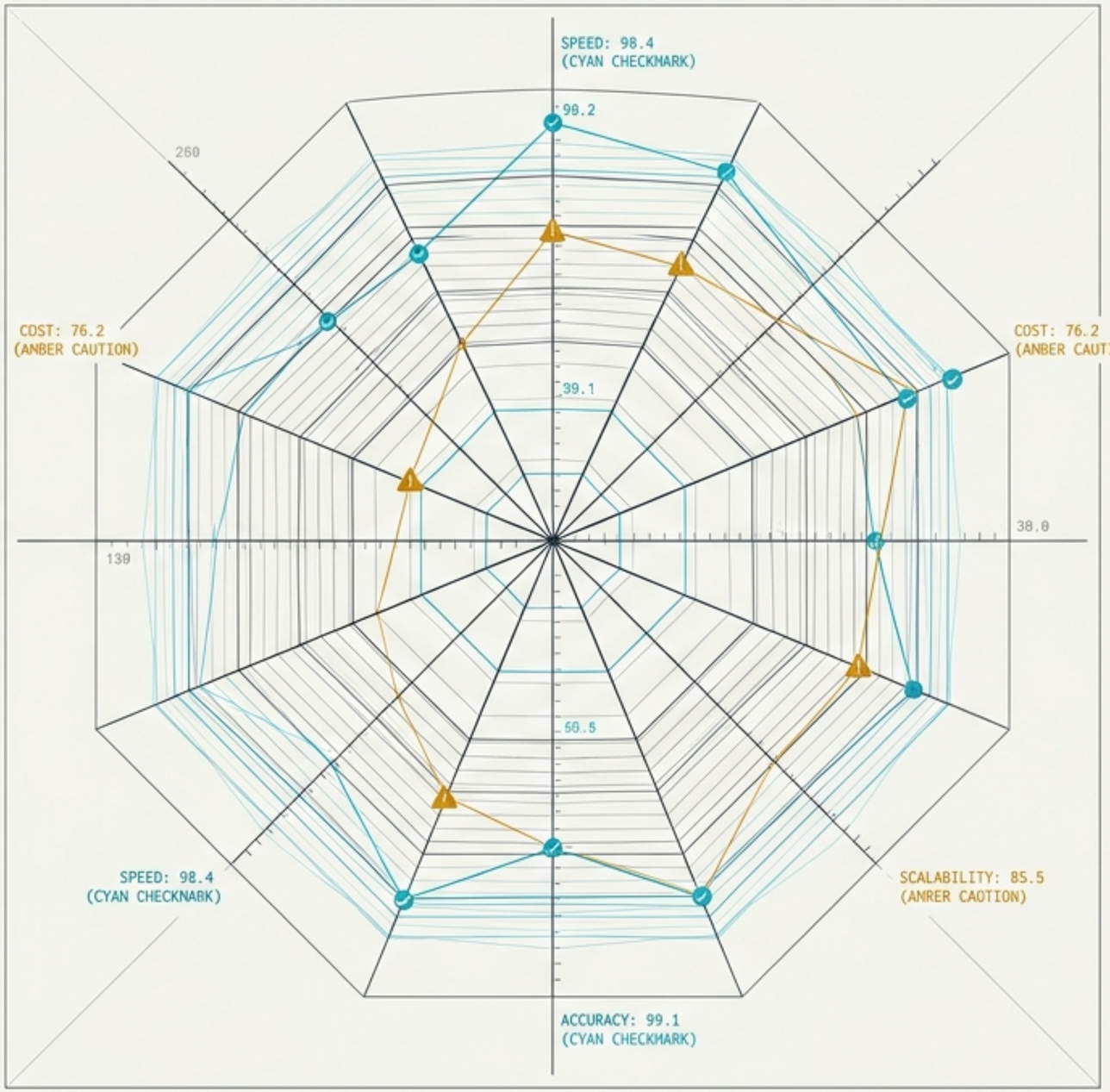
[CLASSIFICATION: STRATEGIC BRIEFING]

