フロンティアLLMの第三者評価

単一モデル依存からの脱却と、用途別「Fit-for-Purpose」展開戦略

評価対象: GPT-6.2 Astra, Claude Fable 5.2, Gemini 4 Neon, Gemini 4 Carbon, Claude Opus 5.5, GPT-6.2 Sol
データソース: Artificial Analysis, Epoch AI, Vals AI, Arena
日付: 2026年11月20日

Gemini Notebook

Premium intelligence briefing

総合スコアの時代は終わった



幻想の同点首位

主要ベンチマークでトップモデルは「174点」で並ぶが、統計的推定の幅と得意領域は完全に異なる。整数のランキングはモデル選定の根拠として機能しない。



領域ごとの明確な勝者

科学研究（Astra）、実務エージェント（Fable）、法務・金融（Neon）。「どの仕事を任せるか」によって第一候補は劇的に入れ替わる。



推論設定の経済的最適解

最上位モデルの「最高設定」はオーバースペック。一段下げた推論設定（例：Astraのxhigh）が、品質を維持しながらコストと時間を半減させる「スイートスポット」となる。

Gemini Notebook

Premium intelligence briefing

評価指標が暴く「局所的優位性」

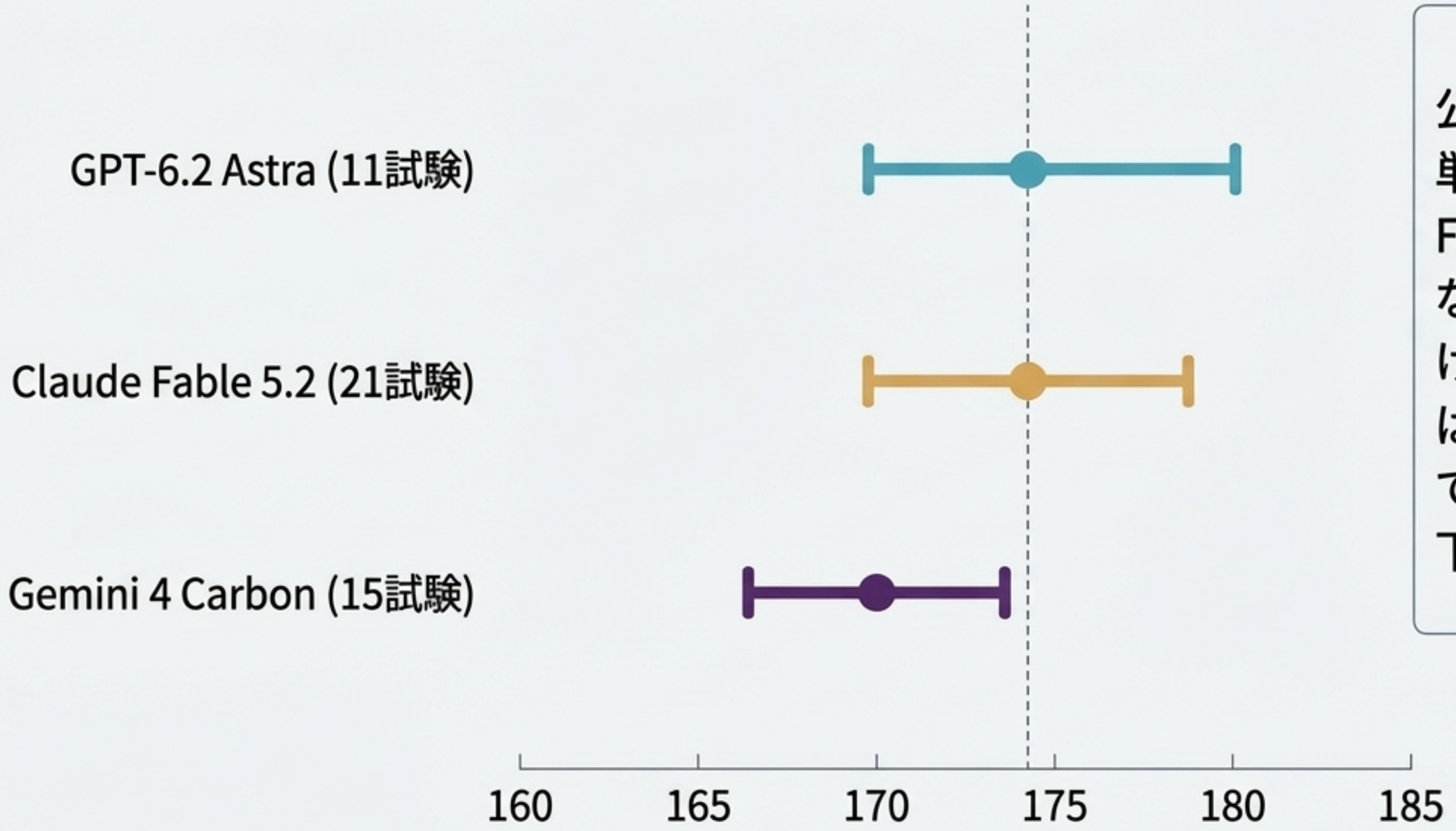
	AA 指数	Epoch ECI	Vals 総合	Vals 科学	Agent Arena	WebDev
GPT-6.2 Astra	62	174	6位	1位	6位	3位
Claude Fable 5.2	65	174	1位	3位	1位	1位
Gemini 4 Neon	61	—	2位	4位	3位	7位
Gemini 4 Carbon	60	170	4位	—	4位	6位
Claude Opus 5.5	59	169	5位	7位	2位	2位
GPT-6.2 Sol	57	—	10位	8位	11位	5位

