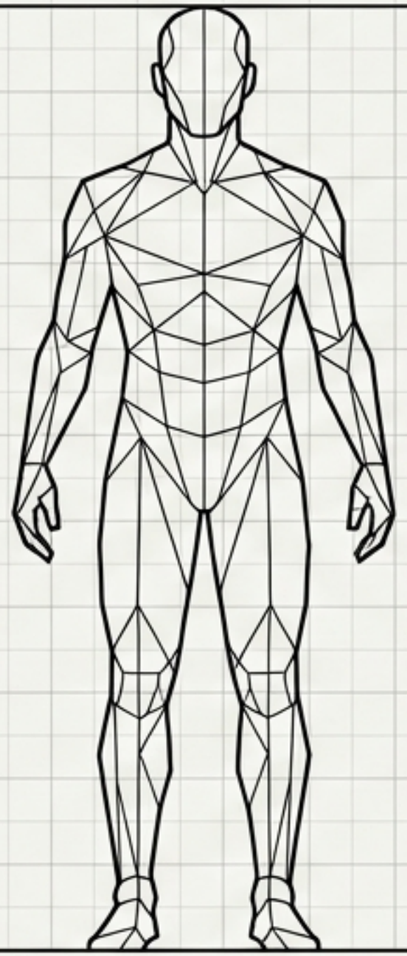
ARC-AGI-3 検証・技術分析レポート

フロンティアAIと人間の「学習効率」を測る新基準

Document ID: REPORT-ARC3-2026
Target: Technical Strategy & Policy
Status: Verified