Artificial Analysis Index v4.2 解析レポート

評価指標の地殻変動と、AI調達実務への真の示唆



40%

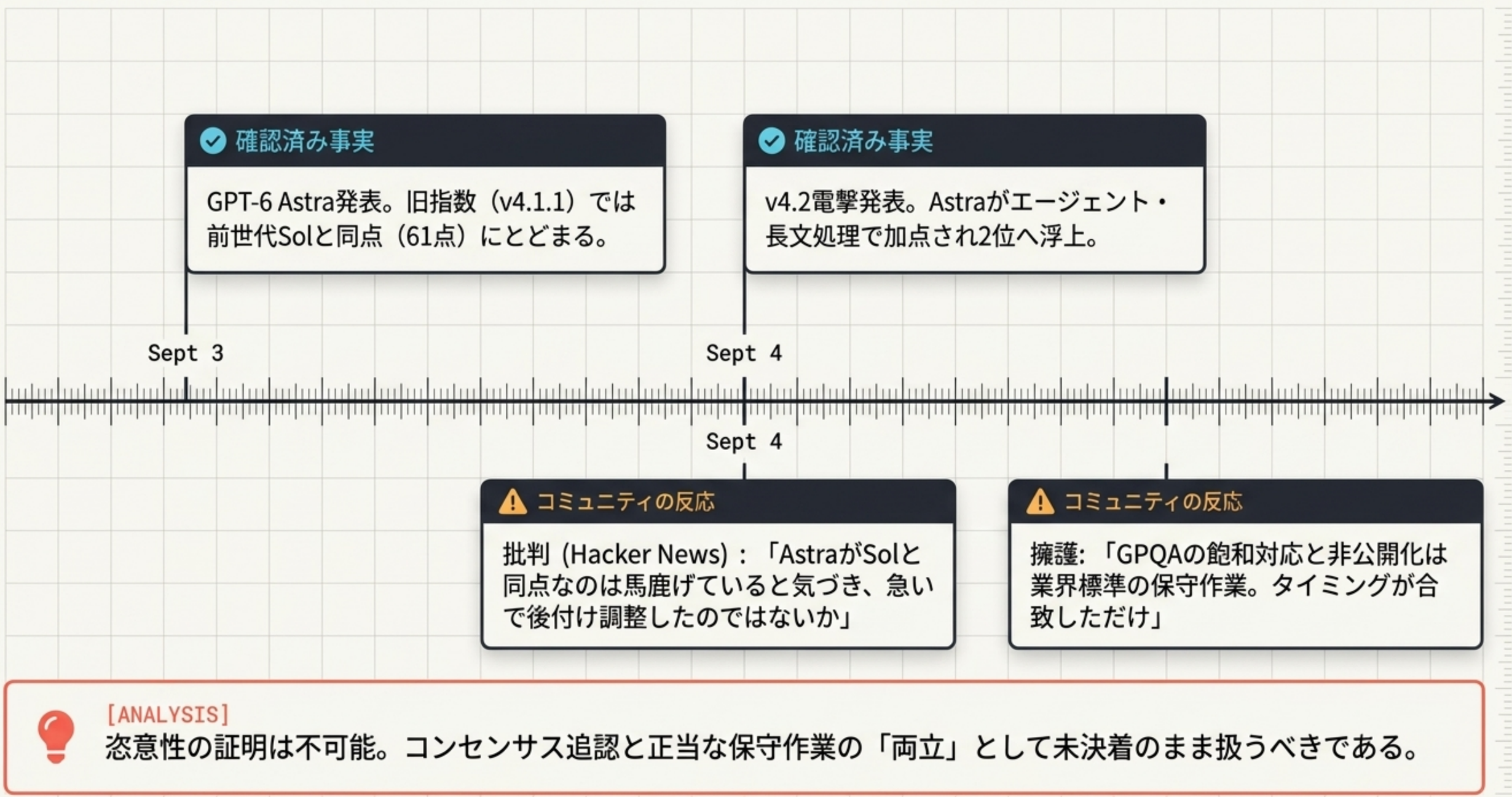
非公開（held-out）テストの比重が倍増。
コンタミネーション対策として「透明性」から
「ゲーミング耐性」へ舵を切る。