視線を集める場所が列ごとに異なる事実に着目してください。
これが『用途特化型』選定が必須である視覚的証拠です。

Gemini Notebook

Premium intelligence briefing

「174点の同点」が意味する統計的現実



公表される整数（174）は単なる中央値。AstraとFableは完全に区間が重なっており、この数値だけで優劣を決めることは統計的に不可能。一方でCarbonは明確に一段下に位置している。

Gemini Notebook

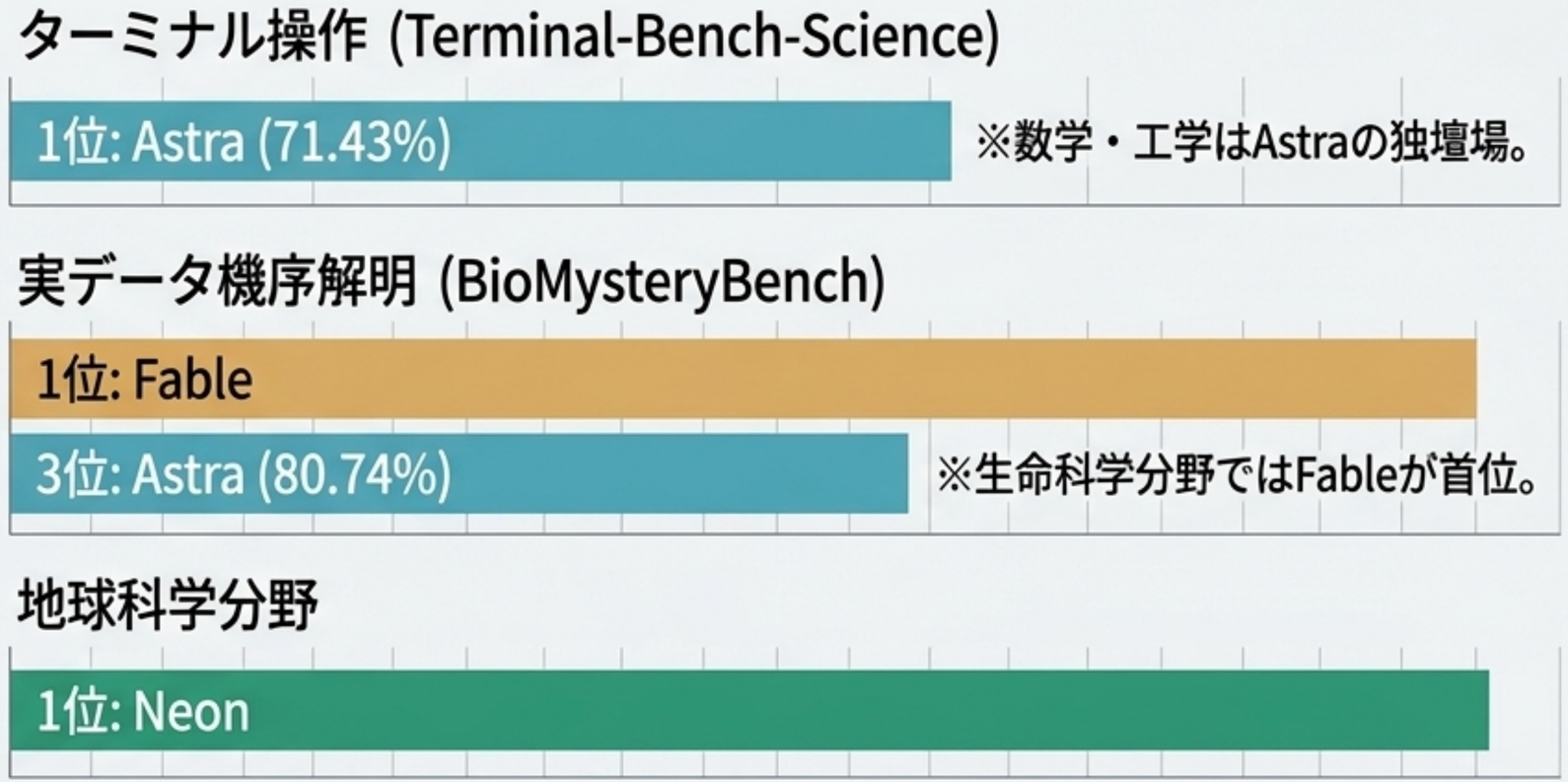
Premium intelligence briefing

科学分野における解像度：数学的処理 vs. 生命科学

68.44%

Vals Science Index
総合首位

GPT-6.2 Astra



科学という一語で括ることは危険。ターミナルと数学に強いAstra、生物学に強いFable、
