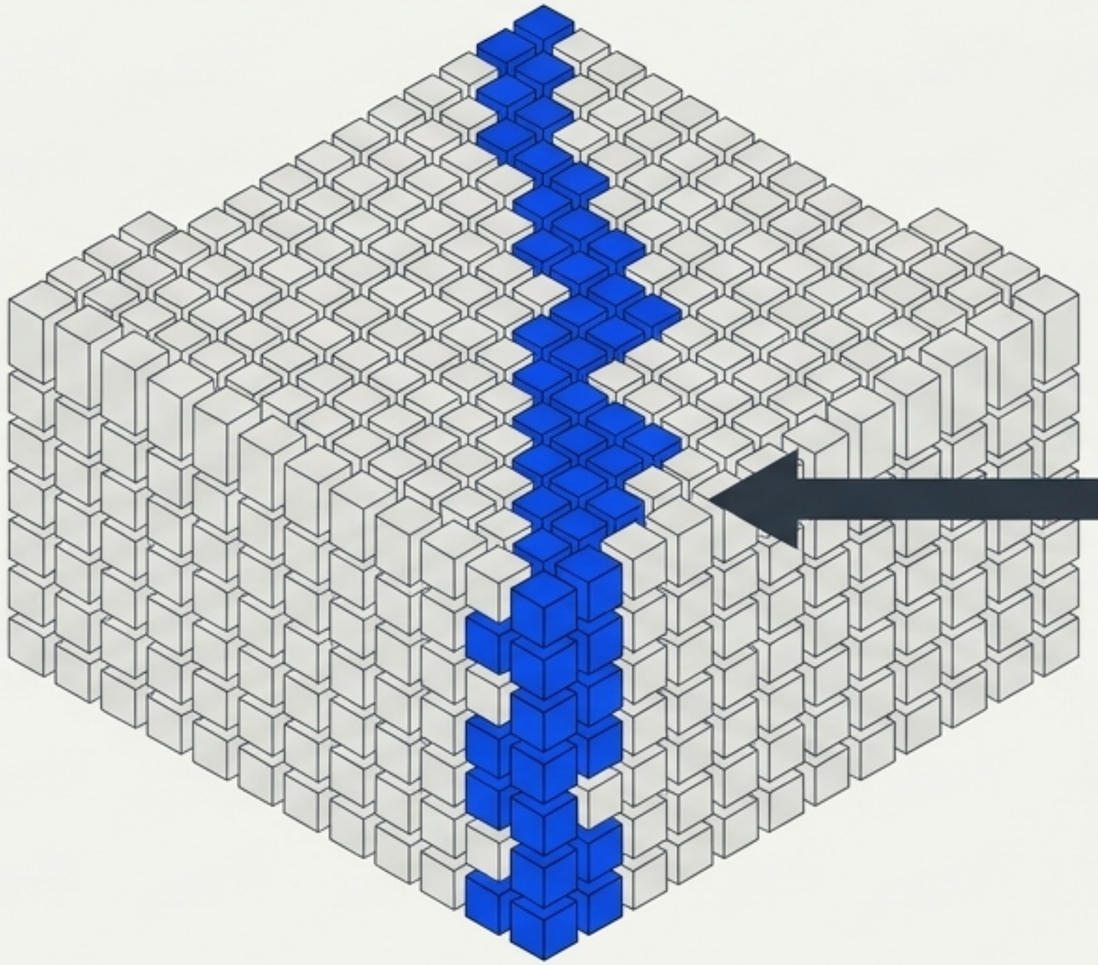
未知の問題に対するハルシネーション率が40.6%へ悪化。
日本語の専門領域（法務・特許）は未検証。

ラインナップの戦略的位置づけ

MiMo-V2.6-Pro (旗艦モデル)	MiMo-V2.6-Flash (コスト重視モデル)	MiMo-V2.6-Distill-9B (エッジ・ローカル)
規模: 総1.02兆 / アクティブ420億パラメータ	**規模**: 総3090億 / アクティブ150.0億パラメータ	**規模**: 90億パラメータ（Qwen3.5ベース）
位置づけ: 最大性能・複雑な推論・エージェント用途	**位置づけ**: 反復タスク・圧倒的な低API単価	**位置づけ**: ローカル環境での検証・軽量タスク
提供: API / 公開ウェイト	**提供**: API / 公開ウェイト	**提供**: 公開ウェイト（教師あり追加学習のみ）