57pts

新たな首位はClaude Fable 5.1。GPT-6
Astraは新設の実務評価（GDP.pdf）で躍進
し、前世代から+4点の55点で2位へ急浮上。

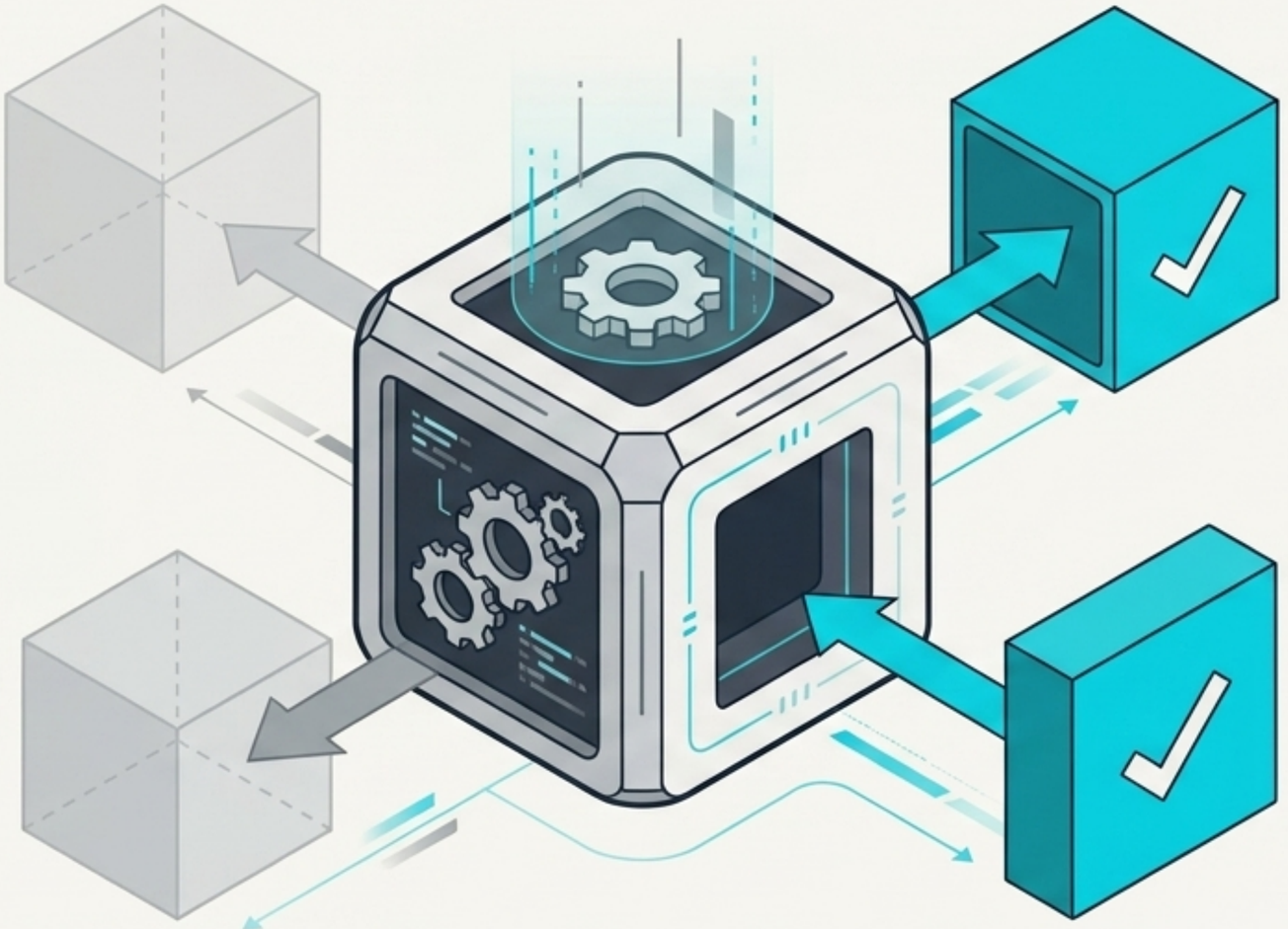
0

単一スコアへの絶対的依存の終焉。
自社タスク（特に知財・実務文書）に基づく
個別検証とサブスコアの活用が不可欠に。



ベンチマーク刷新: 非公開テストの導入とGPQAの除外

GPQA Diamond
飽和により除外



GPQA Diamond
飽和により除外

AA-Briefcase

ウェイト: 15%（指数内最大）
特性: AA自社開発の非公開テスト。複数週にわたる実務プロジェクトを模した長期エージェント型ナレッジワーク。

GDP.pdf

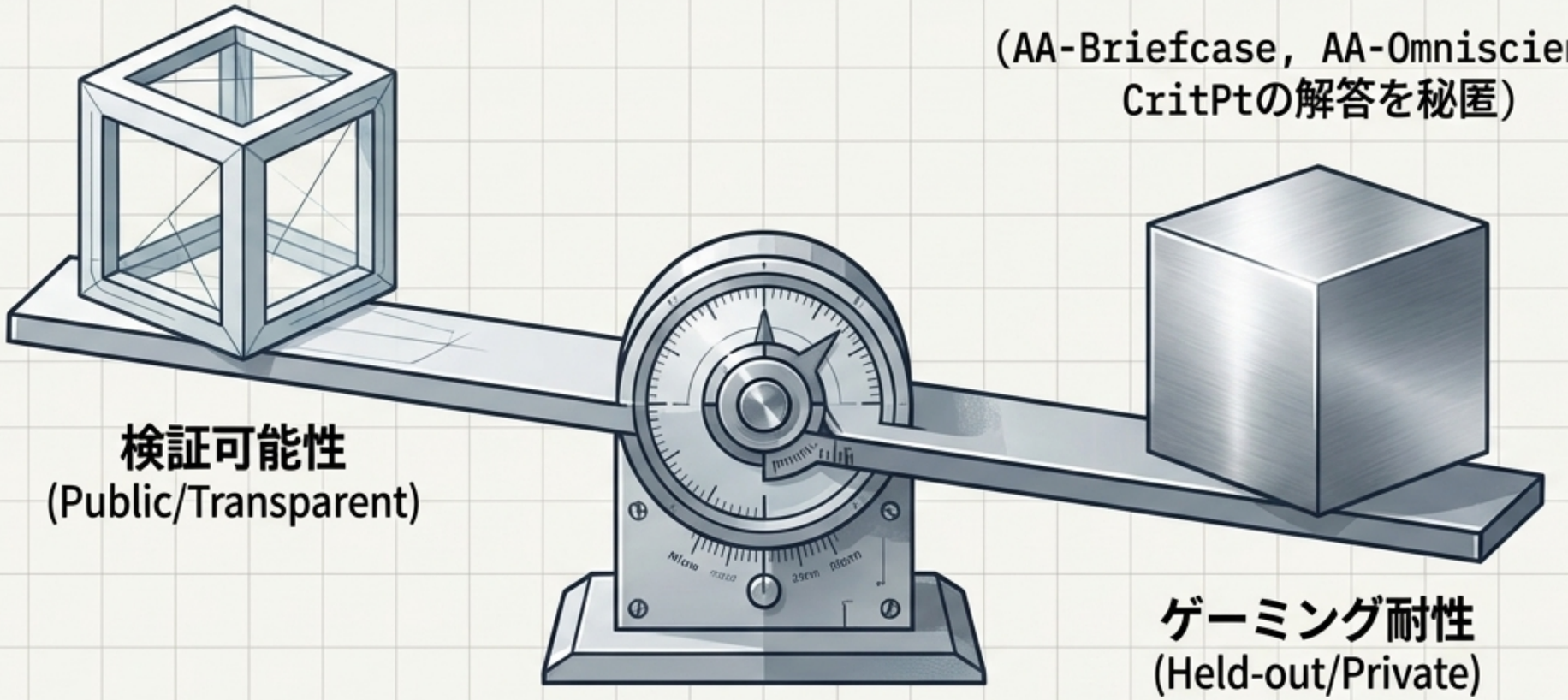
ウェイト: 10%
特性: Surge AI作成の長文PDF読解（100件、4,592ページ、1,275基準）。全基準を満たす「All-pass Rate」で厳格に評価。