地球科学に強いNeonと、分野ごとに明確な棲み分けが存在する。

Gemini Notebook

Premium intelligence briefing

研究設計から最終報告までの「完遂力」の格差

データ分析・ターミナル操作

強み: 局所的な研究作業、数学的処理

優位: GPT-6.2 Astra

最終レポートの執筆・プロジェクト完遂

データセンター調査タスク（Epoch人手採点）

1位: Claude Fable 5.2 (98%)

2位: Claude Opus 5.5 (88%)

3位: GPT-6.2 Astra (65%)

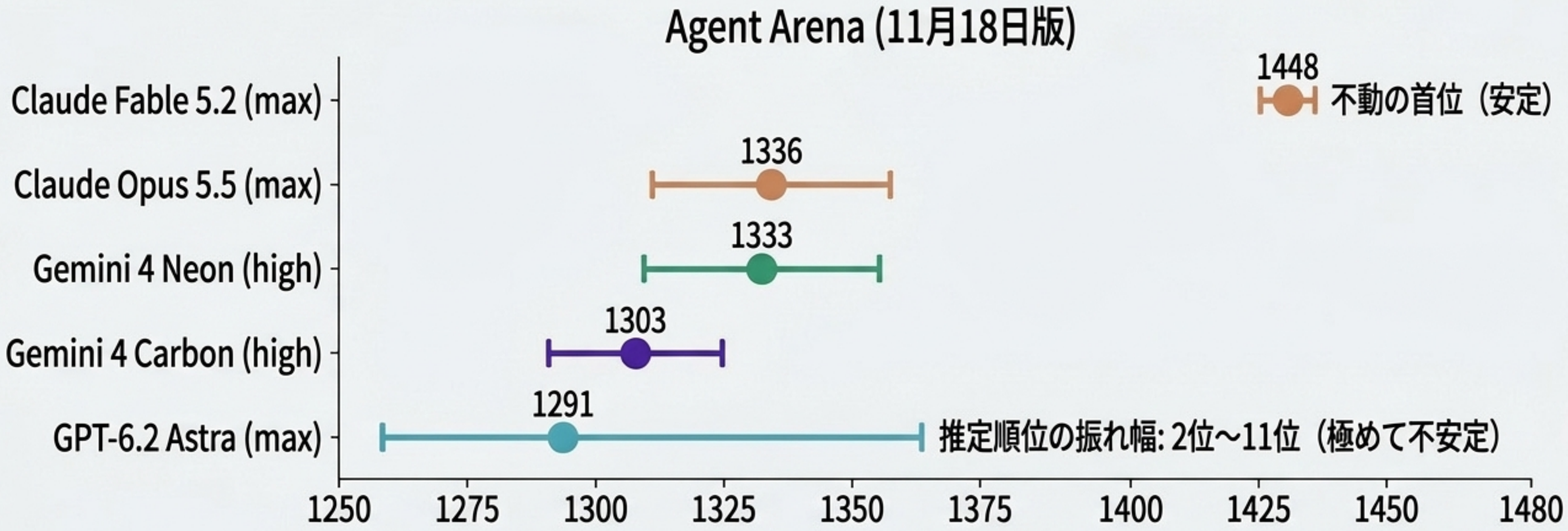
Astraはターミナルでの局所的な研究作業には優れるが、最終的な研究報告書をまとめる段階で論理破綻（冒頭の結論と後段の条件付き結果の食い違い）を起こす傾向がある。