アーキテクチャの優位性：Mixture of Experts (MoE)

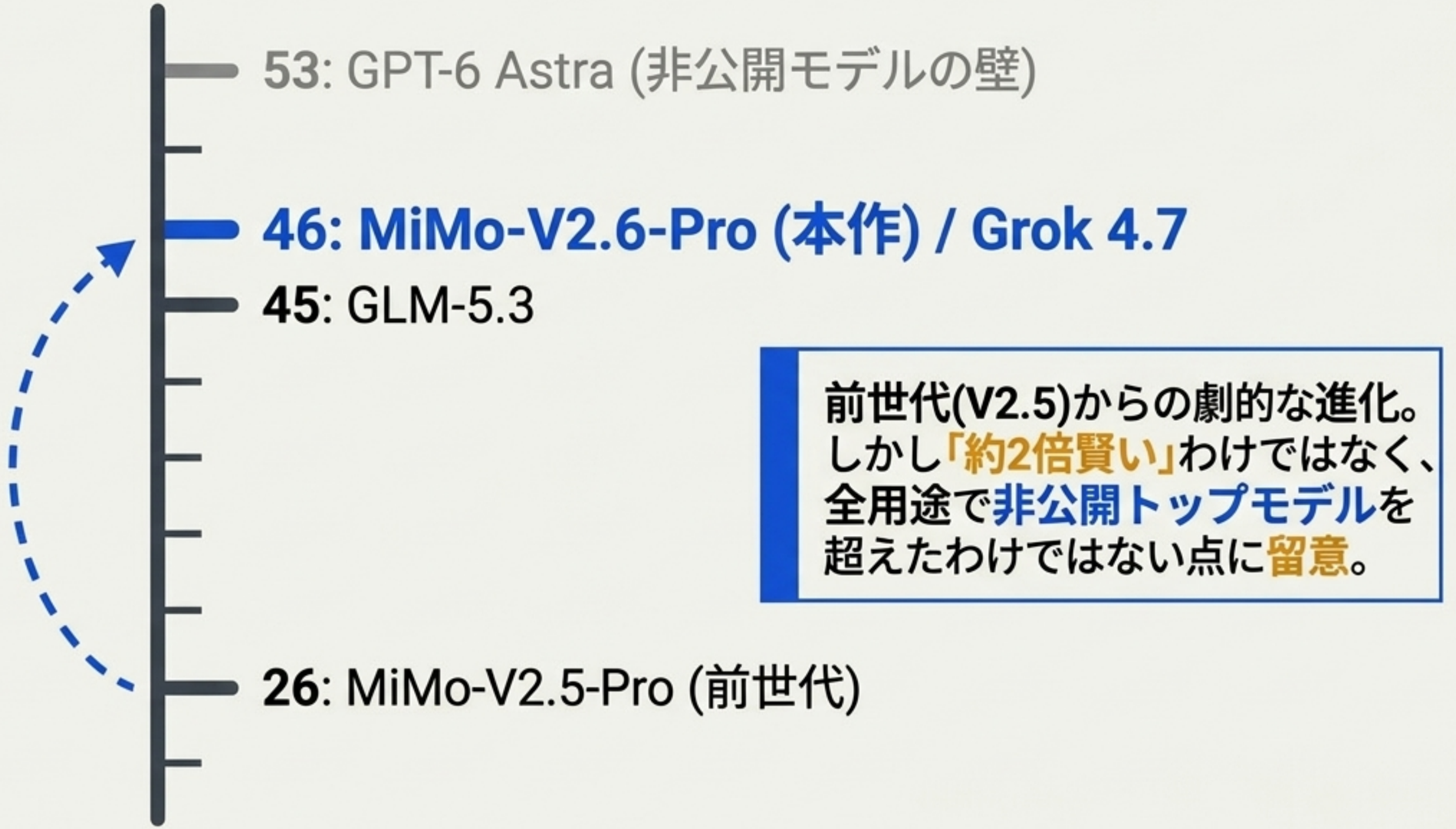
総パラメータ数: 約1.02兆
（保持する知識と能力の総量）



アクティブパラメータ: 420億
