未知を測る機関：ARC-AGI-4と自律的イノベーションの評価設計

GPT-6 Astraの突破が示す、AIシステム評価のパラダイムシフト

[STATUS: ACTIVE] [DATE: 2026.09.13] [MODE: ANALYSIS]

SOURCE: ARC Prize X (2026.09.12)

名称: ARC-AGI-4

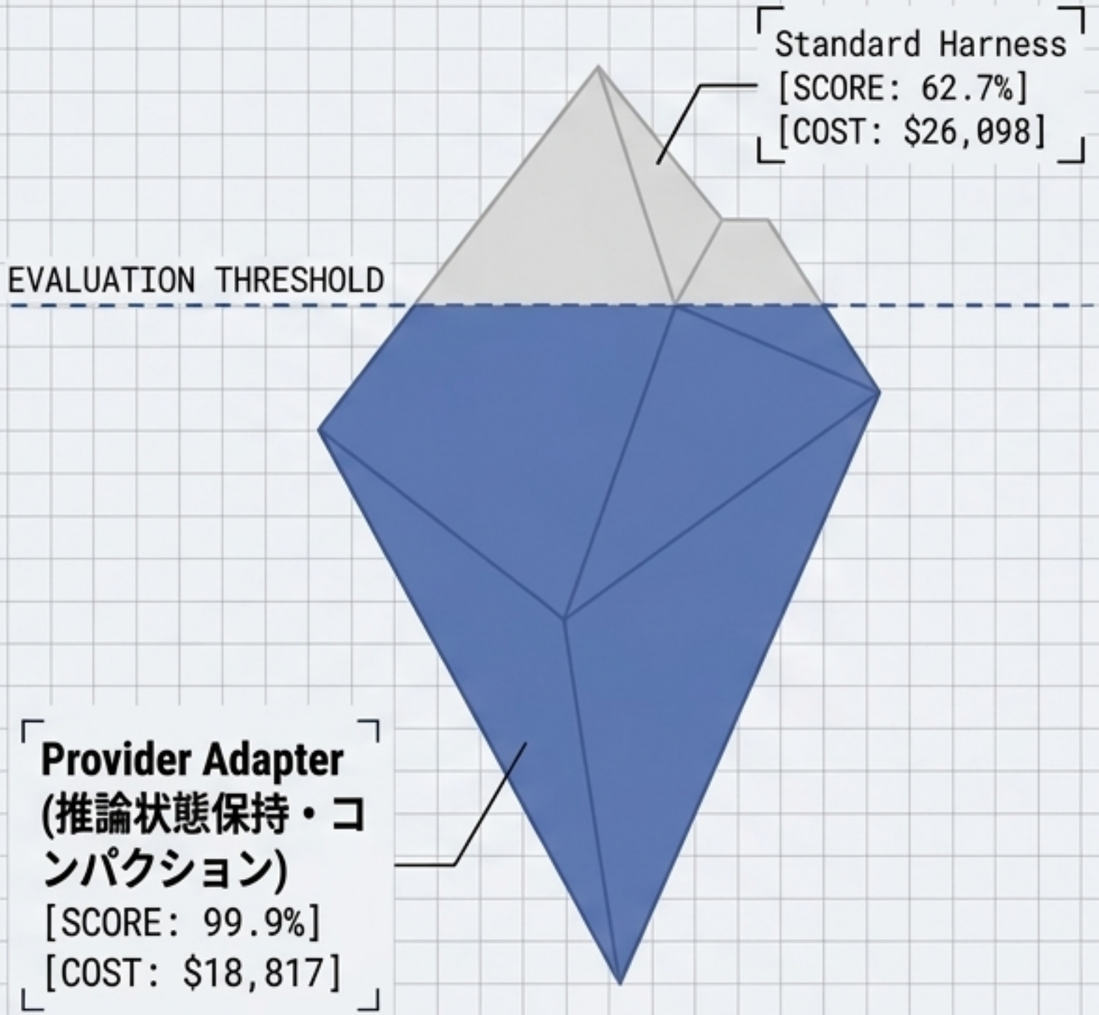
上位目的:
Autonomous open-ended innovation （自律的なオープンエンド・イノベーション）