部品としての知能（Astra）と、プロジェクトマネージャーとしての知能（Fable）の違い。
業務を完遂させるにはFableが圧倒的に優位。

Gemini Notebook

Premium intelligence briefing

自律型エージェントの実利用：最高到達点より「分散の少なさ」



4,785セッション、43万観測の実利用データ。Astraは推定順位の振れ幅が大きく、実務の自動化に組み込むには不確実性が高い。一方でコーディング限定（Agentic Coding）であれば3位と健闘する。

Gemini Notebook

Premium intelligence briefing

プロフェッショナル・サービス（法務・金融・税務）

ベンチマーク	1位	他の主要モデルの順位
Vals Legal Index (米国法務)	Gemini 4 Neon (78.36%)	4位 Carbon (75.93%) / 23位 Astra / 24位 Fable
Vals Finance Index	Gemini 4 Neon (68.31%)	2位 Carbon (67.05%) / 23位 Astra
Tax Agent Bench	Gemini 4 Neon (54.89%)	56位 Astra (21.84%)

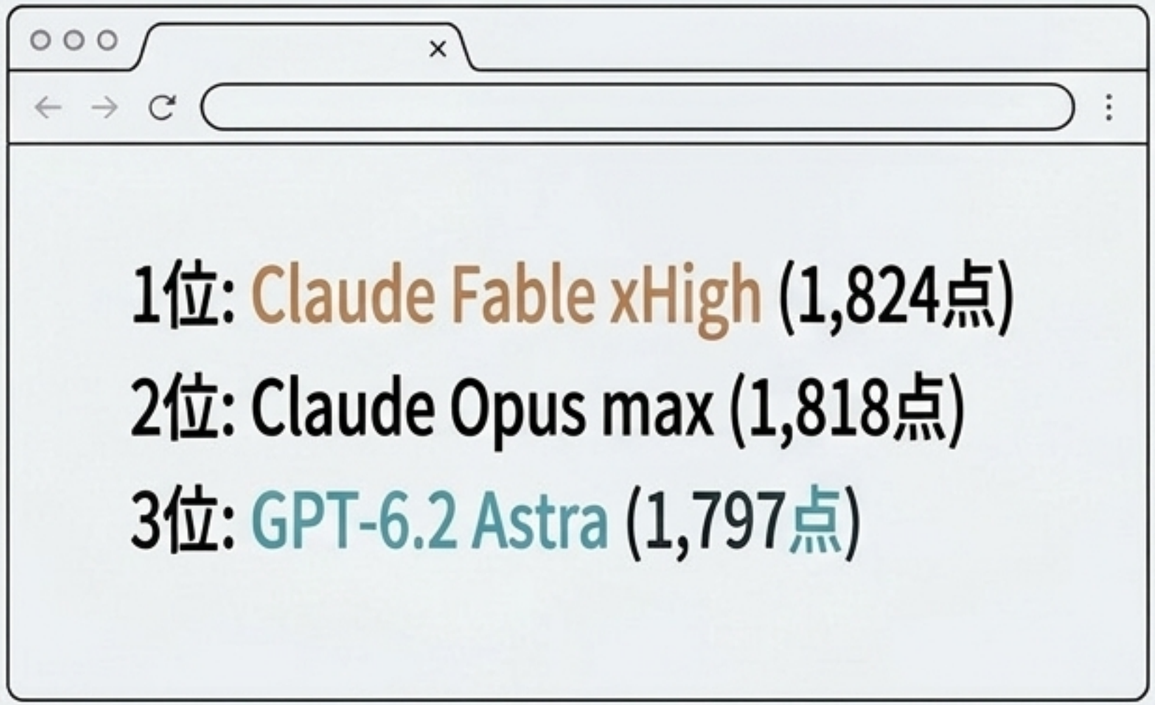
※Harvey法務課題の採点指標に関する注意点
AA版（誤情報を含む全要件合格率）ではAstra 15.0% / Fable 7.2%。一方、Vals版ではAstra 16.25%で全体の12位。単一の指標ではなく、実装と採点指標を併記して評価することが必須。

