フロンティアLLM 第三者評価ランドスケープ 2026

GPT-6.2 Astra公開後の勢力図：5大評価機関のデータから紐解く、次世代AIモデルの「真の実力」と「最適解」

調査基準日：2026年11月20日（日本時間）
対象機関：　Artificial Analysis, Epoch AI, Vals AI, Arena, ARC Prize

[Confirmed]　確認済み事実
[Secondary]　二次情報
[Analysis]　分析

Gemini Notebook



[◎ Analysis]

絶対的王者の君臨

Claude Fable 5.2が総合首位。Artificial Analysis、Vals、Arenaの各指数で1位。妥協なき**最高精度**を誇るが、コストは**最大**。



[◎ Analysis]

理数特化の破壊的イノベーター

GPT-6.2 Astraは**科学・数学・抽象推論**で他を圧倒。高い推論能力を持ちながら、1タスクコスト ($4.00) は極めて効率的。



[◎ Analysis]

実務特化の高ROIダークホース

Gemini 4系がValsやArenaで上位を席巻。圧倒的な**単価**の低さで、実務チャットやルーチン作業における**コストパフォーマンスの最適解**に。



[✓ Confirmed]

「最強」のパラダイムシフト

単一の最強モデルは存在しない。**評価ハーネス**、**推論努力**（Inference Effort）の設定、タスクの性質によって順位は完全に**逆転**する。

Gemini Notebook

[Analysis]

フロンティアLLM ポジショニング・マトリクス

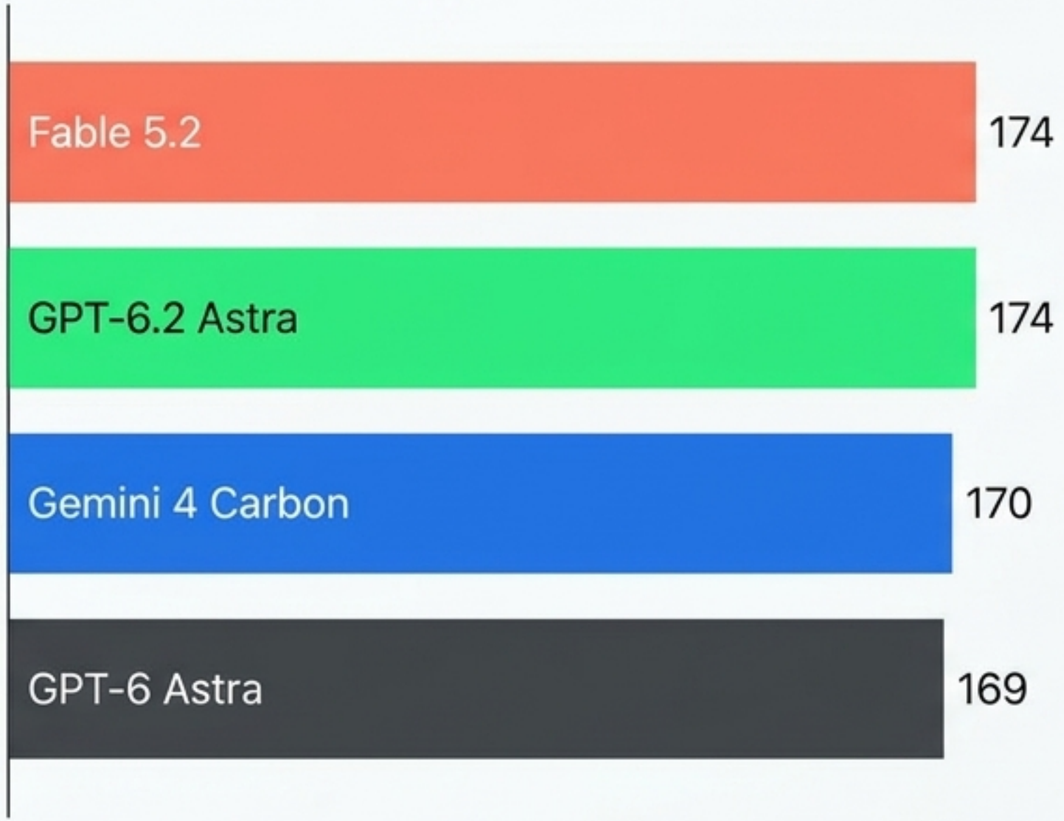
	Claude Fable 5.2	GPT-6.2 Astra	Gemini 4 Series
総合評価	最高水準	同点1位 / 特化型	上位食い込み
得意領域	• 実務・コーディング • 開放タスク	• 科学・数学・ • 未解決問題	• 実務・一般チャット
弱点・課題	• 最高設定時のコスト高	• 金融・法務・ • 事実知識・幻覚	• 最高難度推論
最適なユースケース	品質最優先の複雑な実務	低予算での高度な論理検証	高ROIが求められる大量処理

