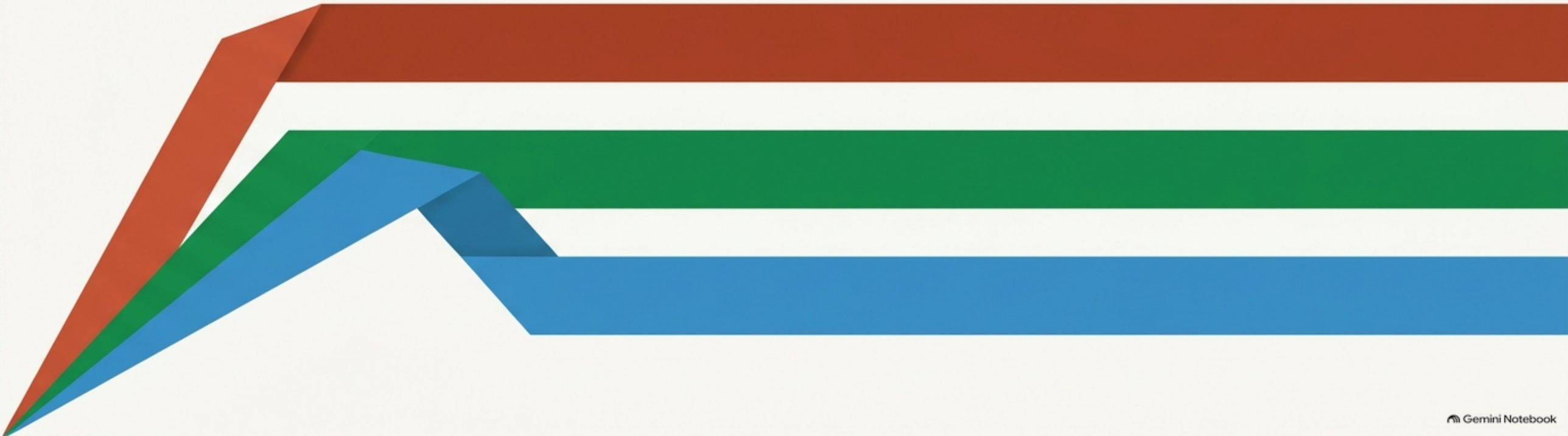
最先端AIの「適材適所」の真実

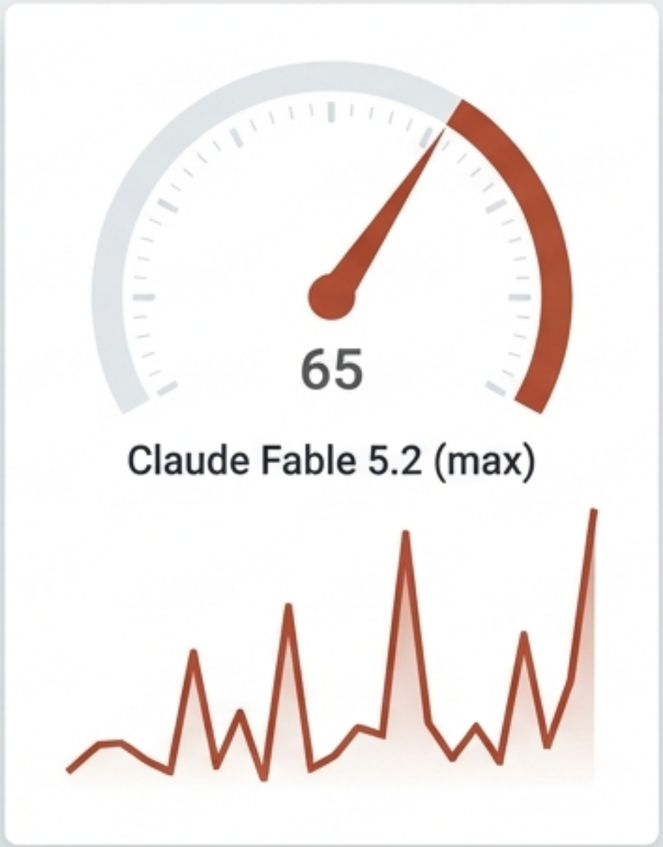
僅差のベンチマークに隠された、3つのフロンティアモデルの
明確な「性格」と戦略的選定プロセス