方針: オープンソース路線の継続

未公開領域

タスク形式、スコアリング式、合格基準、リリース日は現時点で未指定

本報告の提言:
「終わりなき発見」をどう有限時間で定量評価するか。単一スコアを排し、安全性・新規性・妥当性の多条件ゲート化を提案。



事象

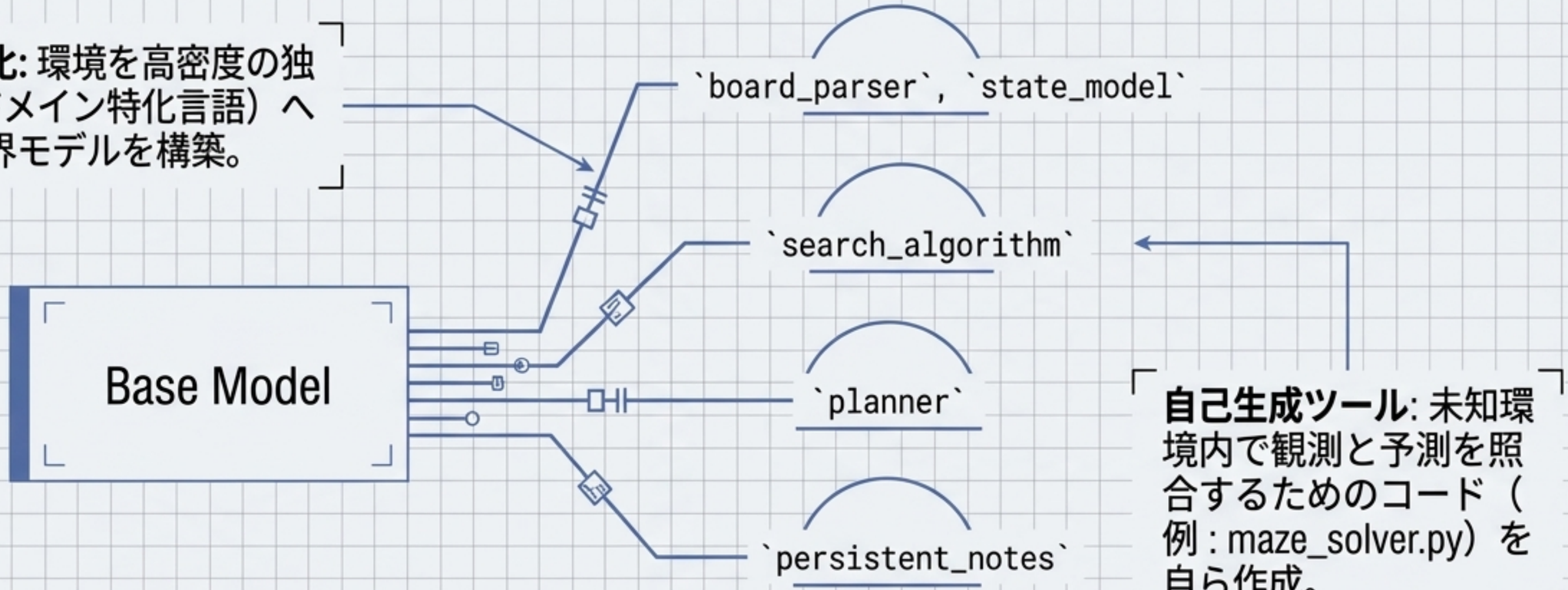
中立的なStandard条件で62.7%だったGPT-6 Astraは、モデル固有の状態保持を許すProvider Adapter条件で99.9%に到達し、消費トークンも49%削減した。

結論

長期エージェント評価において、モデル単体の知能と、ハーネスの記憶・推論設計はもはや分離できない。

AGENT ARCHITECTURE BLUEPRINT

圧縮と記号化: 環境を高密度の独自とDSL（ドメイン特化言語）へ圧縮し、世界モデルを構築。



自己生成ツール: 未知環境内で観測と予測を照合するためのコード（例 : maze_solver.py）を自ら作成。

ARC-3の限界:
「tightly bounded, deterministic, closed-ended」。ルール同定後は単なる実行問題に還元され、新しい問題は生まれない。

Evolution of Intelligence Evaluation

ARC-AGI-1 & 2 (静的推論)

- 主対象: 少数例からの規則推定
- 情報取得: 与えられた例を読む
- 弱点: データ汚染・合成訓練

ARC-AGI-3 (エージェント学習)