Gemini Notebook

[✓ Confirmed]

Epoch Capabilities Index (ECI)

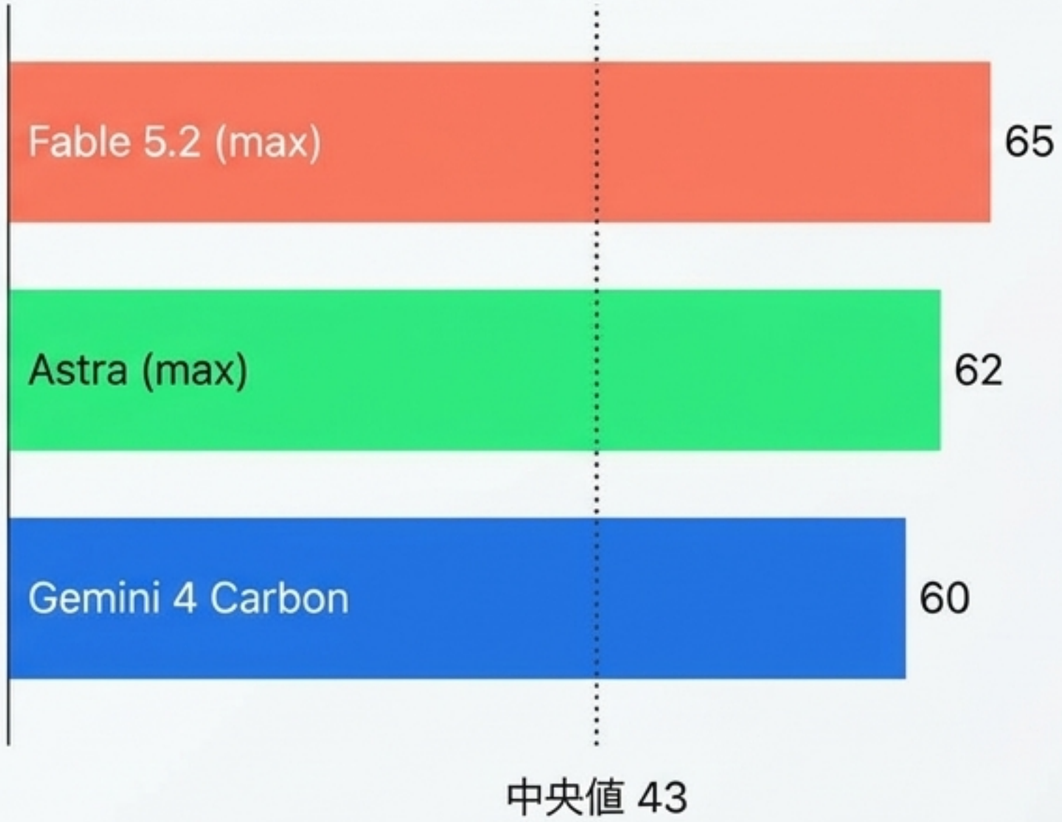
241モデル中の頂点。同点1位だが、内訳は全く異なる



[✓ Confirmed]

Artificial Analysis (Intelligence Index v5.0.1)

Astraはモデル単位で2位。前モデル(57)から確実な進化



Gemini Notebook

[✓ Confirmed]