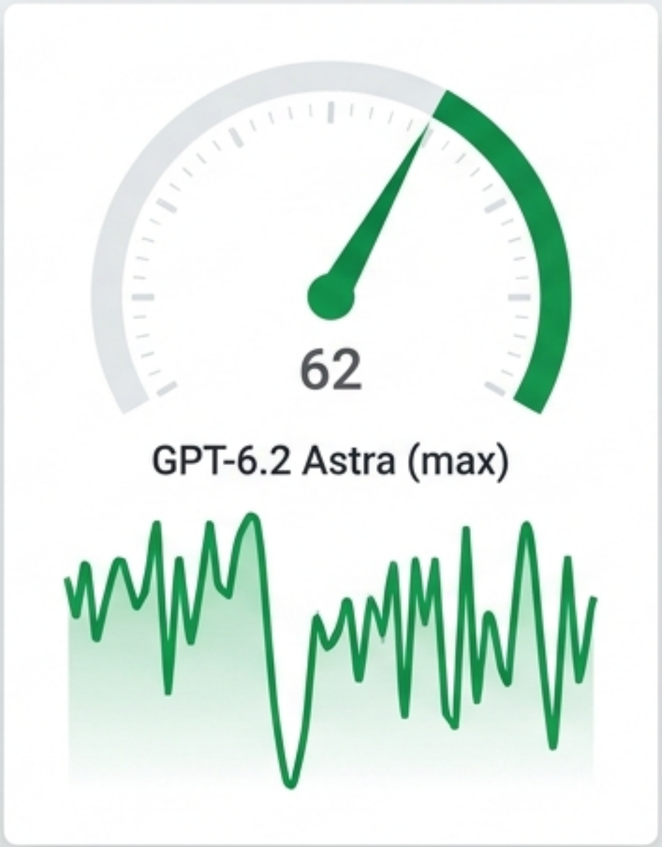


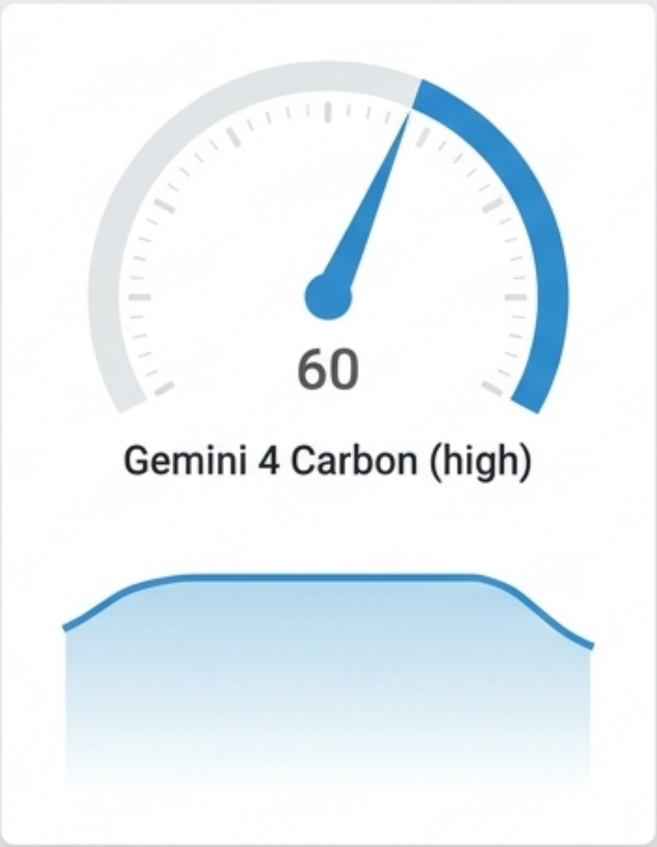
Gemini Notebook

総合スコアの僅差は「錯覚」に過ぎない

AA Intelligence Index v5.0.1 において、3つの最先端モデルは５点差以内の総合スコアを記録しました。しかし、単一の「最強モデル」を探すアプローチはすでに時代遅れです。各モデルの内部パラメータを分解すると、これらは全く異なる用途に向けて設計された、固有の「性格」を持つ別物であることがわかります。



