

2026年9月 新世代AIモデル横断比較

Claude Fable 5.1 / GPT-6 Astra / Gemini 3.8 Flash / Muse Spark 1.3

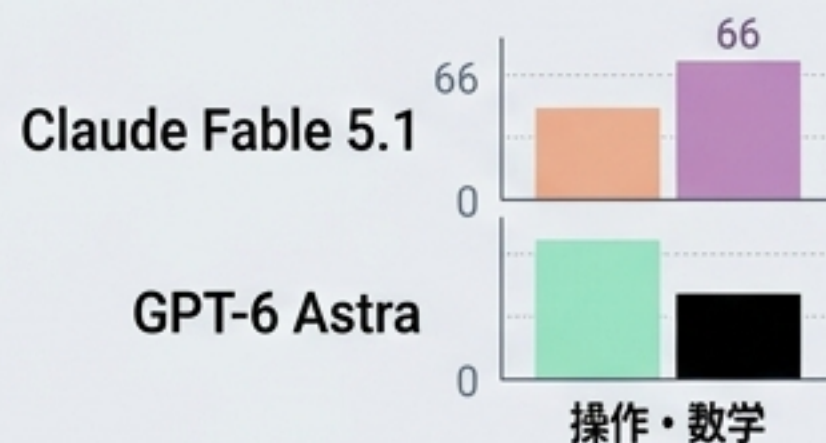
一次ソース検証に基づく独立評価と、日本企業実務への戦略的含意

Executive Summary: 3つの戦略的結論



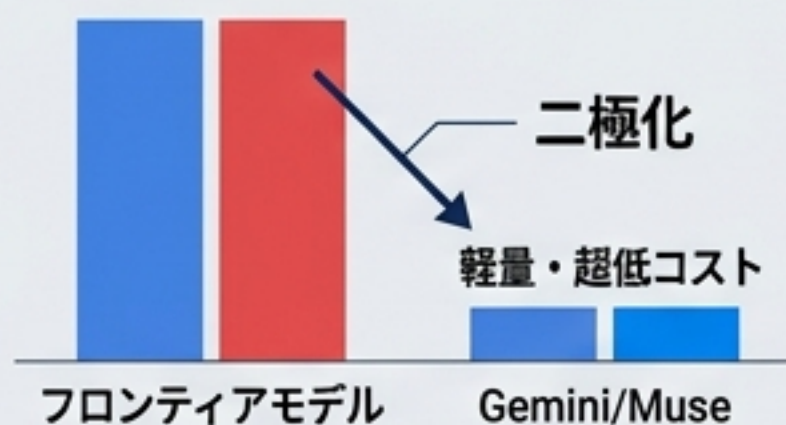
独立評価の勝者

総合知能ではClaude Fable 5.1が首位を維持（AA Index 66点）。GPT-6 Astraは特定領域（操作・数学）で突出するも総合2位。



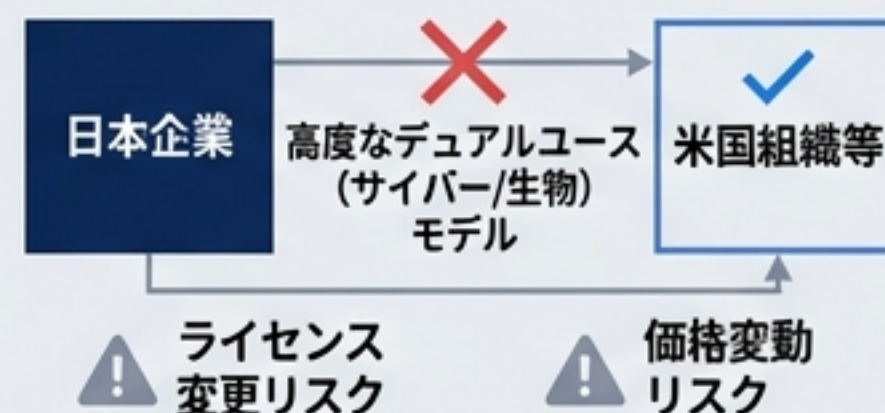
ティアの分化と価格破壊

フロンティアモデルと、Gemini/Museによる『軽量・超低コスト』ティアの明確な二極化。用途に応じた使い分けが必須に。



日本企業への警告

高度なデュアルユース（サイバー/生物）モデルは米国組織等に限定。ベンダーのライセンス変更や価格変動リスクへの備えが急務。



前提の検証：タイムラインと「神話」の解体

Anthropic

Myth (主張/噂)