理数・推論の絶対王者

Vals Science Index



FrontierMath Tier 4 v2



FrontierMath Erdős



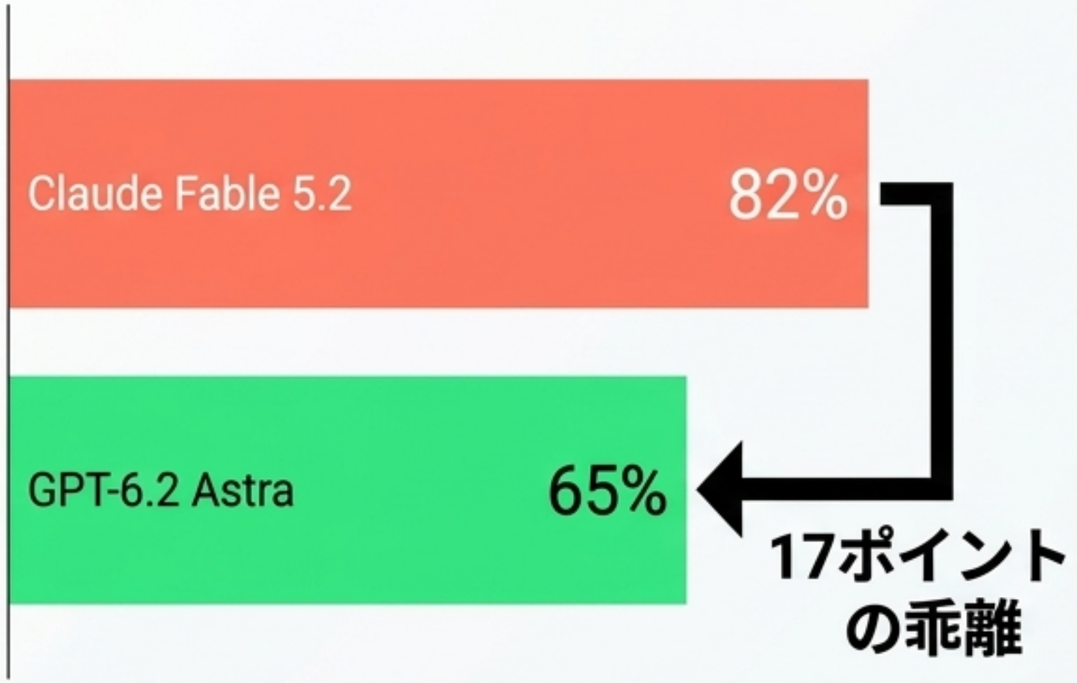
インサイト分析：
OpenAIの「科学者向け最強モデル」という主張はValsのデータ（Terminal-Bench-Science 1位等）と完全に一致する。未知の抽象推論領域において、Astraは現在絶対的な優位性を持つ。

Gemini Notebook

ビジネス実務・知識労働における明暗

[☒ Secondary]

Epoch 実務タスク評価 (Automation Reports)



データセンター分析課題ではFable 98%に対しAstra 65%

[✓ Confirmed]

Vals AI 専門実務領域

	Finance	Legal	Tax
Fable	Fable Finance: 1位	Fable Legal: 1位	Fable Tax: 2位
Gemini	Gemini Finance: 8位	Gemini Legal: 5位	Gemini Tax: 10位
Astra	Vals Finance: 23位 / 45中	Vals Legal: 23位 / 43中	Tax Agent: 56位 / 90中

金融・法務・税務においてAstraは中位〜下位に沈む

Gemini Notebook

[✓ Confirmed]

コーディング・エージェント領域

Arena 指数ランキングボード（11/18導入の新指数）

Coding Index

1位: Claude Fable 5.2 (1,489) ■ ↑

2位: Claude Opus 5.5 (1,373) ■ ↑

3位: GPT-6.2 Astra (1,338) ■ ↓

Work Index

1位: Claude Fable 5.2 (1,445) ■ ↑

2位: Claude Opus 5.5 (1,317) ■ ↑

6位: GPT-6.2 Astra (1,280) ■ ↓

Code Arena WebDev

1位: Claude Fable 5.2 (xHigh: 1,824) ↑

3位: GPT-6.2 Astra (Max: 1,797) ↓

[☒ Secondary]