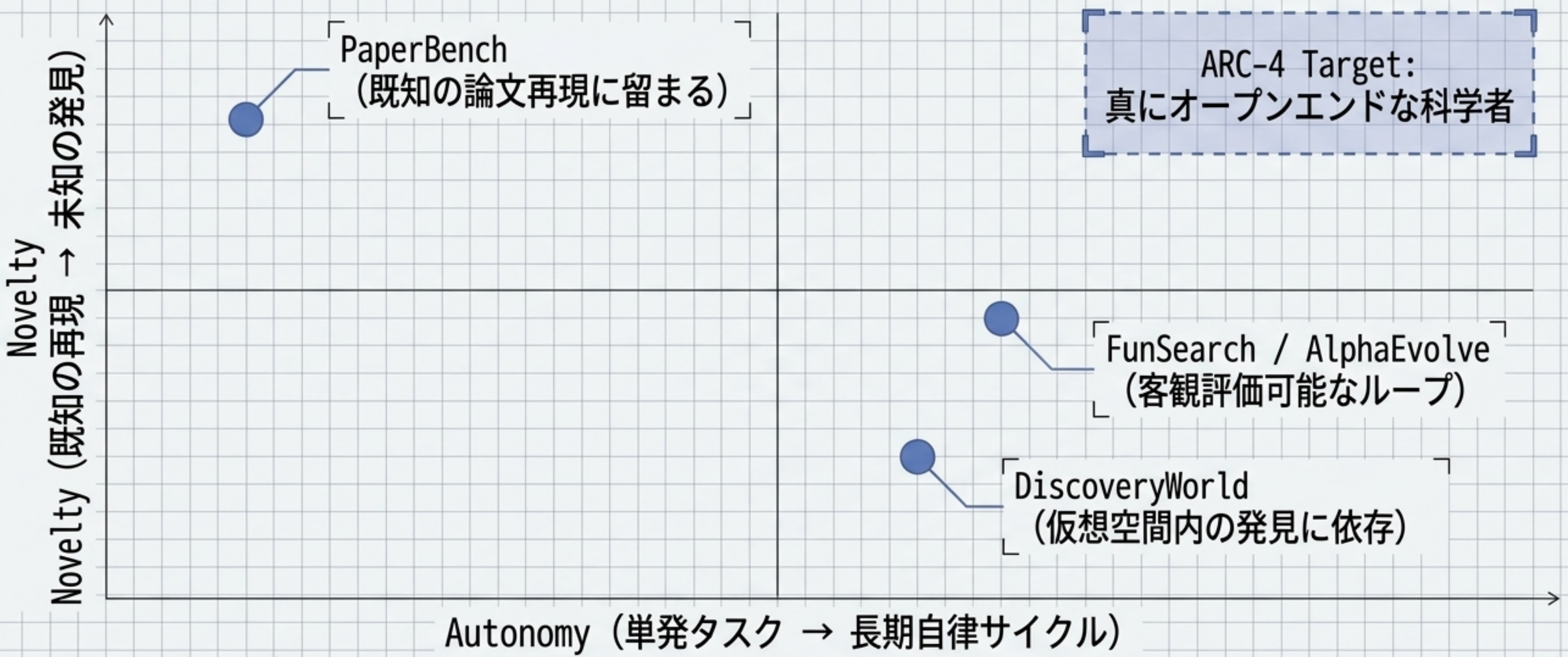
- 主対象: インタラクティブ環境
- 情報取得: 行動して探索
- 弱点: Closed-ended, ハーネス依存

ARC-AGI-4 (オープンエンド)