Claudeのメジャーアップデート

✓ **Fact (事実)**

「5.1」は最上位系列「Fable」の版番号に過ぎない。

Meta

Myth (主張/噂)

Muse Spark 1.3はFable 5に匹敵

✓ **Fact (事実)**

コーディング特化（自社評価）の話であり、総合的な独立評価では最新のFable 5.1に及ばない。

OpenAI

Myth (主張/噂)

全く新しい謎のモデル

✓ **Fact (事実)**

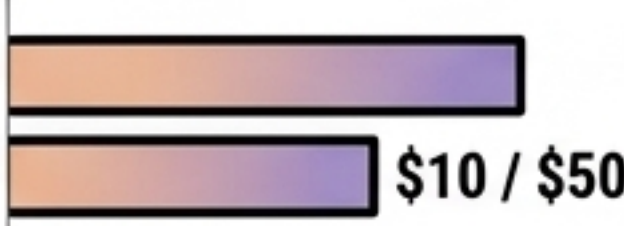
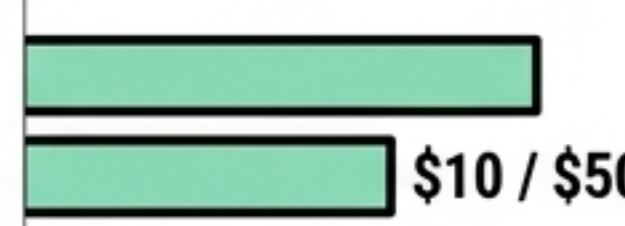
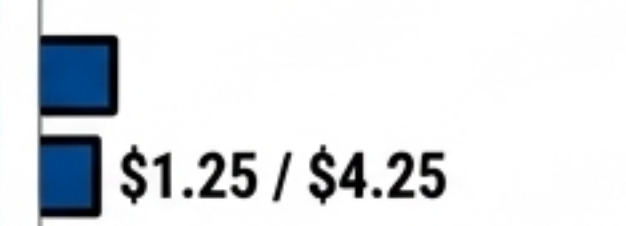
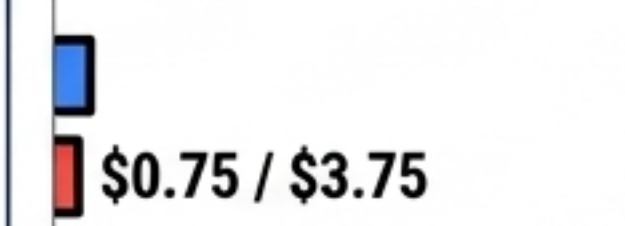
8月に未解決数学問題を解いた内部モデルの正式リリース版である。