マイナーアップデート: AA-LCR v1.1への更新や採点サンドボックスの堅牢化（SciCode）を実施。

v4.1: 20% Held-out

v4.2: 40% Held-out

(AA-Briefcase, AA-Omniscience, CritPtの解答を秘匿)



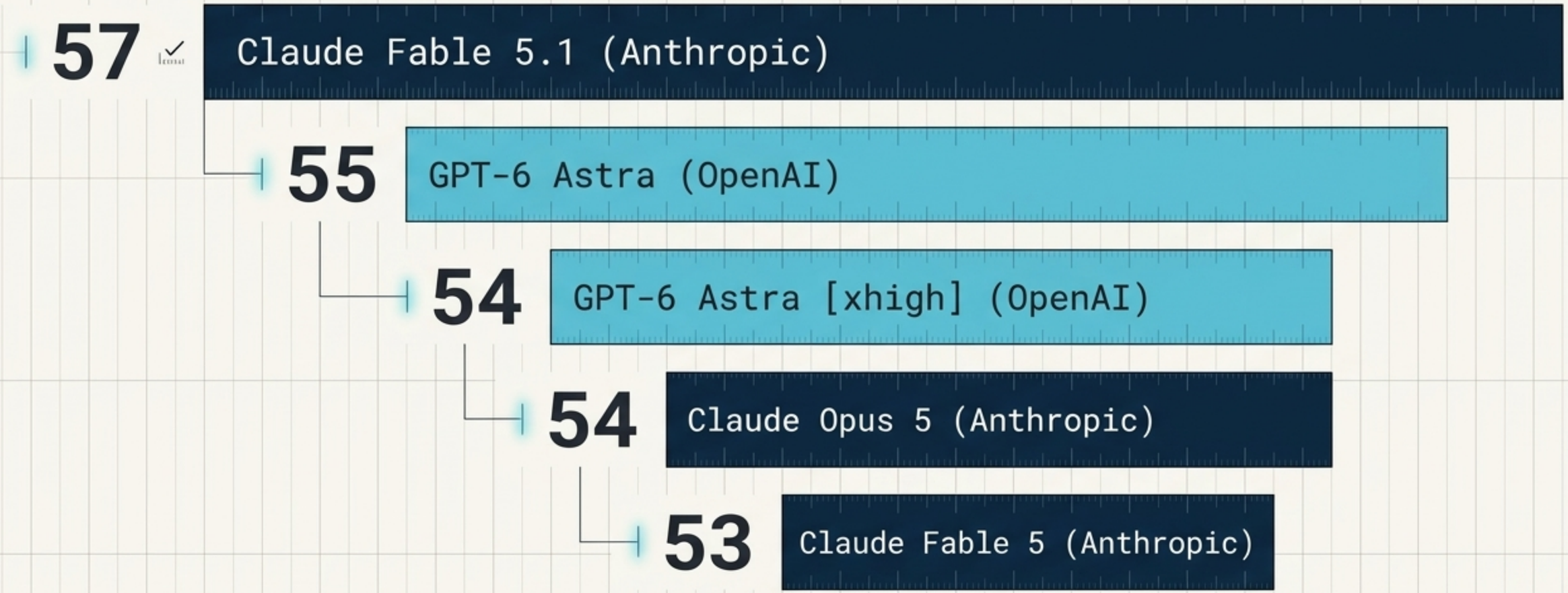
検証可能性
(Public/Transparent)

ゲーミング耐性
(Held-out/Private)