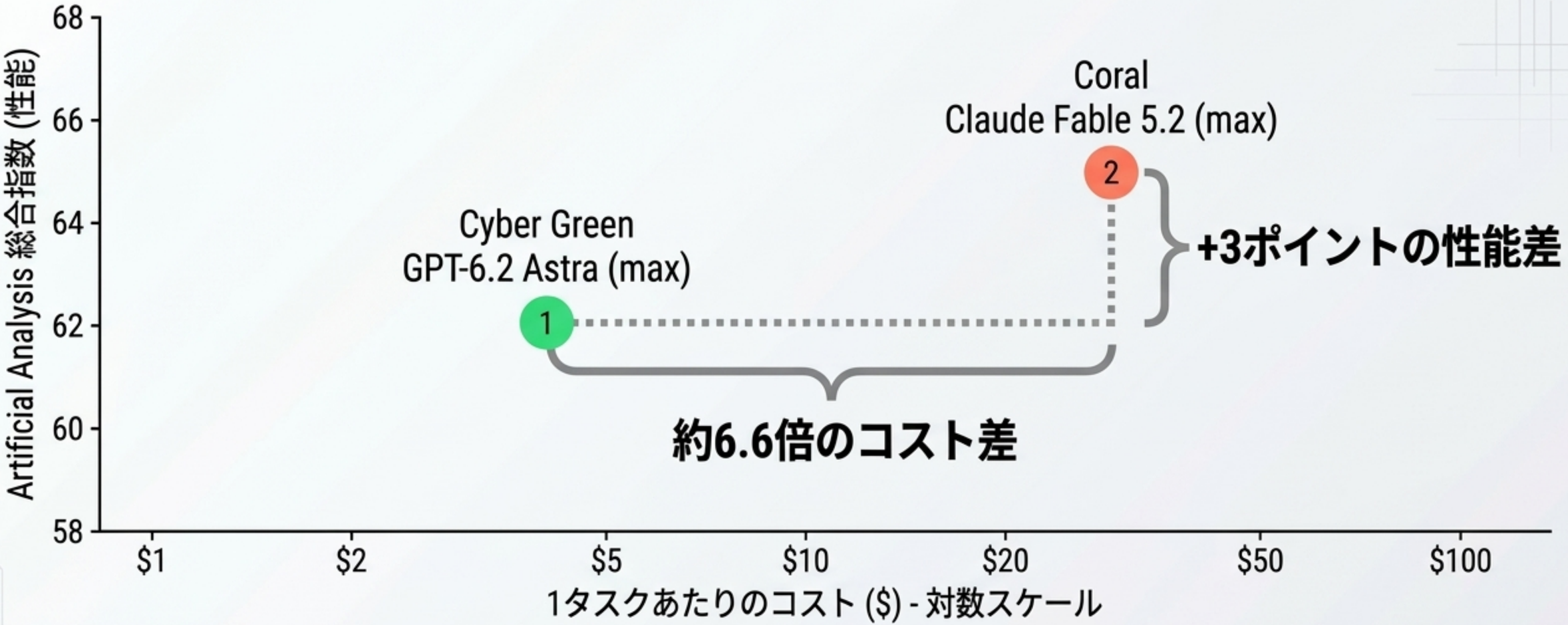
インサイト: AAのCoding Agent Index (v2.1) においてもAstra (53) は前モデルから変化せず、Fable (58) を下回る。コーディング領域の絶対王者は依然としてAnthropic陣営。

Gemini Notebook

[✓ Confirmed]

[☒ Secondary]