2026年9月1日

9月2日

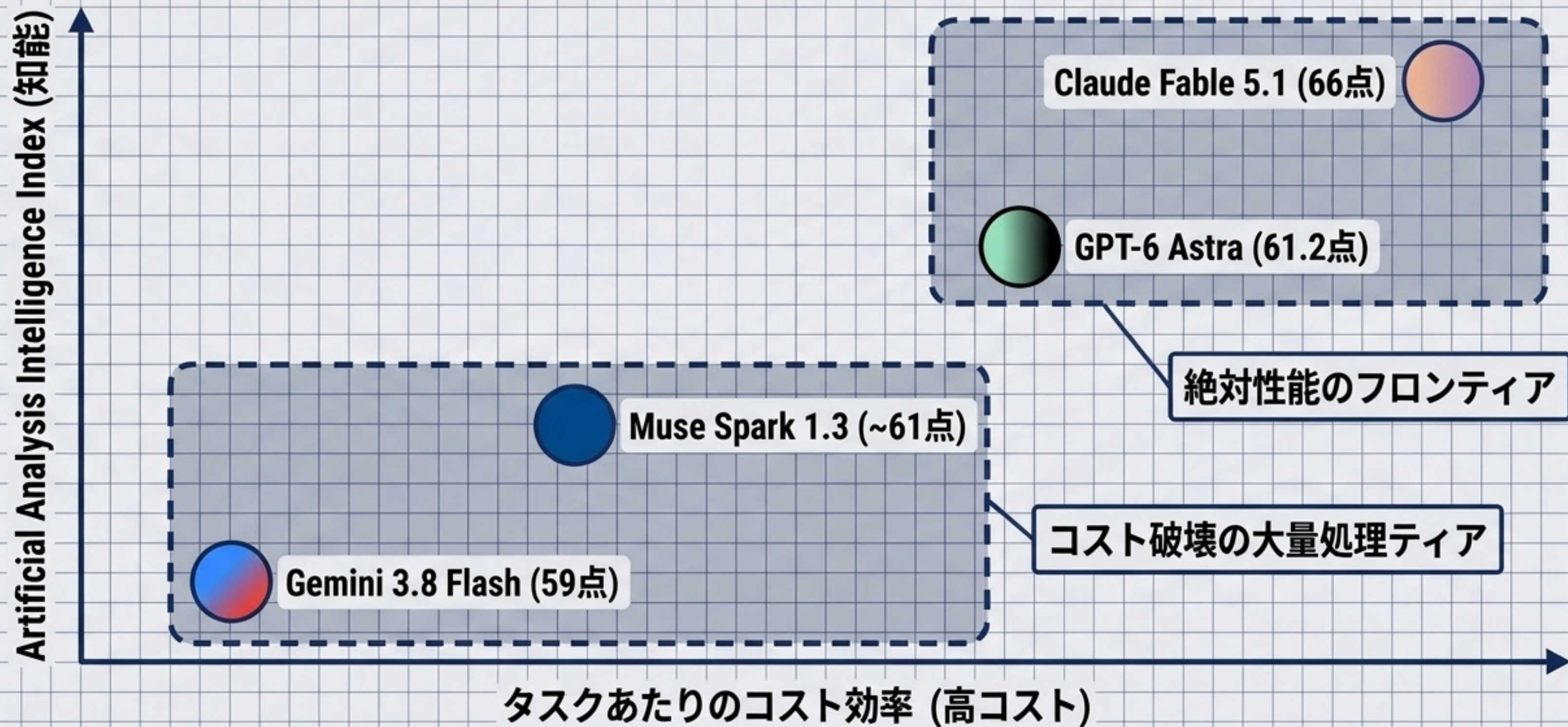
9月3日

基礎プロフィールとコスト構造比較

	Claude Fable 5.1	GPT-6 Astra	Muse Spark 1.3	Gemini 3.8 Flash
位置づけ	フラッグシップ	フラッグシップ	フラッグシップ級	軽量系列（Flash）
コンテキスト長	1M トークン	1M トークン	1M トークン	1M トークン
重み公開	クローズド	クローズド	クローズド（近日オープン未履行）	クローズド
入出力コスト/ 1M token	 \$10 / \$50	 \$10 / \$50	 \$1.25 / \$4.25	 \$0.75 / \$3.75

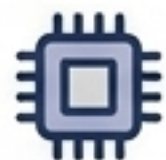
Gemini 3.8 Flashの入出力コストは、フロンティアモデルの約1/13。
圧倒的な低コストを視覚的に証明。

コストパフォーマンス・ランドスケープ：分化するティア



ベンチマークの不都合な真実：自己申告 vs 第三者評価

ベンダーの主張（自己申告）



ARC-AGI-3: 99.9%

FrontierMath Tier 4: 98% 飽和



自社有利なハーネス条件による測定。
OpenAI資金提供による利益相反が存在。

独立機関の真実（Epoch / AA）



ARC Prize 標準ハーネス: 62.7%

高難度のErdősセット解決率: わずか3%



AA Index v4.2 総合首位: Claude Fable 5.1 🏆



独立評価による客観的データ。

ボトムライン: 測定条件の非対称性に注意。「飽和」という言葉は特定の条件下でのみ成立するマジックワードである。

能力プロフィール1：知能のフロンティア

Claude Fable 5.1

サイエンス・エージェント

深い推論



タンパク質設計



長時間自律作業



実績 (Track Record)