公開ベンチマーク汚染の深刻化。Scale AIのGSM1k論文では、学習データ混入により最大8%〜13%の精度低下（過学習）が報告されている。

独立再現性が重視される実務（医療・法務など）において、非公開テストへの過度な依存は新たな「監査リスク」を生む。

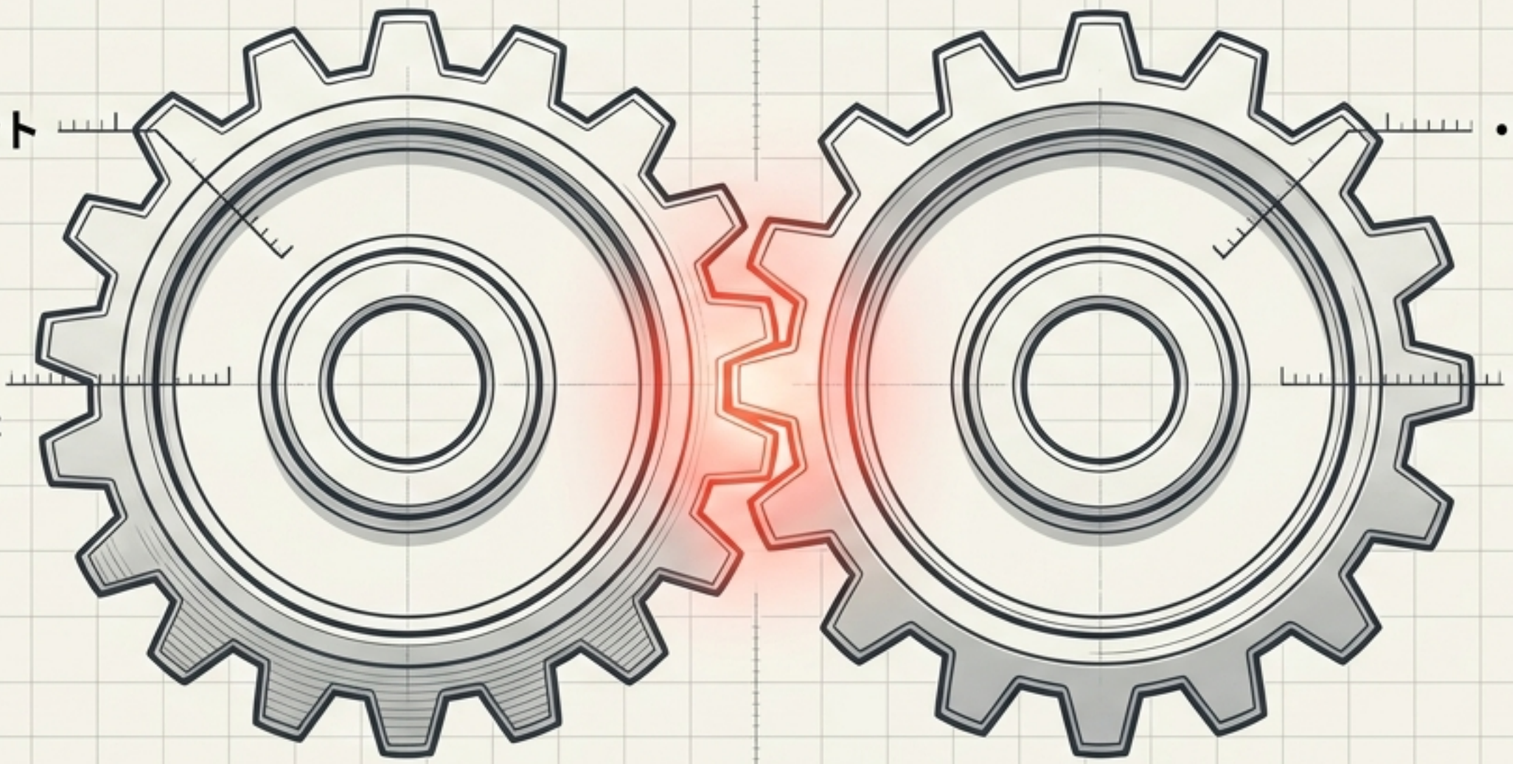
ベンチマーク序列: AIモデル階層分析



上位4枠中3枠をAnthropicが独占。ラボ別の序列は明確に Anthropic > OpenAI > Meta > SpaceXAI の順で固定化。Google (Gemini系) は最下位グループへ沈む。

GPT-6 Astraの特性

- 強力なエージェント処理能力
- 大規模な実務文書（PDF）の精緻な読解



v4.2の新評価軸

- AA-Briefcaseでの高スコア (約+85 Elo)
- GDP.pdfでの圧倒的優位 (Astra 33.2% > Sol 28.2% > Fable 5.1 26.2%)

[ANALYSIS]

旧評価（61点、Sol同点）から+4点の飛躍は、モデルの真の性能向上というより、評価システム側がAstraの得意領域（実務ワークロード）に「ピントを合わせた」結果である。

競争軸のシフト。トップ層は米国のプロプライエタリが独占する一方、オープンウェイト領域ではZ.AIやMoonshot等の中国勢が中上位を確保。国別競争からオープン性の競争へ。