驚異的な学習効率のギャップ

「人間には100%解ける未知の環境で、最強のAIは1%も解けない」



100%

人間の完遂率（easy for humans 基準）

<1%

フロンティアAI最高値
（0.37%）

パラダイムシフト

従来のAI評価は「事前知識からの回答」を測っていた。**ARC-AGI-3は「未知の環境におけるインタラクティブな学習効率」を測定する。**

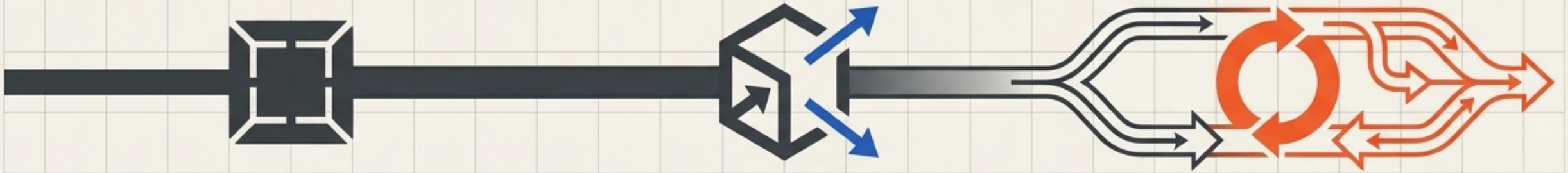
知識量（LLM）から、適応力（エージェント）へ。
評価の軸が根本的に変わる。

知能評価の進化: ARCの系譜とタイムライン

2019年11月
（ARC-AGI-1）

2025年7月〜8月
（プレビュー）

2026年3月
（ARC-AGI-3 ローンチ）



思想的基盤の確立

Chollet氏による定義「知能 = スキル獲得効率」。少数例から変換規則を推定する静的パズルとして登場。