（1トークン生成時に実際に計算される量）

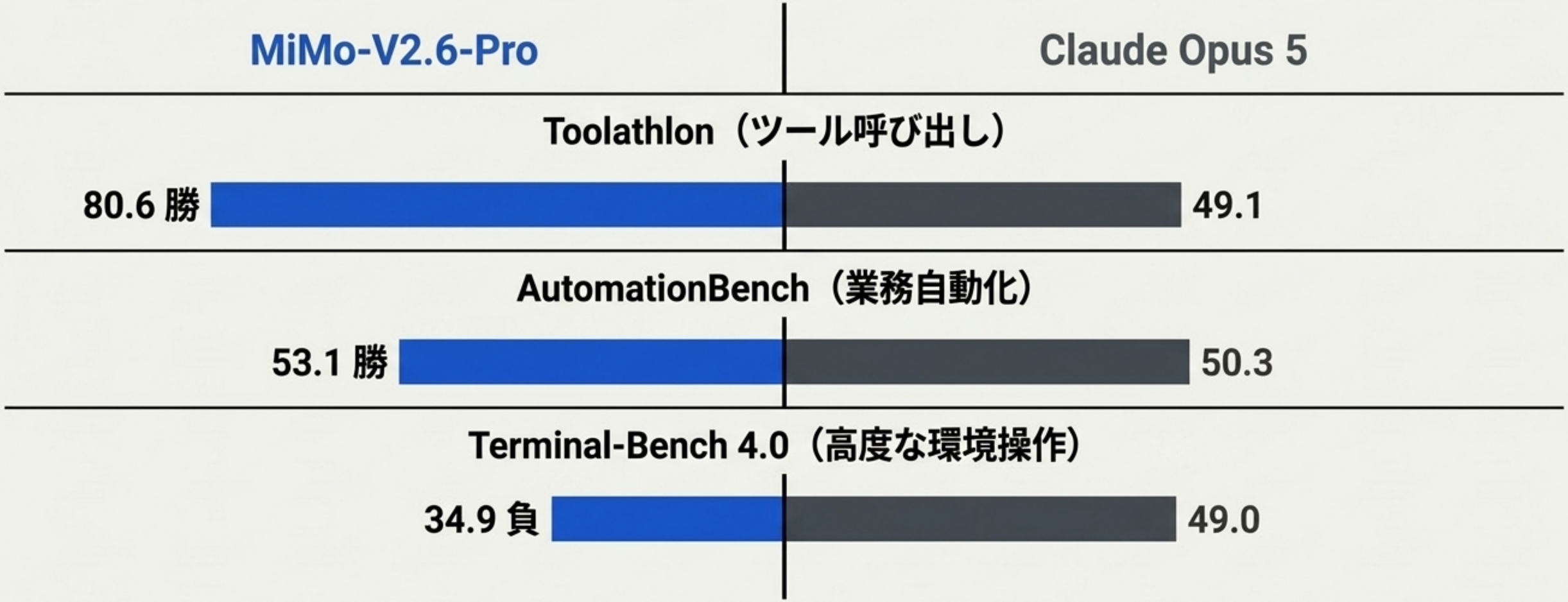
コンテキスト長: 100万トークン（最大出力 128K）	MoE構造により、巨大な知識ベースを持ちながら、推論時の計算コストと遅延を劇的に削減。