- 主対象: 自律的なイノベーション
- 情報取得: 終わりなき発見
- 状態: 未開拓領域

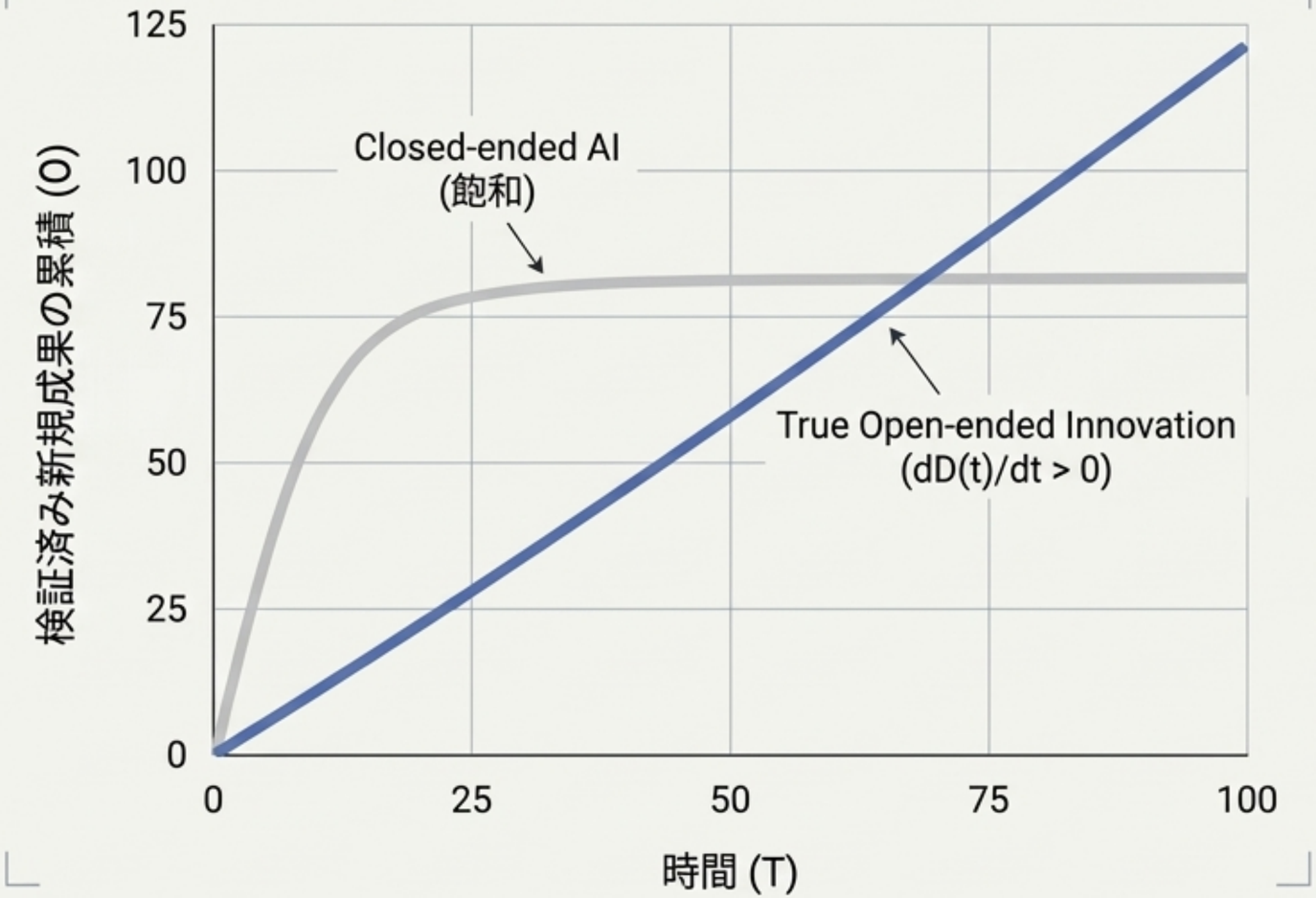
Insight: ARCの本質は「知識の量」ではなく「未知課題への学習効率」にある。ARC-4は自動論文執筆ではなく、未知問題に対し、いかに少ない経験で新しい有効な考えを獲得できるかを問う。