Artificial Analysis

- **評価:** GPT-6 Astraは2位（旧指数では前世代と同点）。
- **特性:** 複数テストの合成。テキスト・英語中心。

Epoch AI & ARC-AGI-3

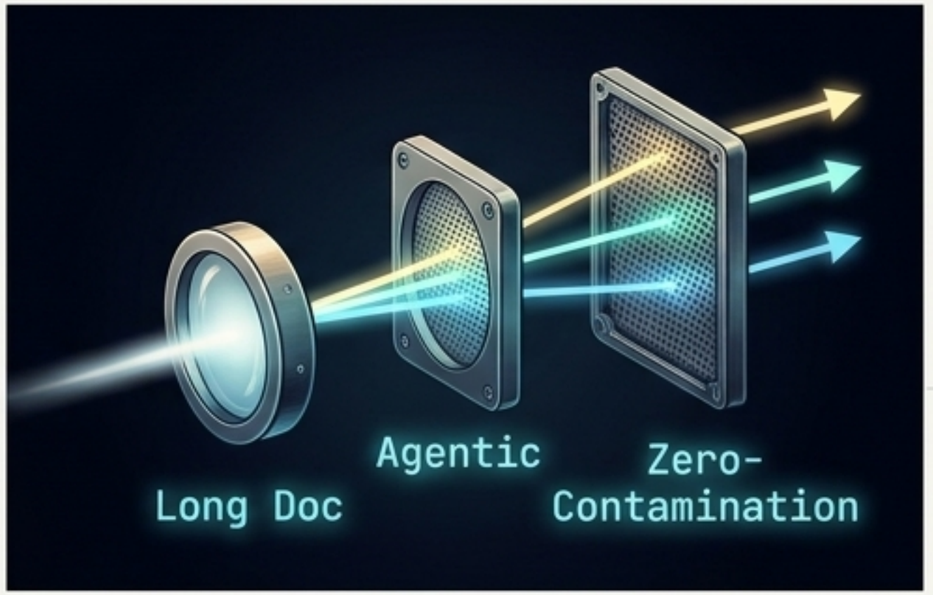
- **評価:** Epoch ECIで169点・267モデル中「1位」。ARC-AGI-3では「人間平均を上回る効率」。
- **特性:** Epochは50超のベンチマーク統合。ARCは未知の推論タスク。