第三者評価：公開ウェイトにおける最高到達点



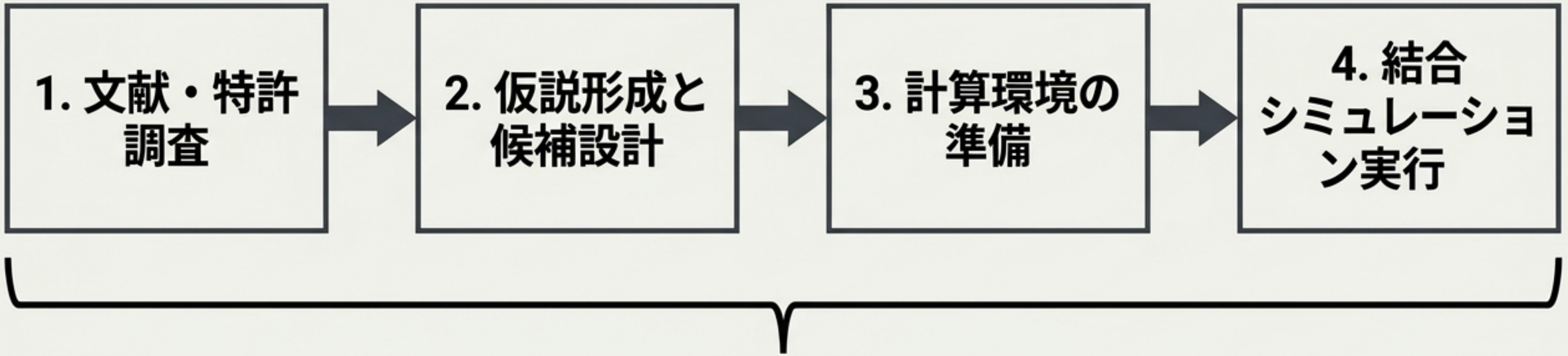
前世代(V2.5)からの劇的な進化。
しかし「約2倍賢い」わけではなく、
全用途で非公開トップモデルを
超えたわけではない点に留意。

メーカー公表値の解剖：Opus 5との実力差



作業の分解とツール実行能力（エージェント力）ではOpus 5を凌駕するが、より複雑な環境（Terminal-Bench 4.0等）での問題解決能力には依然として差が残る。