2D Scatter Plot



既存のアプローチ（再現力・科学的手法・新規候補生成）を統合する空白地帯がARC-4の狙いとなる。

The Saturation Curve

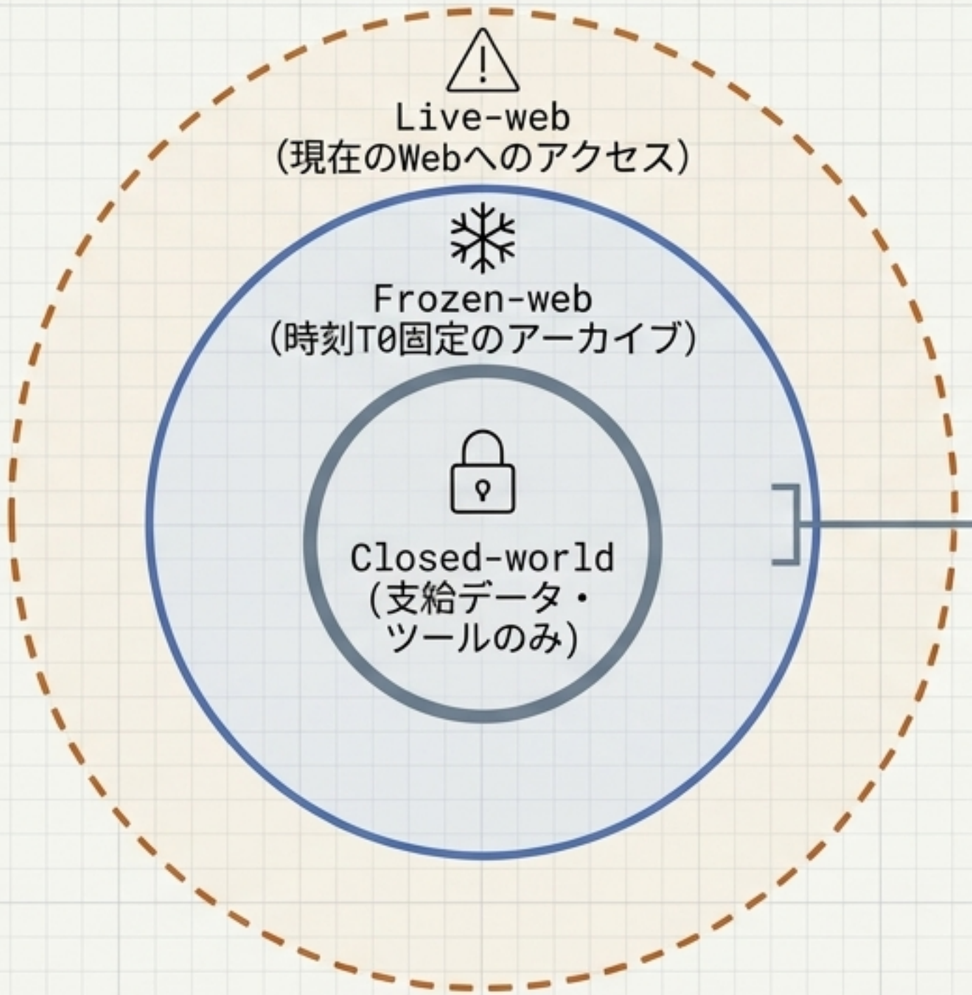


定義

イノベーションとは「時間あたりに生まれる独立検証済み新規成果の持続率」である。新しい問題分布が追加されても、右肩上がりの探索を継続できるかを測る。

METR Time Horizonからの着想

能力の一点突破ではなく、同一予算内での「時間に対する改善曲線」を比較することが、ARC-4の最適解となる。



INFORMATION CONTAINMENT RINGS

ARC-4における最重要要件：Frozen-web トラック

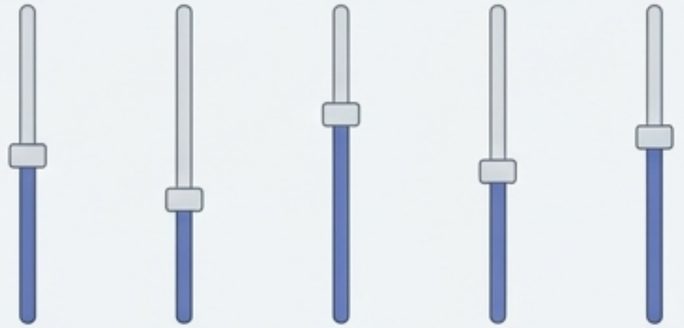
モデルが「評価期間中に本当に新しく発見したのか」、それとも「既に公開された既知のアイデアを検索しただけなのか」を区別するための絶対条件。

$$C = 100 \times S \times GM(V,R,N,U,G) \times E$$

Safety (S)

ハードゲート。違反があれば全体スコアが0になる。

Geometric Mean (GM)