コストとパフォーマンスのトレードオフ



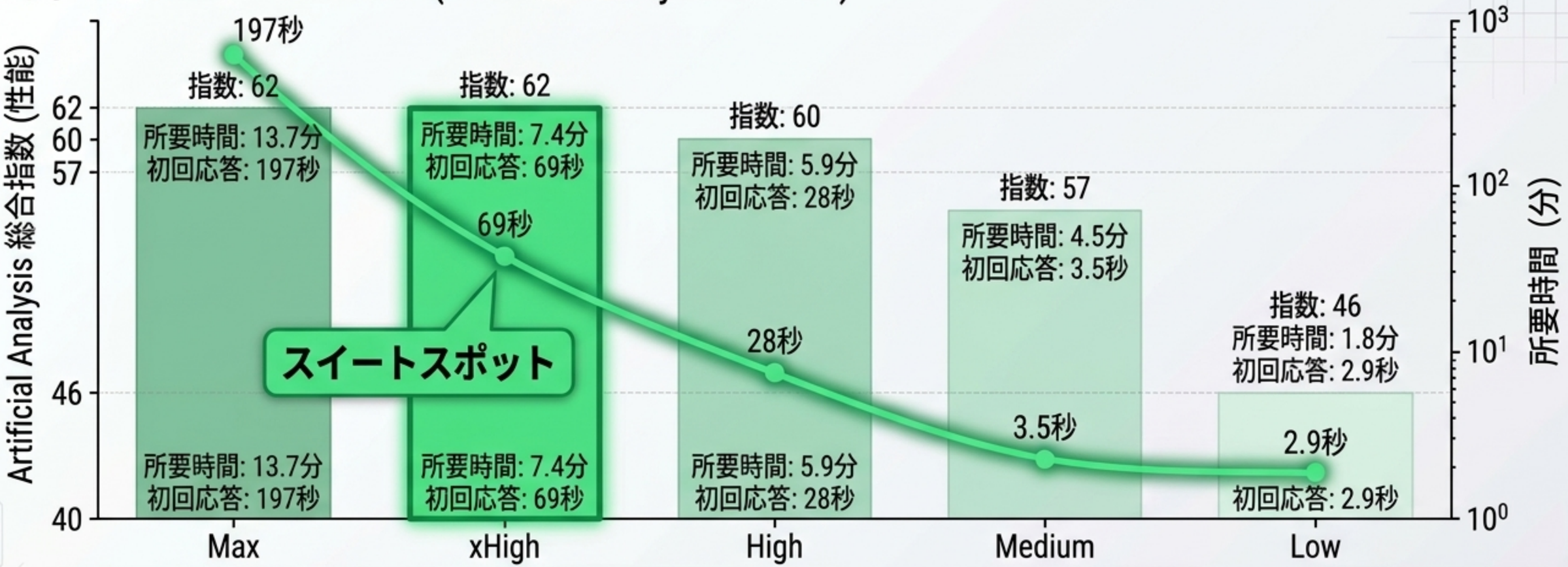
インサイト: +3ポイントの性能を得るために、コストは6倍跳ね上がる。Astraのコストパフォーマンスは特筆に値する。

Gemini Notebook

[✓ Confirmed]

推論努力（Inference Effort）がもたらす変化

対象モデル: GPT-6.2 Astra (Artificial Analysis データ)



インサイト: xHighはMaxと同じ性能（指数62）を出しながら、時間とトークン消費を半分以下に抑える。設定によるトレードオフの最適化が不可欠。

Gemini Notebook

[✓ Confirmed]
[☒ Secondary]

「100%」のカラクリ：評価ハーネスの違い

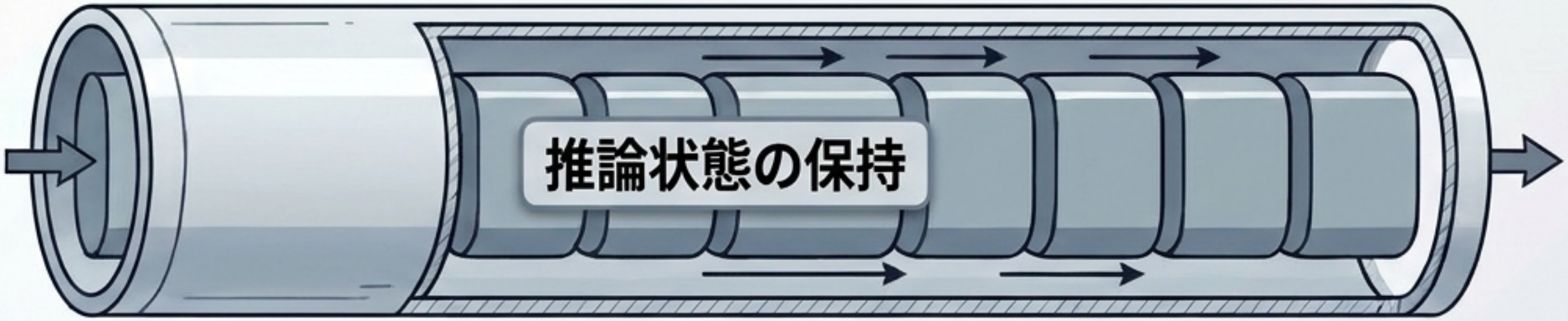
ARC-AGI-3 における検証

標準ハーネス
(Standard Harness)



Astra = 70.3% / Fable = 81.17% (Fableの勝利)

Provider
Adapter



Astra = 100.0% (前人未到の満点)

インサイト: Astraの100%は専用ハーネス（xhigh）の賜物である。測り方を変えれば、Fableが勝つ。評価の数字は環境依存であり、絶対的なものではない。

Gemini Notebook

[✓ Confirmed]
[☒ Secondary]