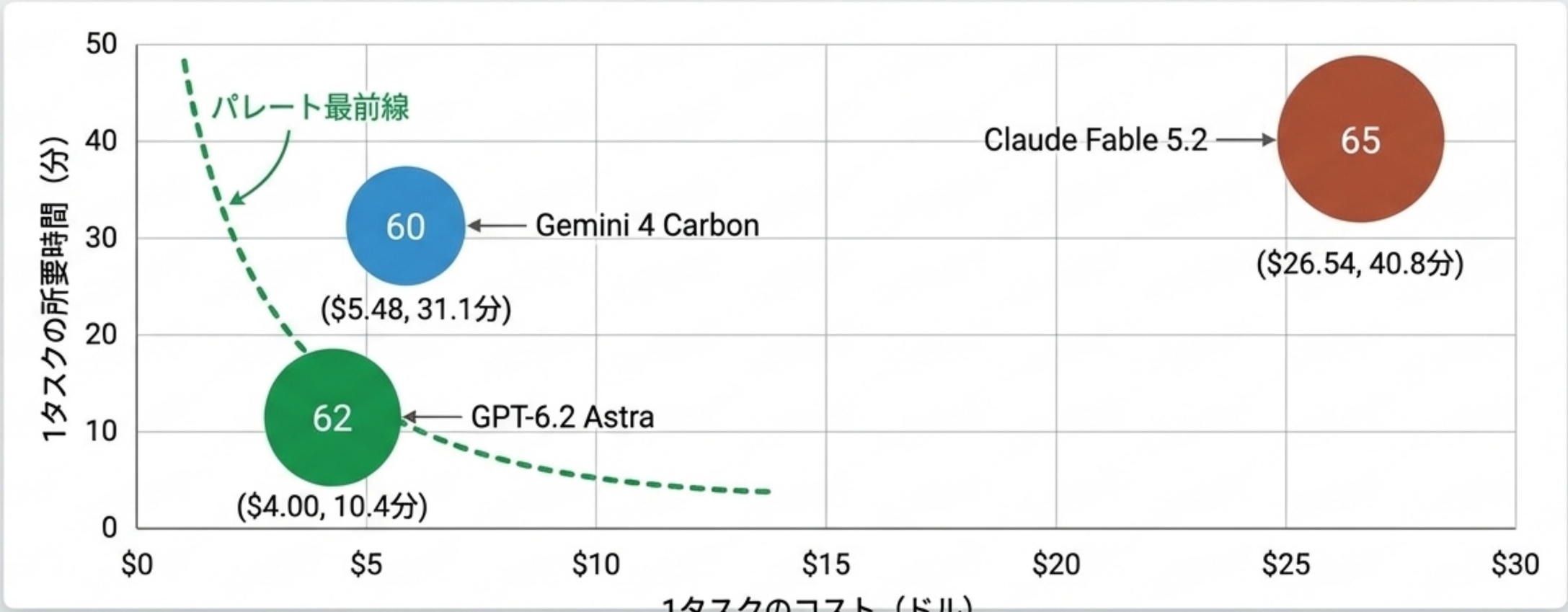




Gemini Notebook

「知能」に対するコストと時間のトレードオフ

各モデルに1つのタスクを実行させるためのコストと時間は、最大で約6倍の開きがあります。
GPT-6.2 AstraはFable 5.2の約15%の費用、4分の1の時間で総合3点差に肉薄しており、
圧倒的なコスト効率のパレート最前線を形成しています。