Gemini Notebook

Premium intelligence briefing

Web開発機能（WebDev）における評価の分岐

公式画面での評価

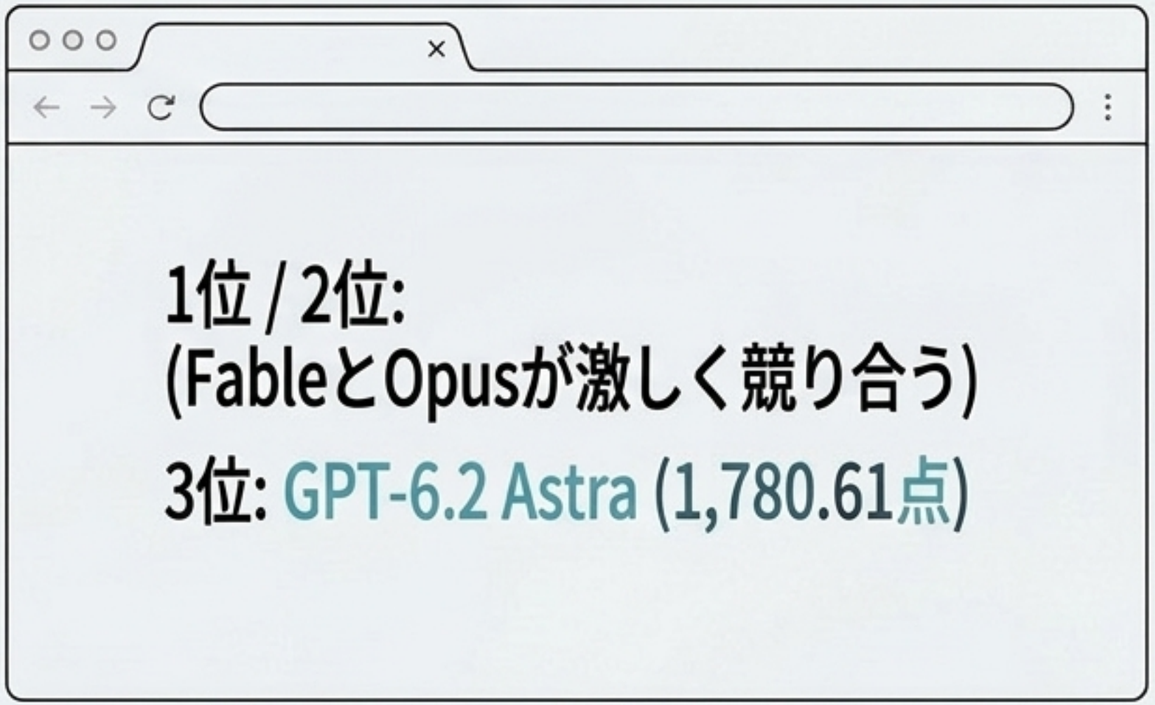


1位: Claude Fable xHigh (1,824点)

2位: Claude Opus max (1,818点)

3位: GPT-6.2 Astra (1,797点)

公開データ（1,102票）での評価



1位 / 2位:
(FableとOpusが激しく競り合う)