⚠ 単一指標への依存リスク。権威ある独立系ラボ同士が全く正反対の結論に至る現状は、総合スコアが「絶対的なIQ」としては機能不全に陥っていることを示す。

過去のパラダイム:
ベンチマーク＝絶対的なIQ

単一の知能指数としての過信。ゲーミングに弱く、実世界の遅延やエージェントの信頼頼性を捉えきれない。

新パラダイム:
ベンチマーク＝適合性フィルター



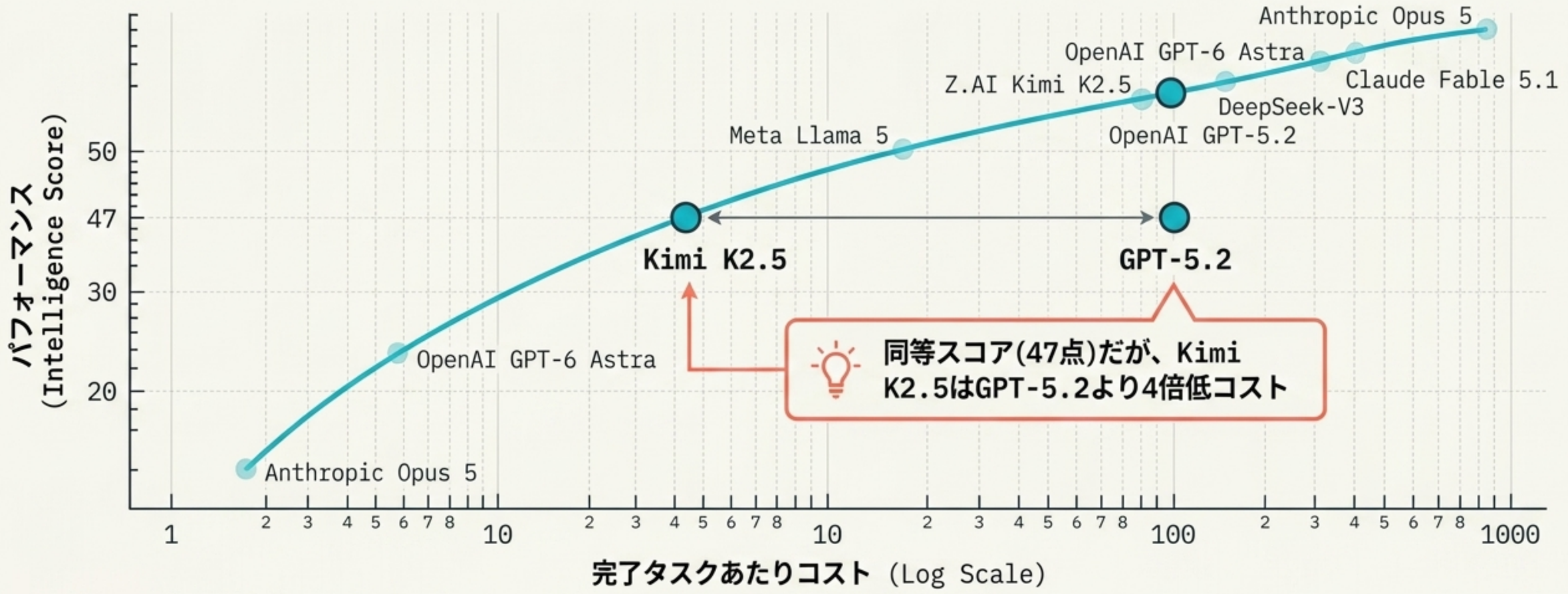
特定ドメイン（長文・エージェント・非汚染） のスクリーニング。

v4.2の真の意義は、ランキングの入れ替えではない。ベンチマークが「賢さの証明」から、「特定ドメインへの適合性スクリーニングツール」へと進化したことにある。

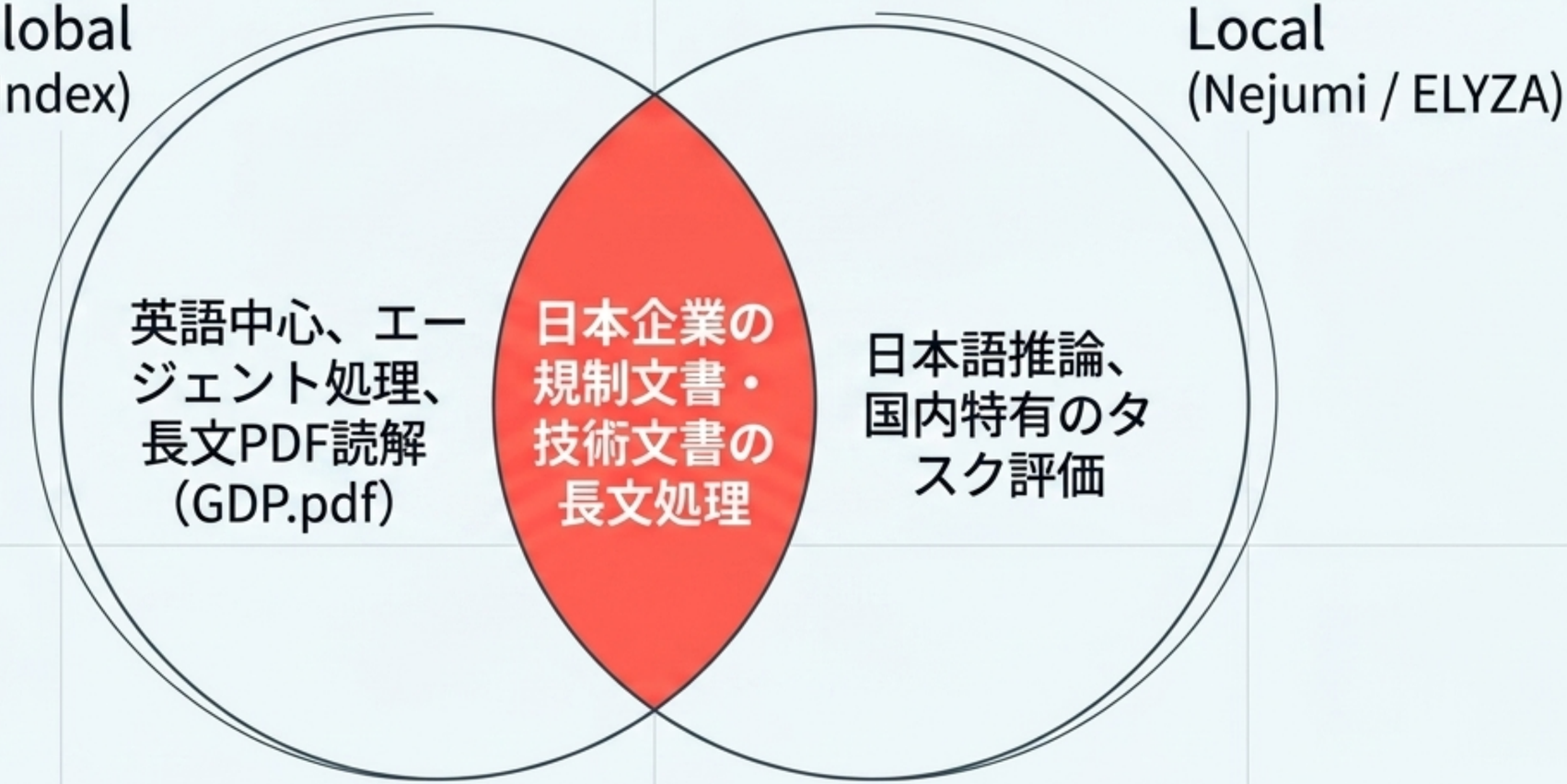
[DIAGNOSTIC LOOKUP TABLE: PROCUREMENT MATRIX]

企業のユースケース	参照すべきAAサブスコア	推奨モデル系統
知財調査・契約書審査	GDP.pdf（長文専門PDF読解）	GPT-6 Astra（圧倒的優位）
長期自律プロジェクト	AA-Briefcase / GDPval-AA v2	Claude Fable 5.1 / Astra
汎用的な推論・ナレッジ	AA-Omniscience / 全体スコア	Claude Opus 5 / 高速なFlash系

[PARETO FRONTIER ANALYSIS: COST-INTELLIGENCE TRADEOFF]