インタラクティブへの移行

30日間の「プレビュー」競争を実施。静的回答から状態空間探索へのシフトが実証される。

「時間的・対話的」新基準の公開

ルールや目標の明示を一切排除した完全ターン制評価へ。

静的知識から動的探索へ：パラダイムの転換

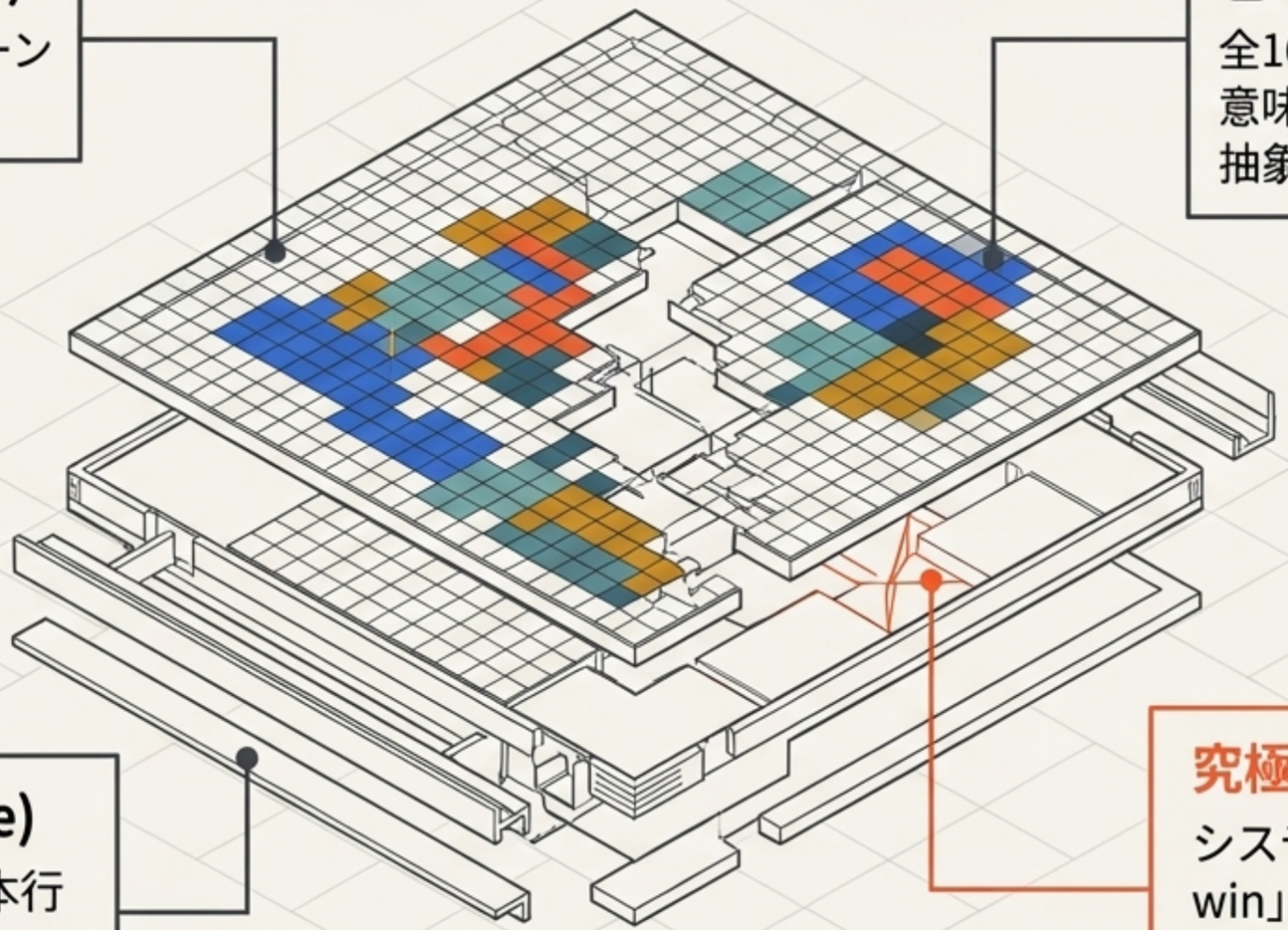
	従来のLLMベンチマーク	ARC-AGI-3
評価対象	事前学習された知識の引き出し	**未知の規則**に対する動的な仮説検証と計画
インターフェース	プロンプトによる一問一答型	**ターン制・マルチターン**の インタラクティブ環境 
ルールの提示	指示やコンテキストが明示される	**説明書・ルール・目標**の明示は一切なし
成功基準	正解率（Accuracy）	人間を基準とした行動効率 （Action Efficiency）