公開直後の「ボラティリティ（変動性）」

静的な数字と動的な体感のズレ

WebDevスコアの急上昇

発表時 1,781点 → 現在表示 1,797点
(数日で16点上昇)

Vals Indexの順位変動

公開当日 7位 → 現在 8位

ハーヴィー法務テストの矛盾

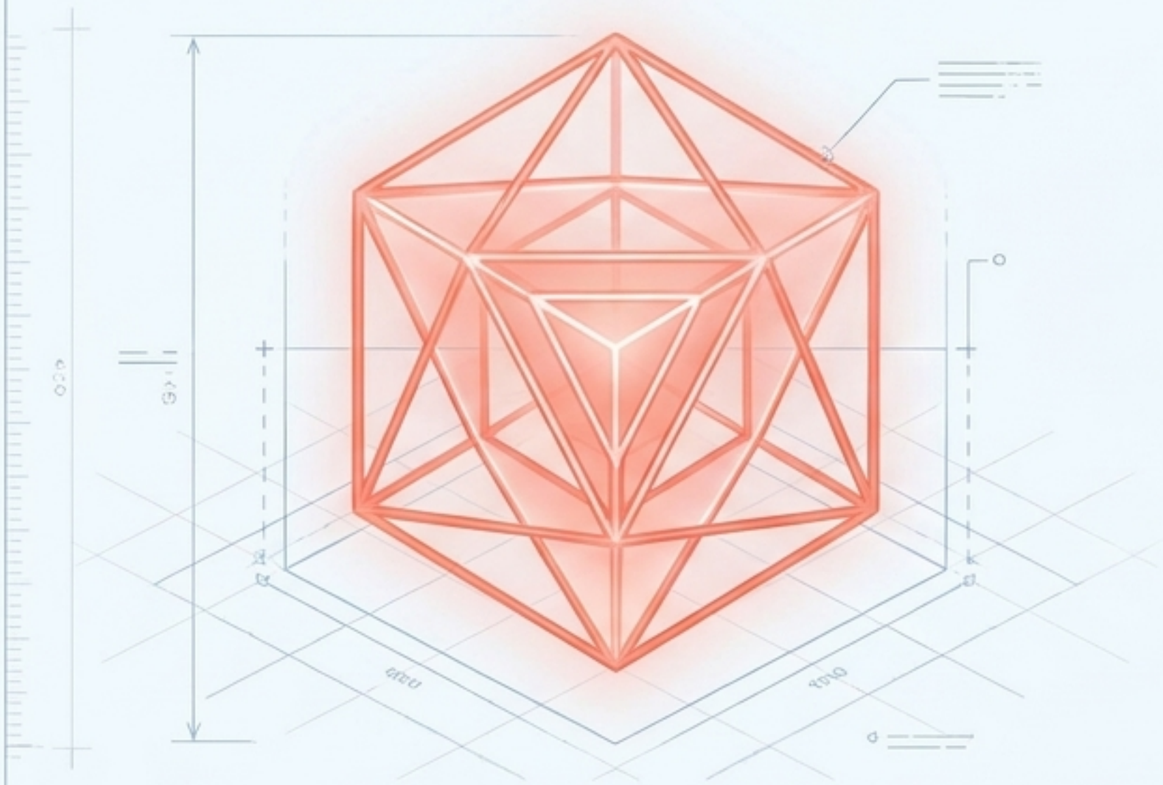
OpenAI発表 (Harvey LAB-AA) ではAstraが最高値だが、Vals設定 (HLAB-Vals)では16.25%で12位に沈む。

市場の体感: VentureBeatが報じた「パワーユーザーの初期の否定的な反応」。ベンチマークの数字と人間の直感的な使い勝手（ハルシネーション率59%など）には確実なズレが存在する。

Gemini Notebook

[✓ Confirmed] [☒ Secondary]