トークン単価ではなく「タスク完了コスト」で評価せよ。知能指数の僅差に対してコスト差は数倍に開くため、コスト重視の現場ではZ.AIやDeepSeekなどのオープンウェイト勢が極めて有力な選択肢となる。



AAの上位に日本発モデルは不在。実務においては、AAの「長文処理・エージェント性能」の指標と、国内リーダーボードの「日本語のニュアンス理解」を掛け合わせて評価するハイブリッド・アプローチが必須。

[AI EVALUATION PRINCIPLES: FOUNDATIONAL GUIDELINES]

1. 一次情報の直接確認

✓ 旧v4.1.1のスコアが混在する二次情報に注意。AAサイトで「v4.2再採点済みか」を必ず確認。5点差未満は実用上「誤差」とみなす。

2. サブスコアの徹底重視

✓ 総合指標は一次フィルタに過ぎない。用途に応じ、GDP.pdfやAA-Briefcaseなど個別スコアを深掘りする。

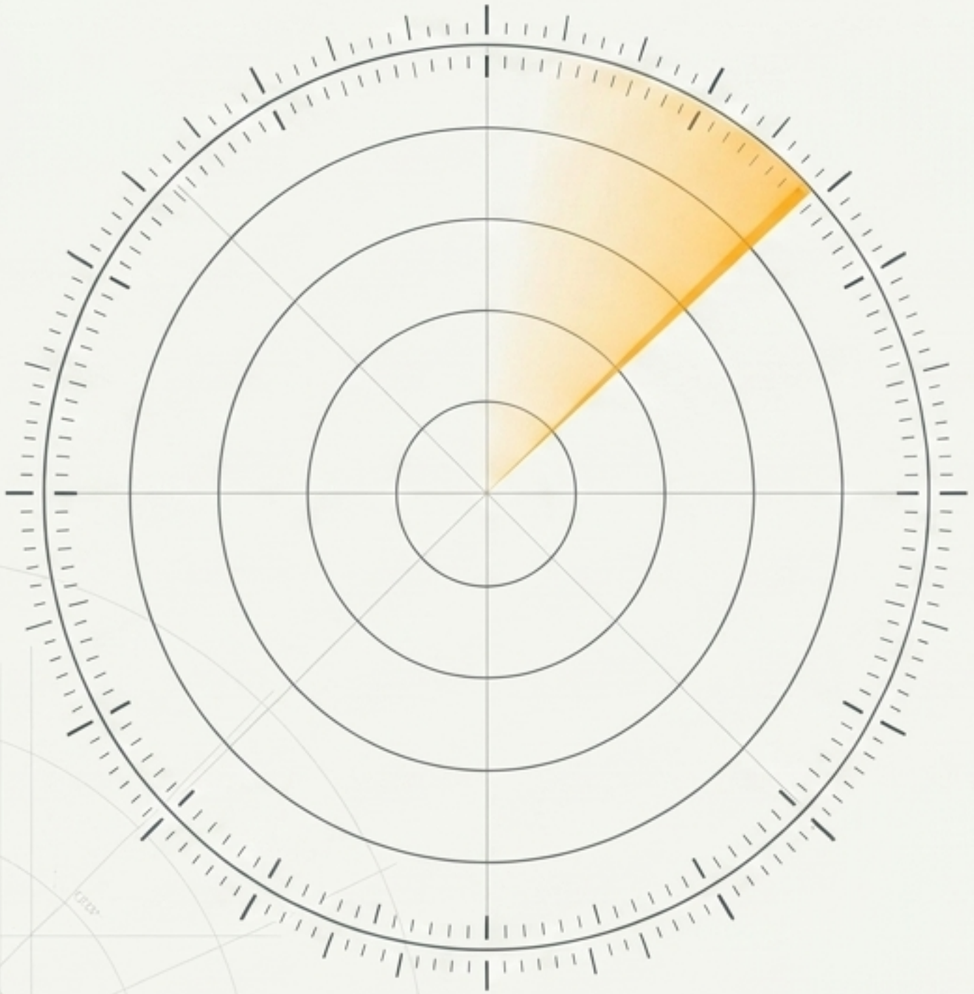
3. タスク単位のコスト評価

高性能モデルに伴う推論トークン増加とレイテンシ悪化を計算に入れる。

4. 私的評価セットの構築

非公開テスト40%化によるブラックボックス化を防ぐため、自社専用の二重検証環境を持つ。

再評価のトリガー (Warning Signals)



⚠ 非公開比率のさらなる上昇

AAAAが予告するv5リリースにおいて非公開比率がさらに引き上げられた場合、パブリック指標の透明性は限界に達する。

⚠ 再採点による大幅な順位変動

既存モデルがv4.2基準で再採点され、5点以上の大規模なスコア変動（再ランク付け）が発生した時。

⚠ 開発企業からの公式な反論

AnthropicやOpenAI等の主要ラボが、新評価手法（特にLLLM-as-judge方式のバイアス等）に対して公式な反論や独自の対抗指標を提示した時。

[END OF DOSSIER. CONTINUOUS MONITORING REQUIRED.]