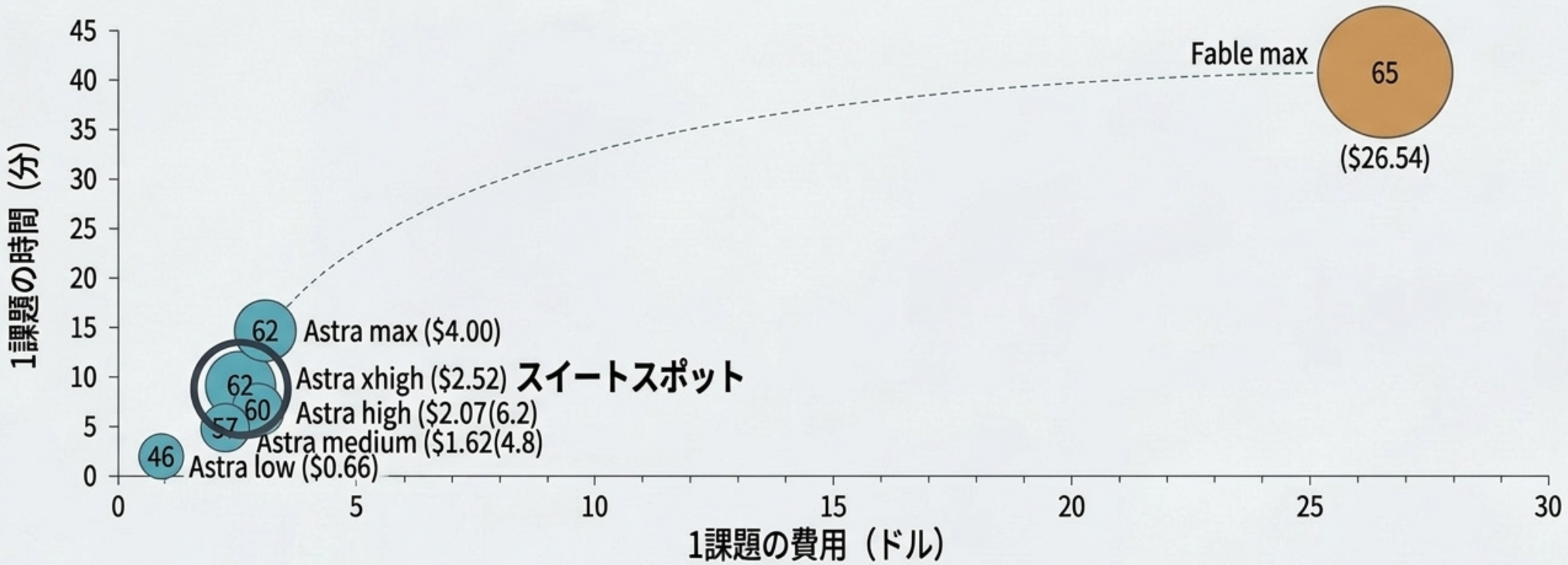
3位: GPT-6.2 Astra (1,780.61点)

指示に基づいてWebアプリを生成するタスクでは、FableとOpusが最高評価を争う。評価プラットフォーム（表示面や収録範囲）によって上位のスコアは変動するが、Astraは一貫して3位グループに留まっている。

Gemini Notebook

Premium intelligence briefing

推論設定の効率性フロンティア：Astraのスイートスポット



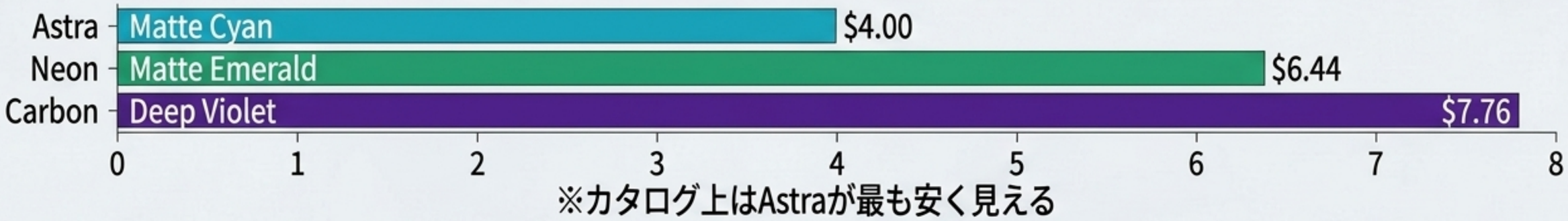
表示上のスコア(62)を維持したまま、maxからxhighに下げることで、費用を37%削減し、処理時間間をほぼ半減させることが可能。ただし、高難度課題ではmaxとxhighで差が出る点に注意。