Millennium社での無人実行安定性、Adaptyv Bioでの結合親和性10倍達成。企業向け安全性における明確な優位。

VS

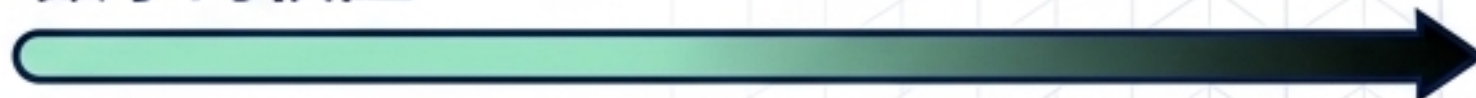
GPT-6 Astra

コンピュータ操作と数学

OS操作



数学的検証



サイバー防御



実績 (Track Record)

OSWorld 2.0で72.6%（圧倒的高速化）、数学の新結果発見、未知のゼロデイ発見。AGI到達への布石とされる強力な突破力。

能力プロファイル2：効率のフロンティア

Gemini 3.8 Flash

コスト破壊の大量処理

圧倒的低コスト

高速大量処理

Google

実績 (Track Record)

DeepSWE v1.1で大型フロンティアを低コストで凌駕。Wiz社においてコストを最大1/5に抑えつつ検出率を向上。

VS

Muse Spark 1.3

エージェント効率最適化

コーディング特化

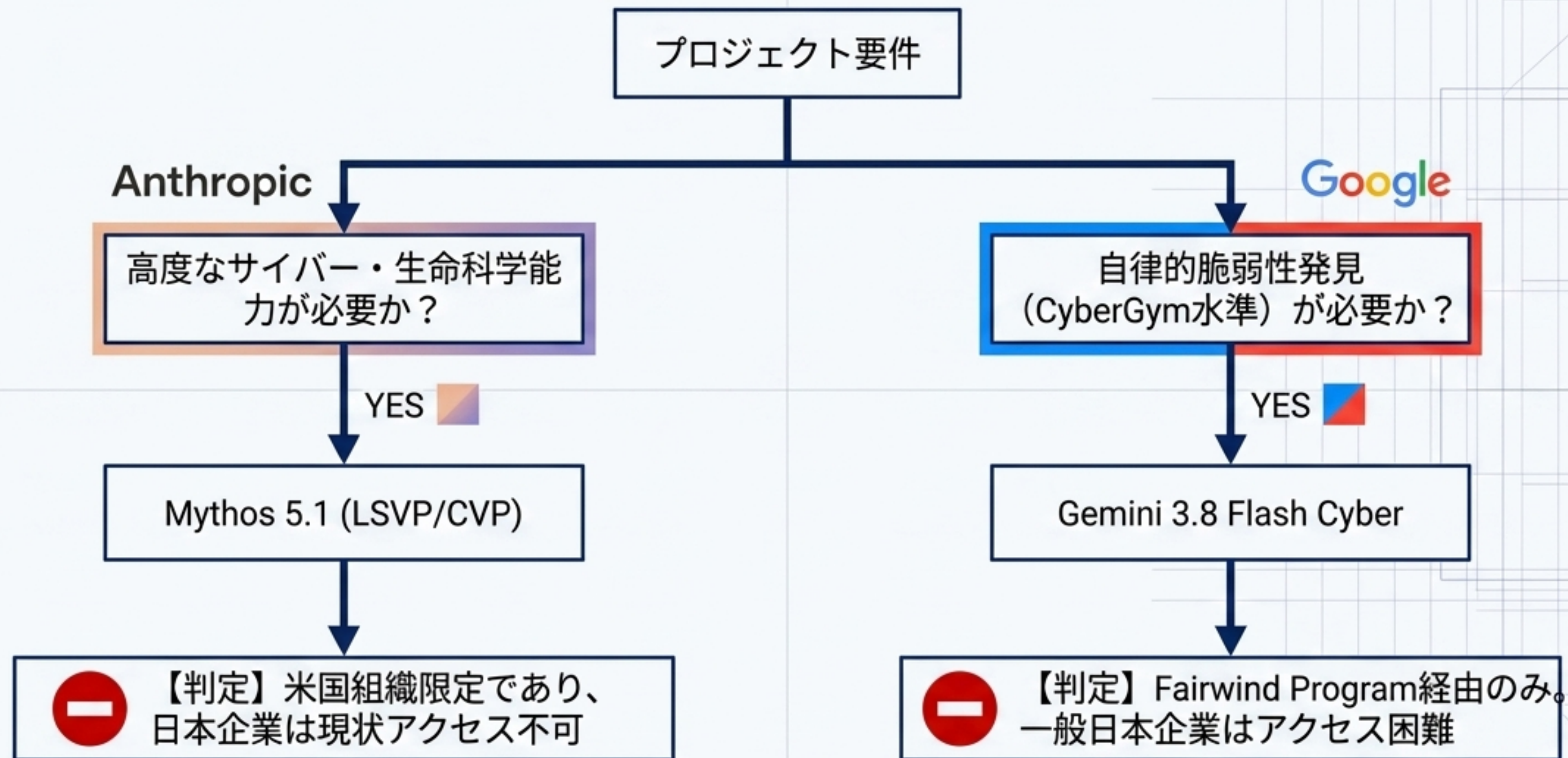
ツール呼び出し削減

Meta

実績 (Track Record)

前版比でツール呼び出し約20%減、トークン消費約25%減。Contributor tierでの極限のコストパフォーマンス。

日本企業における最大の障壁：デュアルユース・アクセス規制



インサイト:

高度なサイバー・生物用途は、当面フル能力版へのアクセスが日本からは制限される前提でR&D計画を組む必要がある。

追従する中国オープンウェイトの影

1

DeepSeek V4-Pro



コスト1/36でSWE-bench上位。しかしエージェント能力（LiveBench）は最下位クラスでscaffolding依存。

2

GLM-5.3



コストパフォーマンスにおいて米3社とのパレートフロンティアを共有。

3

Qwen3.8-Max



オールラウンダー志向（AA Index 58点）。

サマリーメッセージ: 総合知能では米フロンティア（Fable/Astra）がリードするが、コスト効率帯（Gemini/Muse）では中国勢が直接競合として肉薄している。

戦略的プレイブック：R&D・知財部門向けユースケースマッピング

業務プロセス (Tasks)

AI Models

	Claude Fable 5.1	GPT-6 Astra	Gemini 3.8 Flash	Muse Spark 1.3
1. 先行技術調査・深い推論・仮説生成	✓ 最適			
2. 社内システム操作自動化・数学的検証		✓ 最適		
3. 大量の公開特許のスクリーニング・要約			✓ 最適, 圧倒的コスト優位	
4. 社内コーディング支援のコスト最適化				✓ 最適, 効率重視

調達リスクと戦略再評価のトリガー

調達リスク (Procurement Risks)