Claude Fable 5.2 & Opus 5.5

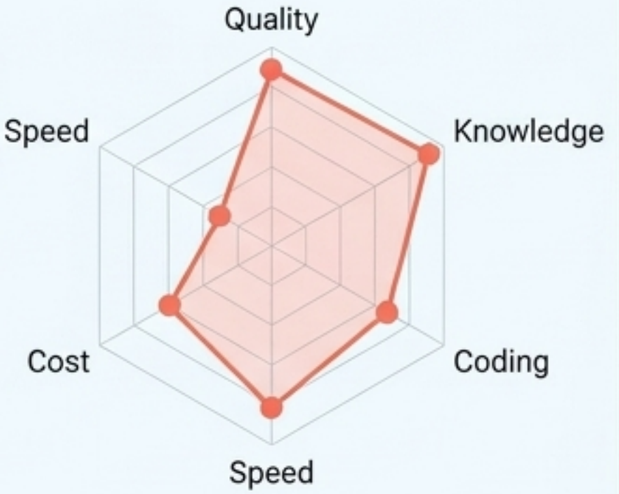


役割: 妥協なき最高品質の追求者
(The Uncompromising Quality Seeker)

- **強み**: オープンエンドな実務タスク、複雑なコーディング、世界知識（SimpleQA 94%）。
- **弱み**: Max設定時のコスト（$26.54/タスク）と推論時間（43分以上）。

最適配置
(Optimal Deployment)

コスト度外視で精度が求められるハイエンドな業務自動化、複雑なデータセンター分析。



Gemini Notebook

モデル・プロファイル：GPT陣営

[✓ Confirmed] [☒ Secondary]

GPT-6.2 Astra



役割: 高難度推論・理数特化エンジン
(The Deep Reasoning Engine)

- **強み**: 科学、数学、未解決問題の突破（Erdős 10%）、圧倒的なコスト効率（$4.00/タスク）。
- **弱み**: 法務・税務などの事実知識ベースのタスク。 Omniscienceでの高い幻覚率（ハルシネーション 59%）。



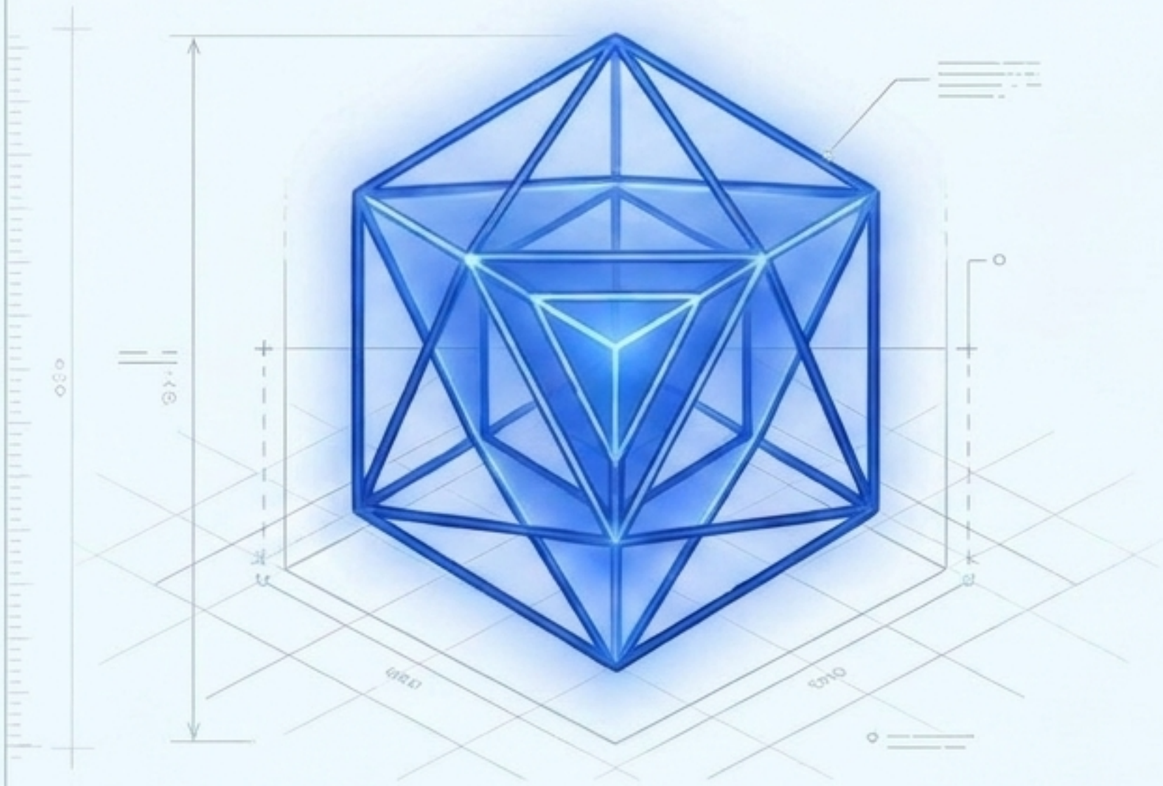
最適配置
(Optimal Deployment)