Gemini Notebook

組織に迎え入れるべき「3つの専門家」



Claude Fable 5.2

「知識と深い推論型」

知識の正確さと検討の深さが成果を左右する調査・分析のスペシャリスト。コストと時間の投資を惜しまない最重要課題向け。



GPT-6.2 Astra

「科学実行とコスト効率型」

科学技術系のエージェント作業や大量データを高速処理する科学者。驚異的なコスパを誇るるが、未知の領域では推測で語るリスクも。

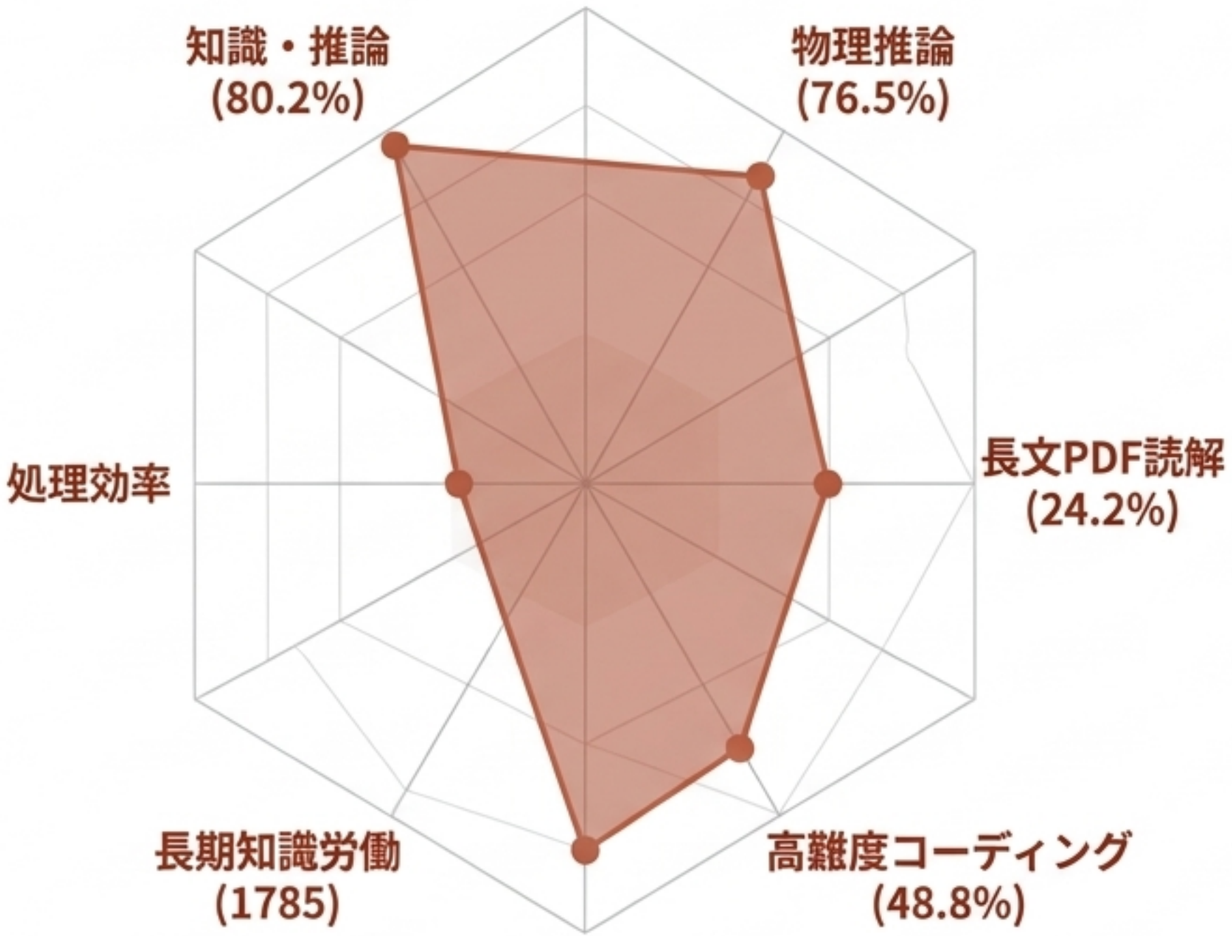


Gemini 4 Carbon

「堅実・自動化型」