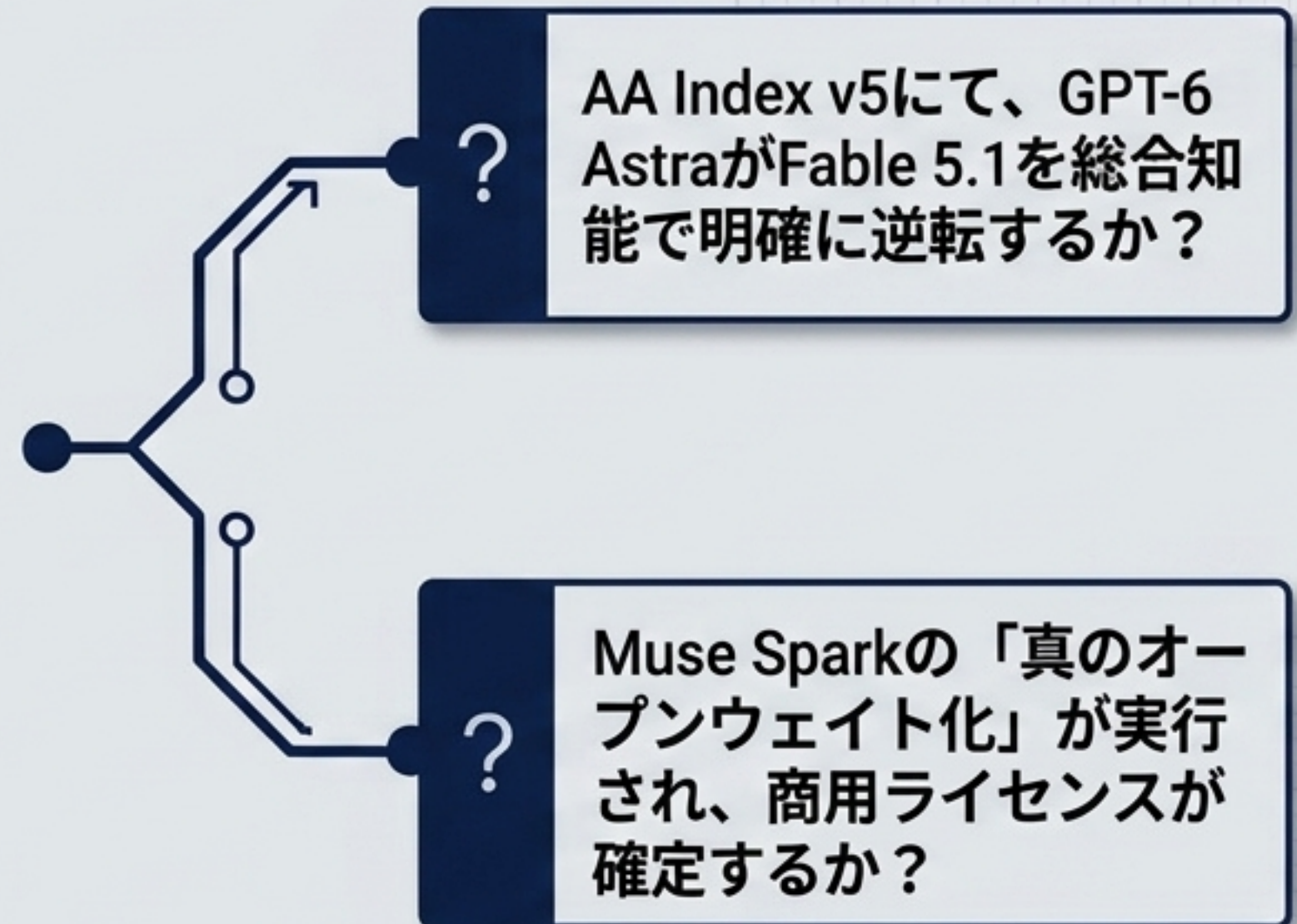


Metaのライセンス不安定性: 6ヶ月で2回変更、オープンウェイト化の未履行の約束に注意。



Gemini 3.8 Flashの価格改定: 2027年1月に価格が倍増する予定であり、コスト前提の見直しが必須。

見直しの閾値 (Re-evaluation Triggers)



総括：適材適所の選択へ

[モデル名]	[強み]	[弱み・制約]	[最も適する用途]
Fable 5.1	総合知能首位・科学生エージェント	高価格・生物/サイバー制約	R&D支援、知財の高度分析
GPT-6 Astra	OS操作・数学・サイバー	高価格・ハーネス依存の誇大表現	業務自動化、数学的検証
Gemini 3.8 Flash	圧倒的コスト効率・高速	最高難度の推論は不可・価格倍増予定	大量文書処理、コスト重視業務
Muse Spark 1.3	エージェント効率	オープン化未履行・予測可能性リスク	コスト最適化されたコーディング