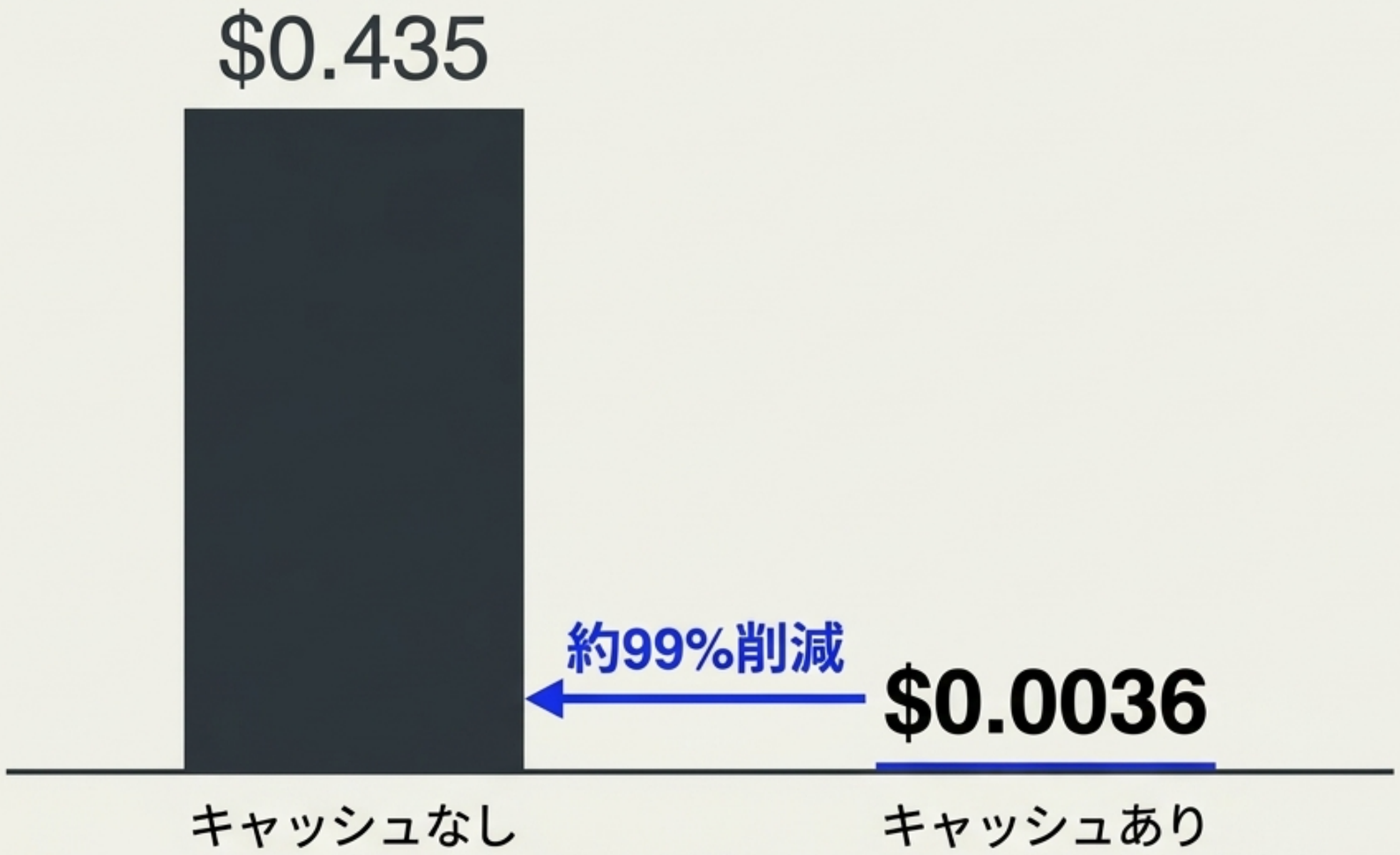
エージェント機能の実証：材料研究（PFAS捕捉MOF）の自律化



従来1か月の工程を「2〜3日」へ短縮（メーカー推計）

注意：現時点ではシミュレーション上の成果であり、実験室での物理的な検証を待つ段階。未知の難問を完全自律で解決したわけではない。

APIエコノミクス：キャッシュがもたらす価格破壊

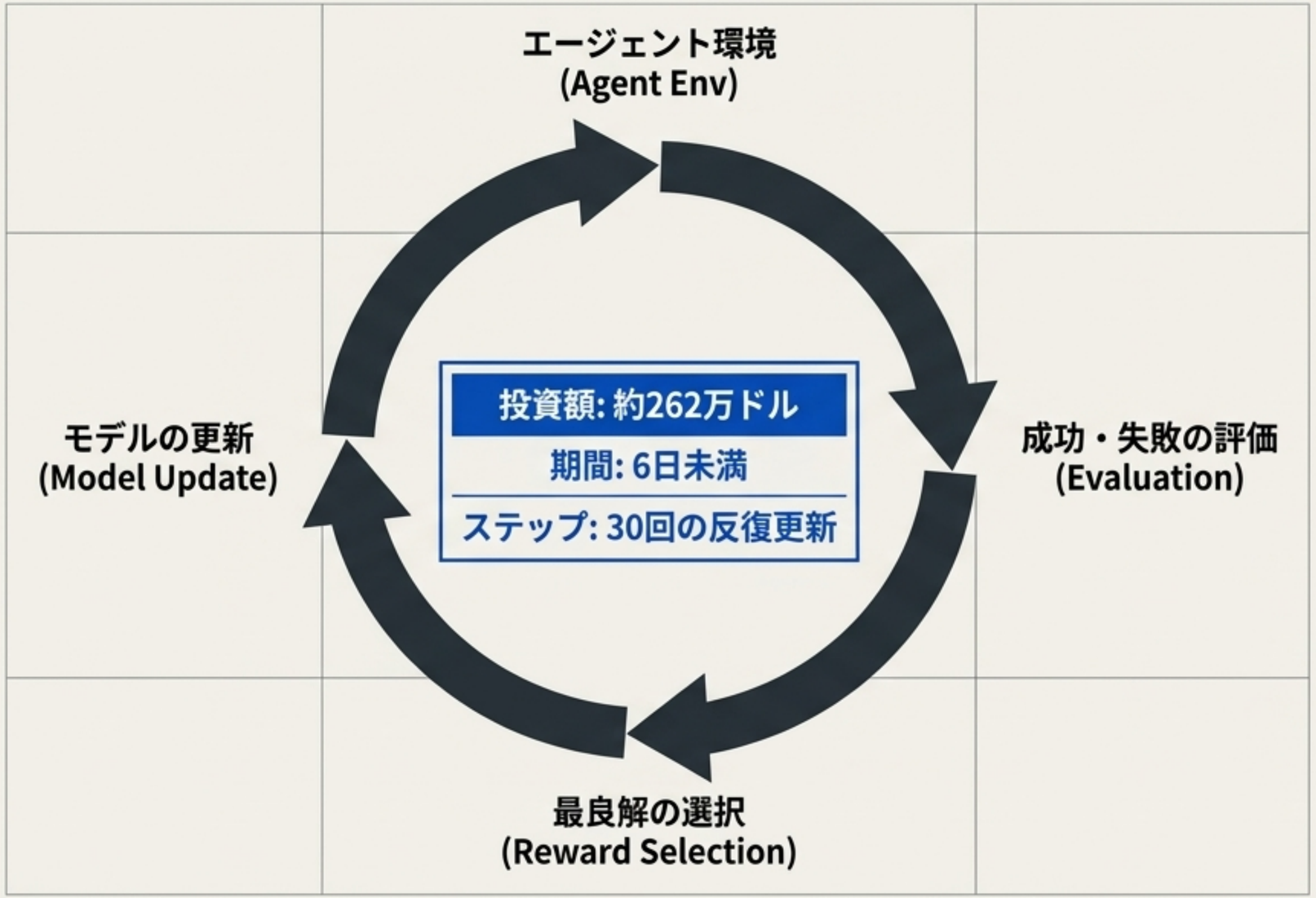


Pro-UltraSpeed API

単価10倍（入力$4.35）で最大20倍の出力速度を謳うオプション。
※全処理の完了時間が1/20になる保証ではない。

*参考出力単価: $0.87

自律的改善の裏側：強化学習（RL）エンジンの力技



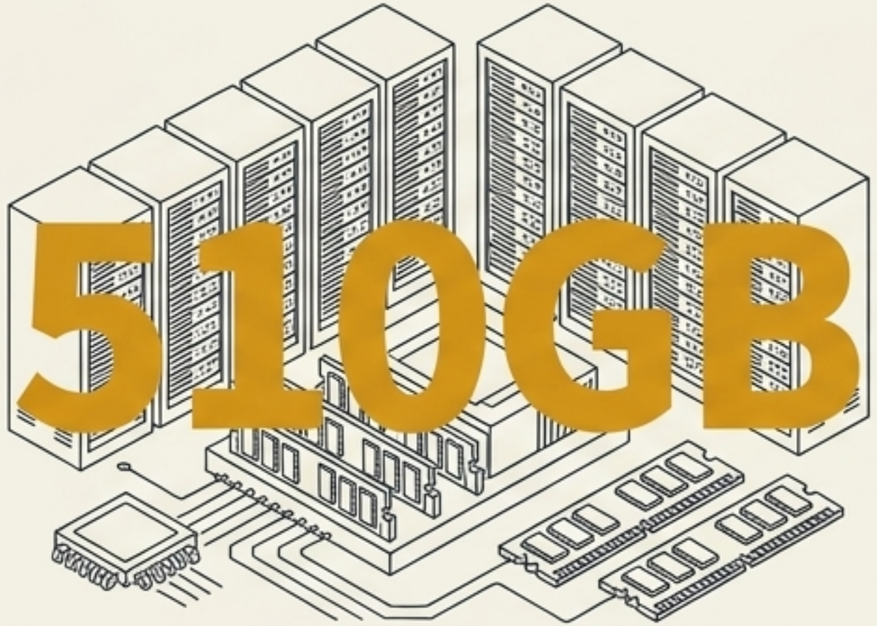
手法:

複数のエージェント実行環境（コード、視覚、サイバー等）での訓練。成功・失敗だけでなく、成功した解法同士も比較し品質を評価。

結論:

汎用人工知能（AGI）の実現ではなく、「人が設計した採点系と環境」を極限までスケールさせた成果。

オンプレミス運用の現実：立ちはだかるハードウェアの壁



- Proモデルのアクティブパラメータは420億だが、総量は1.02兆。
- 4bit量子化の単純計算でも、重みの保持だけで約510GBのメモリが必要。

通常のPC/単体GPU	✕ 実行不可
ローカルでの手元検証	9B蒸留版 を推奨
Proモデルのエンタープライズ利用	公式API または大規模クラスター運用が必須

ハルシネーションのパラドックス：賢さの代償

AA-Omniscience 前版 V2.5 → V2.6

正答率

約22.4% → 約34.9%

(知識量の増加)