知らないことは答えない守りの姿勢を持つ実務家。誤答を絶対に避けたい定型業務や端末自動化に特化。

Gemini Notebook



Claude Fable 5.2：知の探求者

【強み】

9評価中5評価で最高値を更新した圧倒的推論力。特にHLE-Diamondの知識問題では他を大きく引き離す。

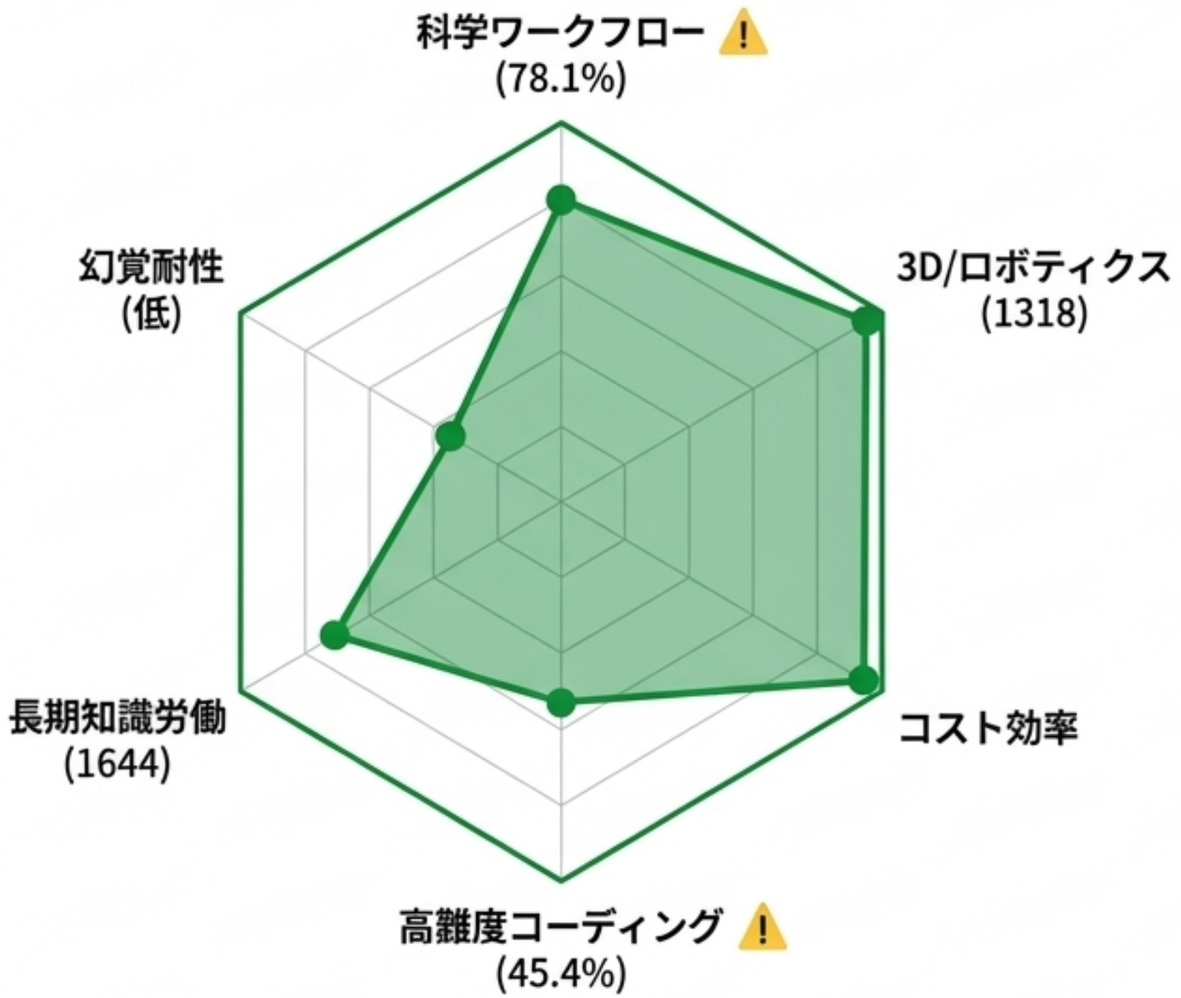
【弱み】

1タスクあたり約40分・$26超という莫大なリソース消費。出力トークンも2億7500万に達し、長文読解（GDP.pdf: 24.2%）や端末操作は今一つ。

※設定：Default Fallback (max)