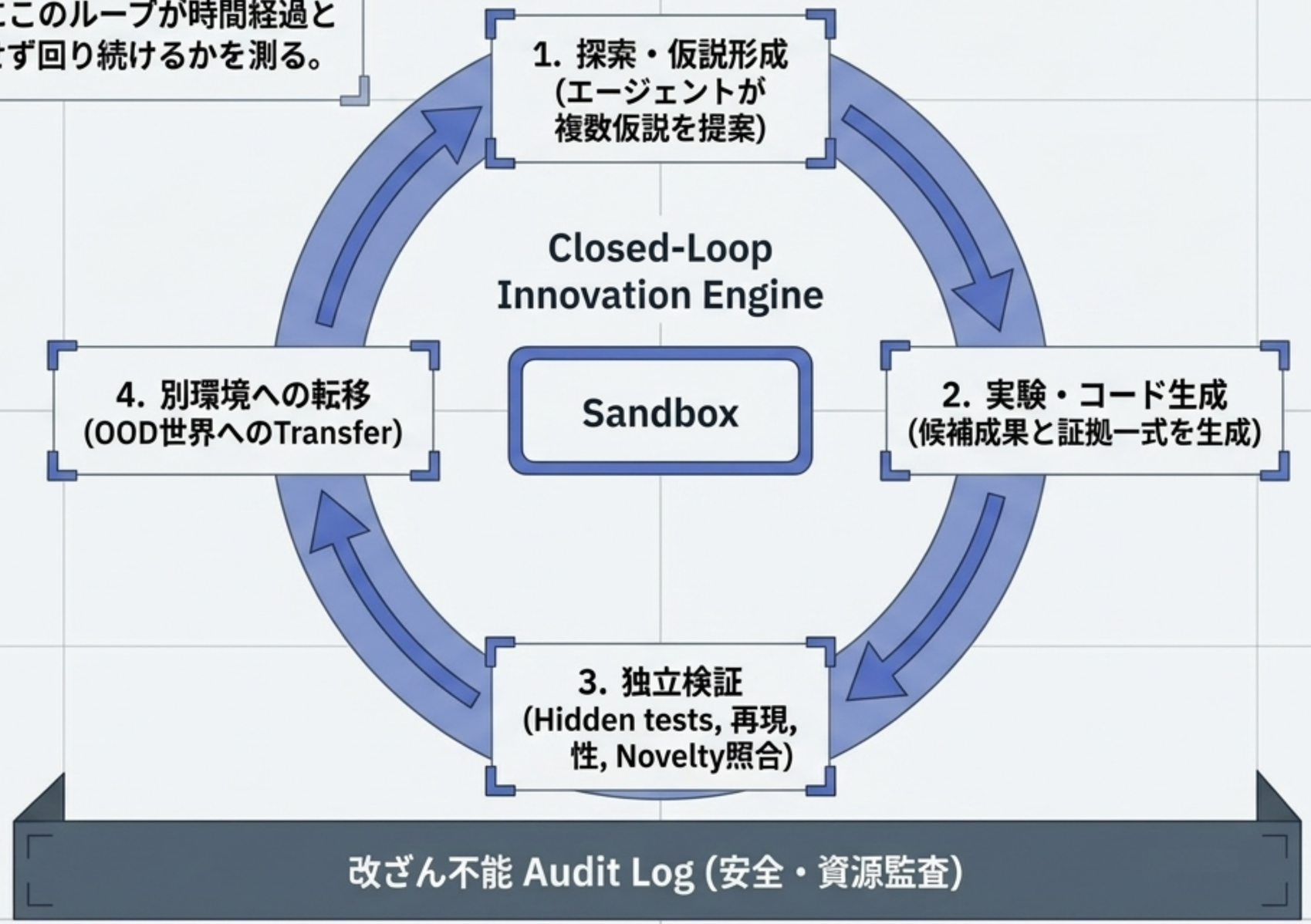
一つの指標が極端に低い場合（例: 新しいが間違っている）、幾何平均により全体スコアが急減。

Efficiency (E)