ハルシネーション率

約24.7% → 約40.6%

(悪化)

「答えられる問題」は増えたが、「わからない時に保留する」能力が低下。未知の問題に対して不適切に回答（推測）してしまう傾向が強まっている。ファクトチェックの工数増に直結。

日本市場における適用ヒートマップ：検証済み領域と死角

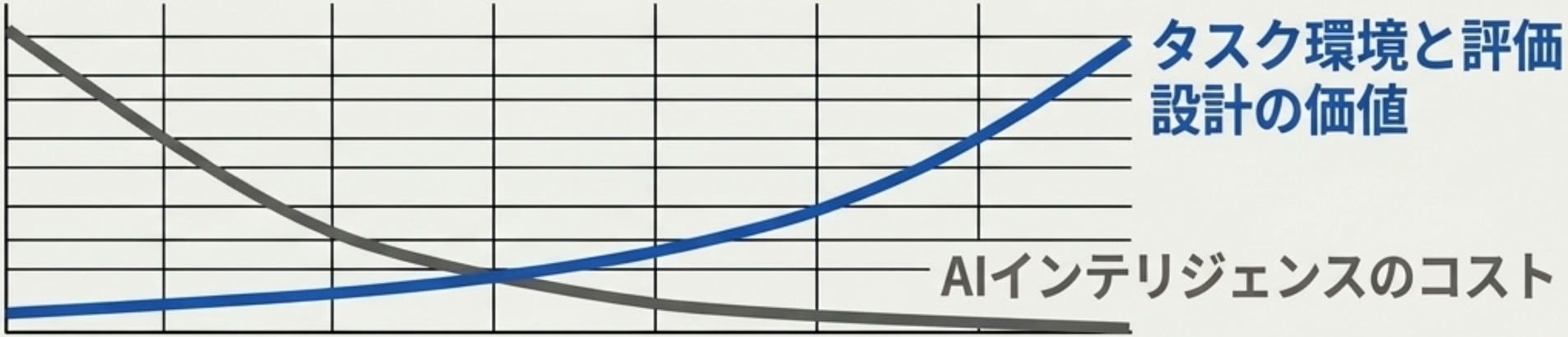
適用推奨 (Safe Zones)

- ✓ コーディング・テストコード生成
- ✓ 英語文献・技術仕様書の調査
- ✓ キャッシュを活かした大量データのRAG処理

要警戒・未検証 (Danger Zones)

- ✗ 日本の法務・契約書レビュー (特有の法的ニュアンス)
- ✗ 特許明細書の作成 (幻覚率40.6%のリスク直撃)
- ✗ 長文編集における不自然な漢字の混入 (初期報告あり)
- ✗ APIの混雑時・障害時の安定稼働

戦略的インサイト：エージェント業務のコモディティ化



圧倒的な低コスト（キャッシュ利用時）と高いツール実行能力により、エンタープライズのボトルネックは変化した。

過去の課題：
「このAIエージェントを動かすコストをどう正当化するか？」

現在の課題：
「AIが自律的に試行錯誤するための『高品質なタスク環境と評価基準』を人間がいくつ設計できるか？」

エンタープライズ導入への3ステップ・プレイブック

1

低リスク領域でのAPI検証（Start Small）

自前ホスティング（510GBの壁）は避け、コード生成やテスト反復など、成果を自動照合できる領域からAPI利用を開始する。

2

日本語実務のバッチテスト（Verify the Nuance）

実案件に近い課題を10〜20件用意し、正答率だけでなく「ハルシネーションの頻度」と「不自然な日本語の混入」を定量評価する。

3

人間参加型（HITL）での真のROI算出（Calculate True ROI）

APIトークン単価の安さだけでなく、「モデルが失敗した際の再試行コスト」と「人間のファクトチェック・修正時間」を合算し、実運用ベースのROIを策定する。