Gemini Notebook

Premium Editorial Intelligence Dashboard



GPT-6.2 Astra：高速な科学者

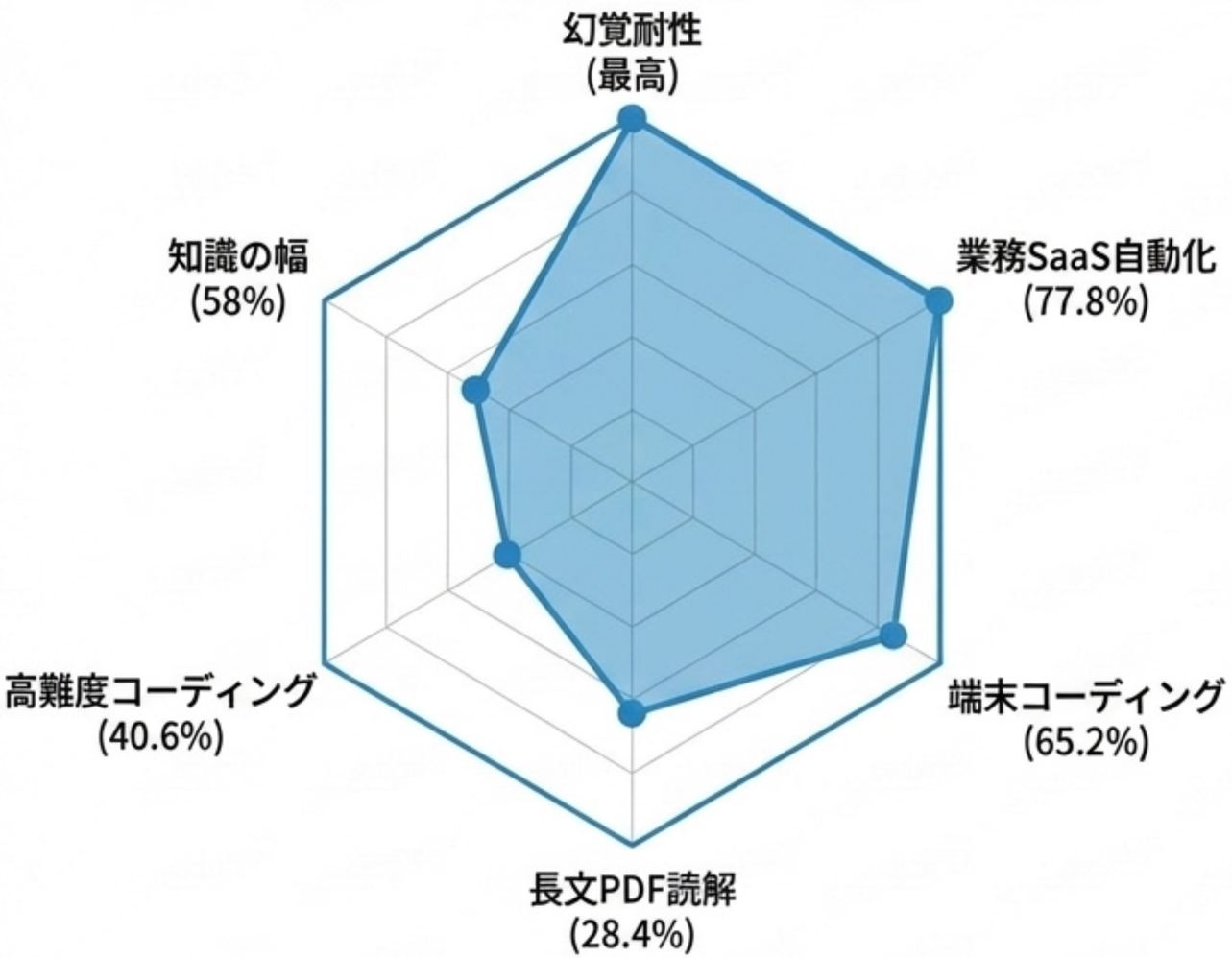
【強み】
科学研究ワークフローや3D/ロボティクス分野で首位。これらを圧倒的な低コスト・低遅延で実現する。