ARC-AGI-3は「知っているか」ではなく、「いかに無駄なく適応するか」を問う。

環境の解剖図: 制約された未知の領域

観測空間 (Observation)
64×64のグリッド空間。ターンごとにフレームを受信。

色と意味 (Semantics)
全16色。文化的記号や言語的意味付けを徹底排除。純粋な抽象空間。

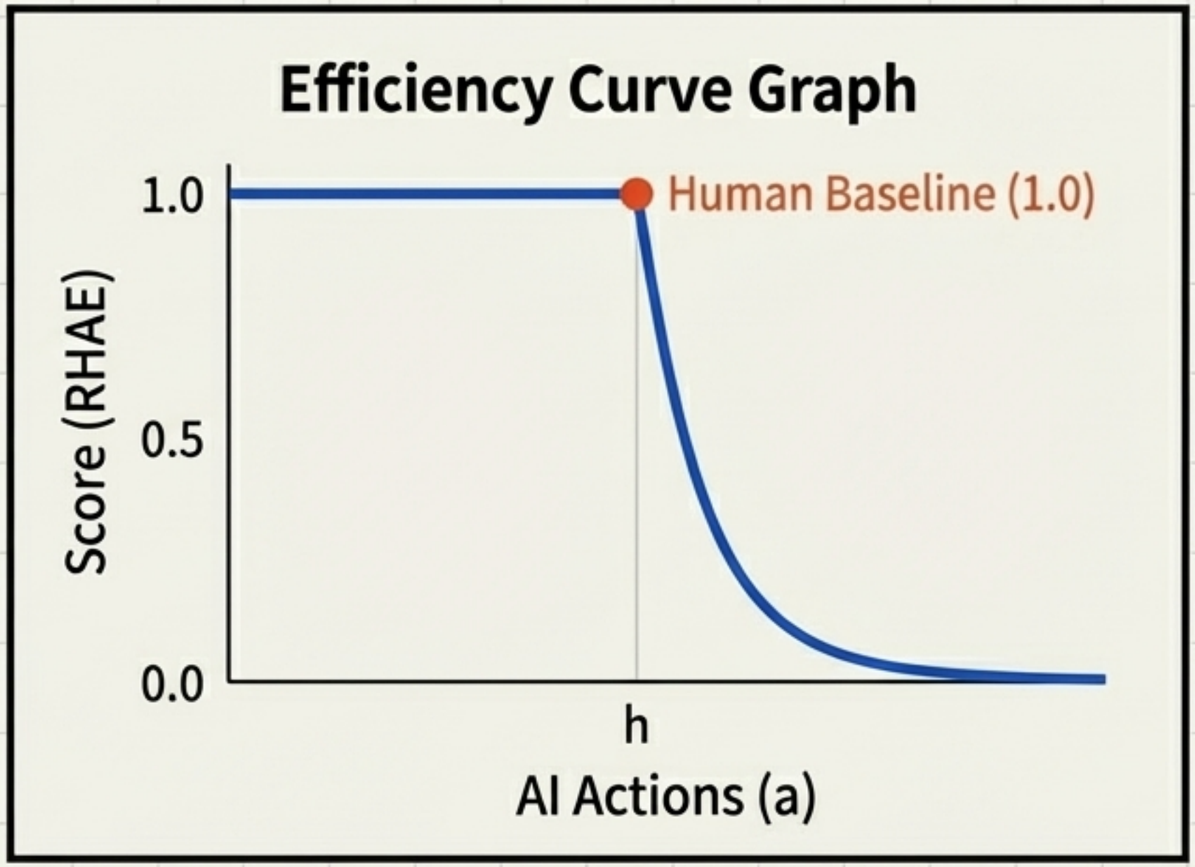


行動空間 (Action Space)
非常に小さな行動集合（基本行動+Undo、座標指定など）。内部計算は含まない。

究極の制約
システムプロンプト「goal is to win」のみ。勝利条件自体をエージェント自ら探索して解明しなければならない。

スコアリング指標：RHAEの仕組み

$$RHAE = \min\left(1, \left(\frac{h}{a}\right)^2\right)$$



h	**(Human Baseline):** 人間10人の試行のうち「2番目に少ない行動数」
a	**(Agent Actions):** AIエージェントが完了までに要した行動数