Gemini Notebook

Premium intelligence briefing

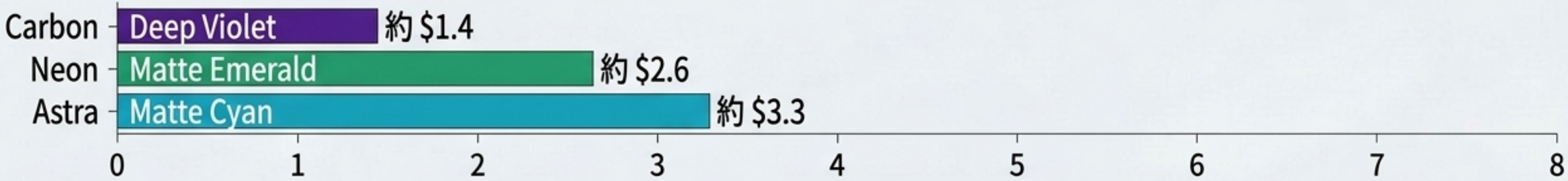
カタログ価格 vs. 現実の運用コスト

カタログ上の標準価格（AA標準価格ベース）



※カタログ上はAstraが最も安く見える

実利用コストの中央値（Arena利用記録ベース）



※実際の作業構成ではNeonとCarbonが逆転し圧倒的に安価になる

APIの標準価格表に基づく試算は、実運用のコストを正確に反映しない。
作業の構成と集計方法により、実際に発生するコストの順位は完全に逆転する。

Gemini Notebook

Premium intelligence briefing

統合診断マトリックス：用途別最適モデルの選定

	TASK	RECOMMENDED MODEL
	数学・工学 / ターミナル操作	第一候補 = GPT-6.2 Astra
	エンドツーエンドの研究報告 / 自律エージェント	第一候補 = Claude Fable 5.2
	米国法務・金融・税務	第一候補 = Gemini 4 Neon
	Webアプリケーション開発	第一候補 = Claude Fable 5.2 または Opus 5.5
	大規模・低コストの日常処理	第一候補 = Gemini 4 Carbon

各モデルのアーキテクチャ特性は、特定のドメインで不可逆的な優位性を生む。組織は単一のAIベンダーに依存するのではなく、このマトリックスに基づくマルチモデル・ルーティングを構築すべきである。

Gemini Notebook

Premium intelligence briefing

エンタープライズ展開の実践：知財・法務戦略への適用

1. 発明発掘・実施例の検討

Assigned Model:

GPT-6.2 Astra

KPI Box
検証: 観測事実と推定した作用機序の切り分け。裏付けのない効果の数え上げ。

2. 複数資料を使う競合調査

Assigned Model:

Claude Fable 5.2

KPI Box
検証: 公報・論文の照合。重要文献の取りこぼし率、根拠まで追える主張の割合。

3. 契約・法務資料の調査

Assigned Model:

Gemini 4 Neon

KPI Box
検証: 日本語での契約レビュー。引用箇所の正確さ、論点の漏れ。Vals(米国法)との差分検証。

4. 新着公報の定期監視

Assigned Model:

Sol / Gemini 4 Carbon

KPI Box
検証: 評価用セットの再現率達成にかかる費用と人間の確認時間の削減率。

Gemini Notebook

Premium intelligence briefing

戦略的デプロイメントの3つの鉄則

1 「整数のランキング」を意思決定から排除せよ

同点や微差は統計的ノイズに過ぎない。必ず「信頼区間」と「ドメイン別の内訳」で評価すること。

2 運用コストは「カタログ」ではなく「実地」で測れ

理論上のトークン単価ではなく、自社タスクを回した際の実利用中央値（Arenaの手法）で予算を組むこと。

3 モデル切り替えの前に「推論設定」を最適化せよ

最新モデルを盲目的に最高設定（max）で回すのは非効率の極み。品質を落とさずコスト・時間を半減できる設定を見極めること。

Gemini Notebook

Source Typology（原典および参照データ）

Artificial Analysis

- LLM Leaderboard
- Release Comparison
- Harvey LAB-AA
- 価格条件

Epoch AI

- Capabilities Index (ECI区間)
- Astra公式同順位
- Astra公式実務検証 (Fable vs Astra)

Vals AI

- Index v2.2 (産業重み付け)
- Science Index (3課題構成)
- Legal/Finance Index

Arena

- Leaderboard (Agent/Coding)
- Code Arena (WebDev)
- Indexシグナル変更
- 公式配布データ(11/18版)
- 公式配布データ(11/18版)

OpenAI

- Astra/Sol 開発者評価

すべての評価基準および数値は2026年11月20日（日本時間）時点の公開データに基づく。最新のライブ表は各評価機関のポータルにて随時更新される。

Gemini Notebook