【弱み】
幻覚率が前世代から9ポイント悪化し59%に達する。長期の知識労働や高難度コーディングも上位モデルに劣る。

⚠ ※データ注記：Science, FrontierCode等は二次情報経由でありAAA公式との照合未了を含む。

Gemini Notebook

Premium Editorial Intelligence Dashboard



Gemini 4 Carbon：堅実な実務家

【強み】
知らないことは答えない「守りの姿勢」（幻覚率11%）。業務自動化や端末コーディングで首位を獲得し、定型業務に極めて強い。

【弱み】
知識の幅が狭く、高難度コーディングや物理推論は最下位。またGPT-6.2 Astraと比較して約3倍の処理時間を要する。

※設定：high設定での評価結果

Gemini Notebook

パフォーマンス・ヒートマップ（詳細評価データ）

Artificial Analysis個別評価の完全な比較。色付けされたセルは3モデル中の最高値を示します。

評価軸	Claude Fable 5.2	GPT-6.2 Astra	Gemini 4 Carbon
HLE-Diamond（知識＋推論）	80.2%	2位（未確認） !	69.1%
AA-Omniscience（知識の信頼性）	正答率 79%	幻覚率 59% !	幻覚率 11%
CritPt v2 Hard（物理推論）	76.5%	71.5%	69.0%
Terminal-Bench-Science（科学研究）	73.3%	78.1% !	64.8%
FrontierCode-AA（高難度コーディング）	48.8%	45.4% !	40.6%
Terminal-Bench 4.0（端末コーディング）	59.6%	64.1%	65.2%
AutomationBench-AA（業務SaaS）	76.4%	77.2%	77.8%
GDP.pdf（長文PDF読解）	24.2%	未確認	28.4%
1タスクのコスト / 時間	$26.54 / 40.8分	$4.00 / 10.4分	$5.48 / 31.1分

※未確認・二次情報に基づくデータには警告アイコンを表示

Gemini Notebook

知識と幻覚のパラドックス：リスク許容度による使い分け

「幻覚率59%」という数値は、モデルの性能が低いことを意味しません。FableやAstraは**「知らないことにも推測で答える」アグレッシブな設定**になっているため、正答率が伸びる一方で幻覚率も高くなります。対照的にCarbonは**「知らないことは答えない」防衛的な設定**であり、極めて難しい事実問題においては、この**リスク許容度の違い**こそが**モデルの性格**を決定づけています。



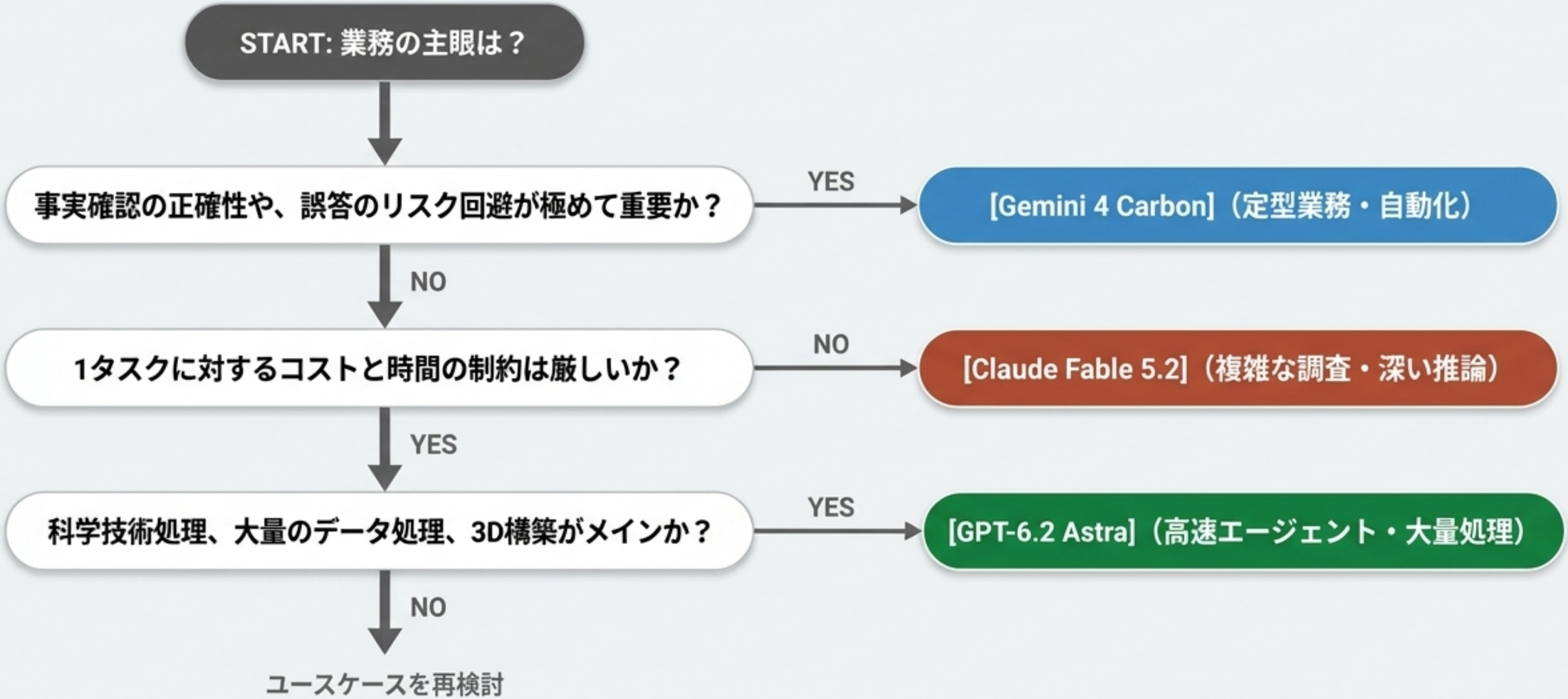
攻めの姿勢（Aggressive）
正答率：高 / 幻覚率：高 (59%)