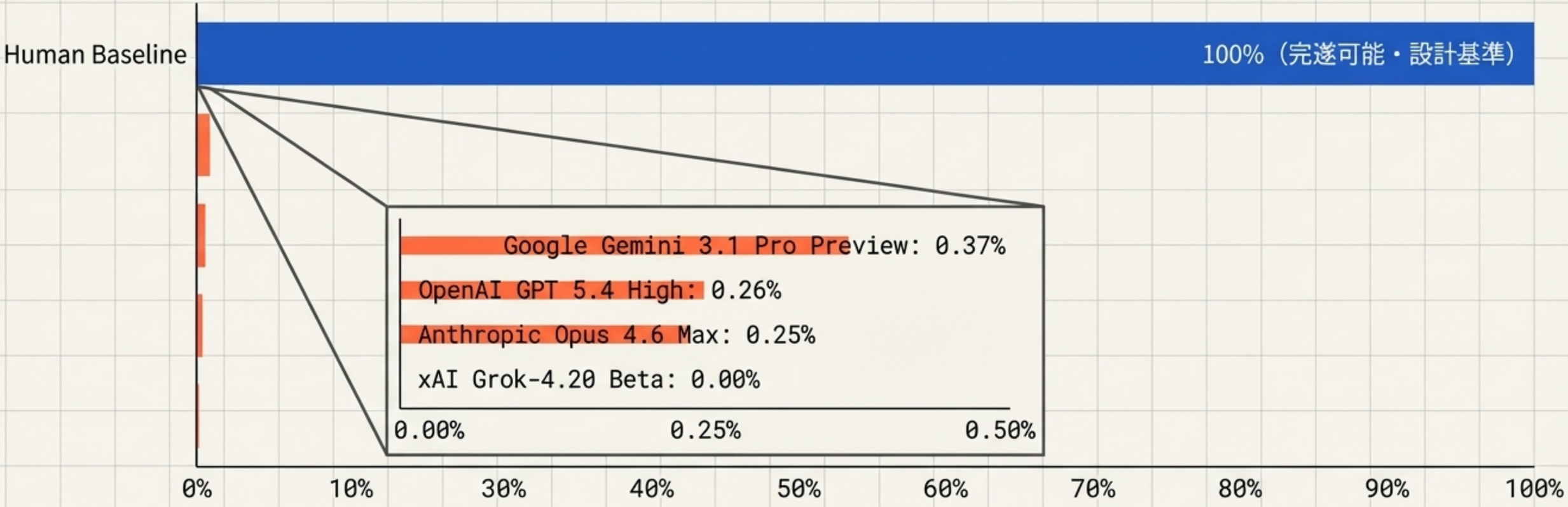
クリッピング:
少ない手数で解いても上限は「1」。スコア暴騰を防ぐ。

ペナルティ:
2乗関数により、無駄な探索（総当たり）を激しく減点。

集計:
レベル重み（1〜5）で加重平均し、全環境の平均で算出。

99%の壁: 現行フロンティアAIの限界

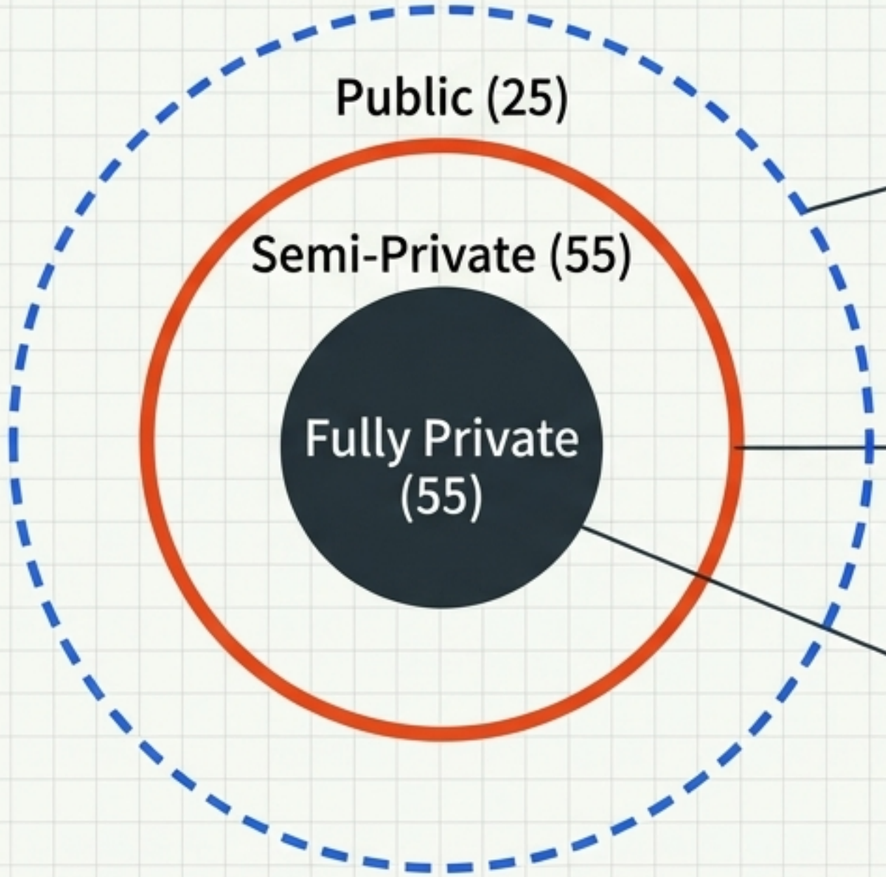
Semi-Private リーダーボード結果 (2026年3月公開時点・ハーネス無し)