閉じた環境での論理検証、数学的モデリング、Provider Adapterを活用した長大な推論タスク。

Gemini Notebook

[✓ Confirmed] [☒ Secondary]

Gemini 4 Series (Carbon, Neon, Argon)



役割: 実務特化の高ROIワーカー
(The High-ROI Practical Worker)

- **強み**: 圧倒的な単価の低さ、実務ワークや一般的なチャットタスクでの安定した上位ランクイン。
- **弱み**: FrontierMathなどの極限の抽象推論。

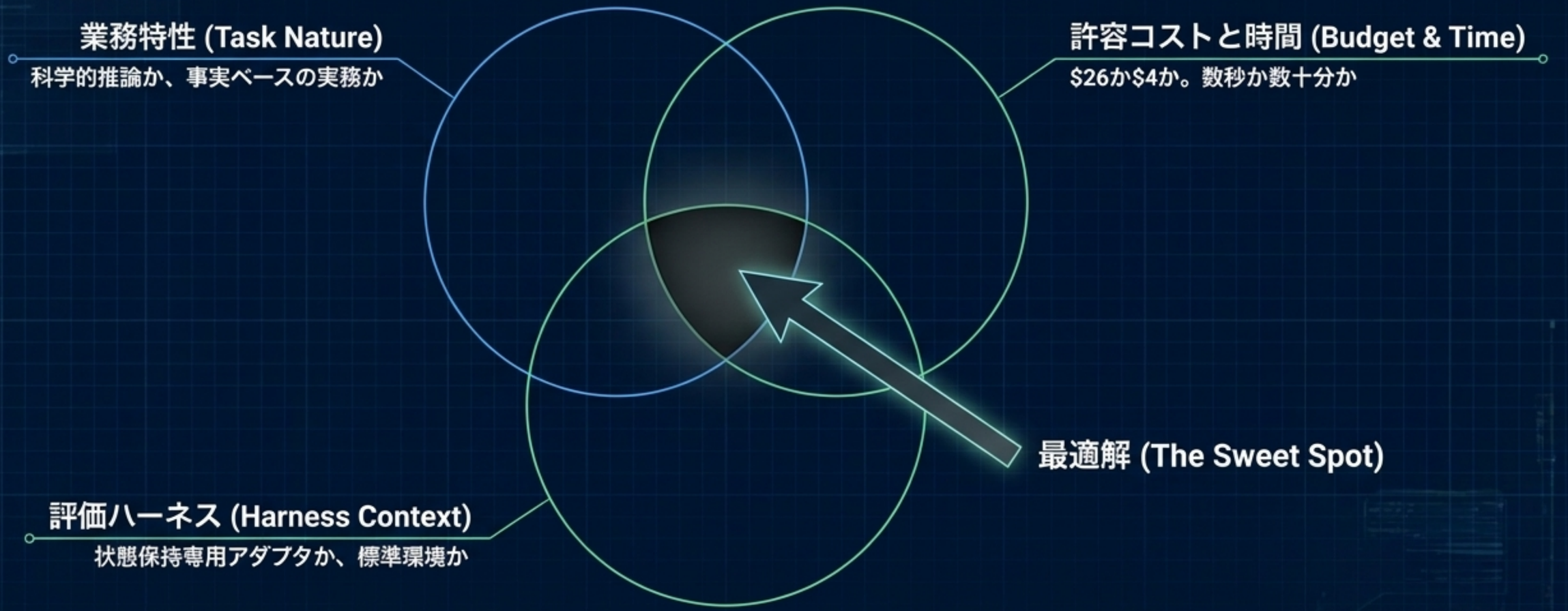
最適配置
(Optimal Deployment)

スケールが求められる大量のルーチン作業、社内AIアシスタントのバックエンド。



[Analysis]

文脈主導のモデル選定へ



結論：「どれが最強か？」という問いは終わった。順位表を横並びにするのではなく、自社の文脈（Context）に合わせた設定の最適化こそが、次世代AI戦略の鍵である。

Gemini Notebook