守りの姿勢（Defensive）
正答率：低 / 幻覚率：低 (11%)

Gemini Notebook

モデル選定のデシジョンツリー

自社のタスク特性とリソース制約に基づいて、最適なAIモデルを導き出すための戦略的フロー。



Gemini Notebook

ユースケース・マッチングマトリクス

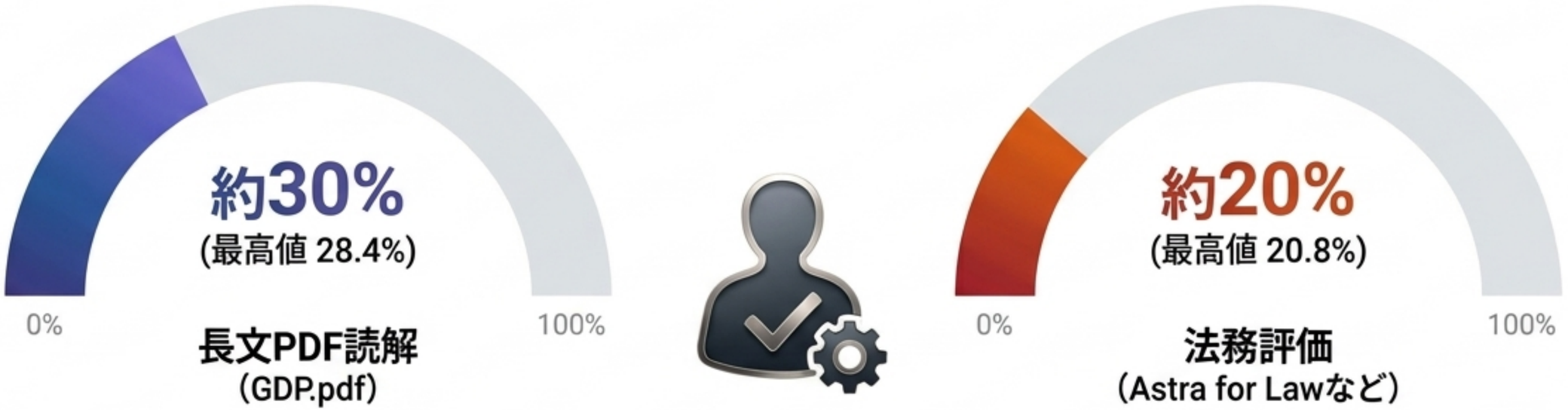
業務カテゴリ	Claude Fable 5.2	GPT-6.2 Astra	Gemini 4 Carbon
深い調査・分析・論理構築	●	◑	◔
定型業務SaaSの自動化	◑	◑	●
科学研究・ロボティクス推論	◑	●	◔
高難度の開発・コーディング	●	◔	◔
数万件の大量データバッチ処理	◔ コスト難	●	◑
厳密な事実ベースのQA対応	◔ 幻覚リスク	◔ 幻覚リスク	●

◎: 最適 (Highly Recommended) / ○: 適 (Suitable) / △: 注意が必要 (Proceed with Caution)

Gemini Notebook

フロンティアモデルの現在地と限界

適材適所でモデルを選定しても、完全自律型の運用にはまだ壁があります。特に長大なPDFの読解や、厳密性が求められる法務評価の合格率は現状2〜3割にとどまります。専門業務へのAI実装においては、これらの能力限界を前提とし、人による検証（Human-in-the-loop）を組み込んだワークフロー設計が不可欠です。



必須要件: Human-in-the-loop（人間の介在）

Gemini Notebook