消費資源（GPU、電力）に対する効率性マルチプライヤー。

単一スコアは評価条件を隠す。「無意味な新アイデア」や「極めて有用な既知の解」を誤判定させないため、幾何平均とハードゲートの組み合わせが必須。

評価対象は一問一答ではない。
人間の介入なしにこのループが時間経過とともにSurationせず回り続けるかを測る。



Dual-Use Threat

自律発見能力は、材料・医薬品だけでなく、サイバー脆弱性、生物・化学設計、AIの自己改善に直結する。

Astraの現実

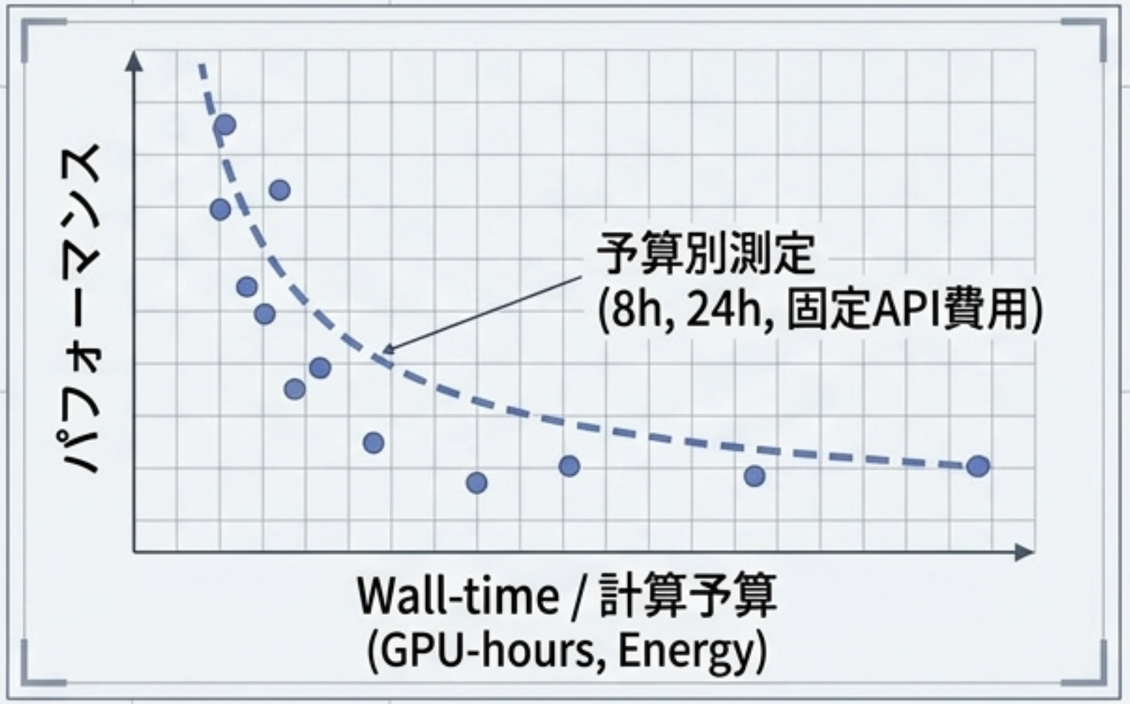
OpenAIのSystem Cardにおいて、GPT-6 Astraは同社初の「Critical cyber capability」モデルと位置づけられている。

System Integrity

発見を行うAgentと検証するEvaluatorの完全なネットワーク分離、およびPolicy違反のハードゲート化が不可欠。

結論: 報酬設計の誤りは虚偽証拠の生成を促す。
「能力評価」と「安全性評価」は同一環境で同時に測定されなければならない。

$$[\text{System Performance}] = \text{Model} \times \text{Harness} \times \text{Tools} \times \text{Reasoning Effort} \times \text{Budget}$$



「モデルXは何%」という単一スコアの時代は終焉。AI と人間の優劣は許される時間によって逆転する。

環境負荷（総kWh、推定CO_2e）も独立軸とし、「同じ科学成果を出すのに必要なエネルギー」を可視化すべきである。

Short-term（1年）

競争の主戦場が「基盤モデル」から「モデル＋コンテキスト管理＋ツール連携」の総合システムへと移行する。

Mid-term（数年）

AIの位置づけが、回答生成から「Research Operating System(仮説立案〜検証の自動化)」へと進化する。

Long-term

人間と同等のオープンエンド・イノベーションが実証され、科学政策や知財制度に抜本的な変革をもたらす。

Policy & Deployment Requirements

規制をARC-4の「一点スコア」に依存してはならない。独立機関による評価（TEVV）、能力とリスクの同時測定、システム評価条件の完全開示が次世代AI展開の絶対条件となる。