大手モデル間の比較可能性は、同一の「半非公開セット」上で評価されることで担保されている。
現行の純粋な推論能力では、未知環境のルール推定において全滅状態にある。

構造的防御: データ分割による過学習防止

全135環境による厳密な分離



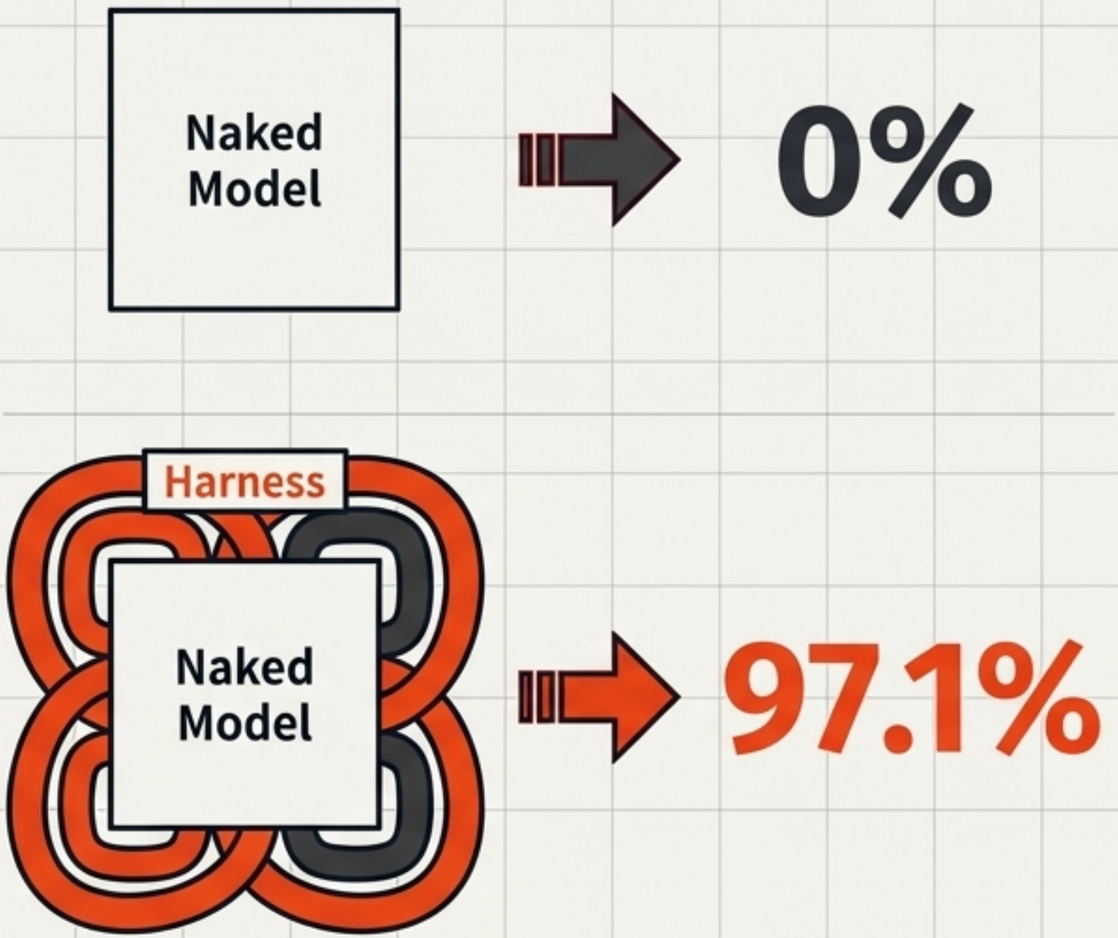
Tier 1: Public Demo (25環境)
- 用途: ローカル開発、テスト、SDKでの検証用。
- 防御: 公式リーダーボードには反映されない。

Tier 2: Semi-Private (55環境)
- 用途: 公式リーダーボード評価用。
- 防御: API経由でのみ評価され、モデル提供者側へのデータ汚染を完全に遮断。

Tier 3: Fully Private (55環境)
- 用途: 将来の検証や最終確認用として完全に秘匿。

独立した再現性を犠牲にしてでも、訓練データへの「暗記（データ汚染）」を構造的に排除する設計思想。

評価の脆弱性と「ハーネス」問題



タスク/ドメイン特化リスク

公開環境や合成環境に過学習（Overfitting）させることで、一般化能力を持たないままスコアを偽装できる危険性。

ハーネスによるスコアの人工的インフレ

- 問題: 外部の専用手続き・推論ループツール（ハーネス）でモデルを包むと、同一モデルでも極端な二峰性スコアが発現する。
- 対策: 公式リーダーボードは「ハーネス無し（Naked Model）」での生パフォーマンス評価を義務付け。

運用上の防御

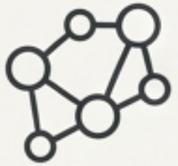
行動数が人間の5倍を超えた時点で打ち切る上限設定により、偶然の総当たり（Brute-force）を防止。

独立再現のための3ステップ技術フロー

Step 1: OFFLINE (ローカル環境整備)

行動: Python SDKの導入（pip/uv）。Public環境での大量実行。

利点: 高速実行（推奨約2,000 FPS）、回数無制限。探索アルゴリズムの基礎検証に最適。



Step 2: BASELINE (ベースライン比較)

行動: ランダム探索や状態グラ探索などを実装し、同一ゲーム上でRHAEの近似値を測定。

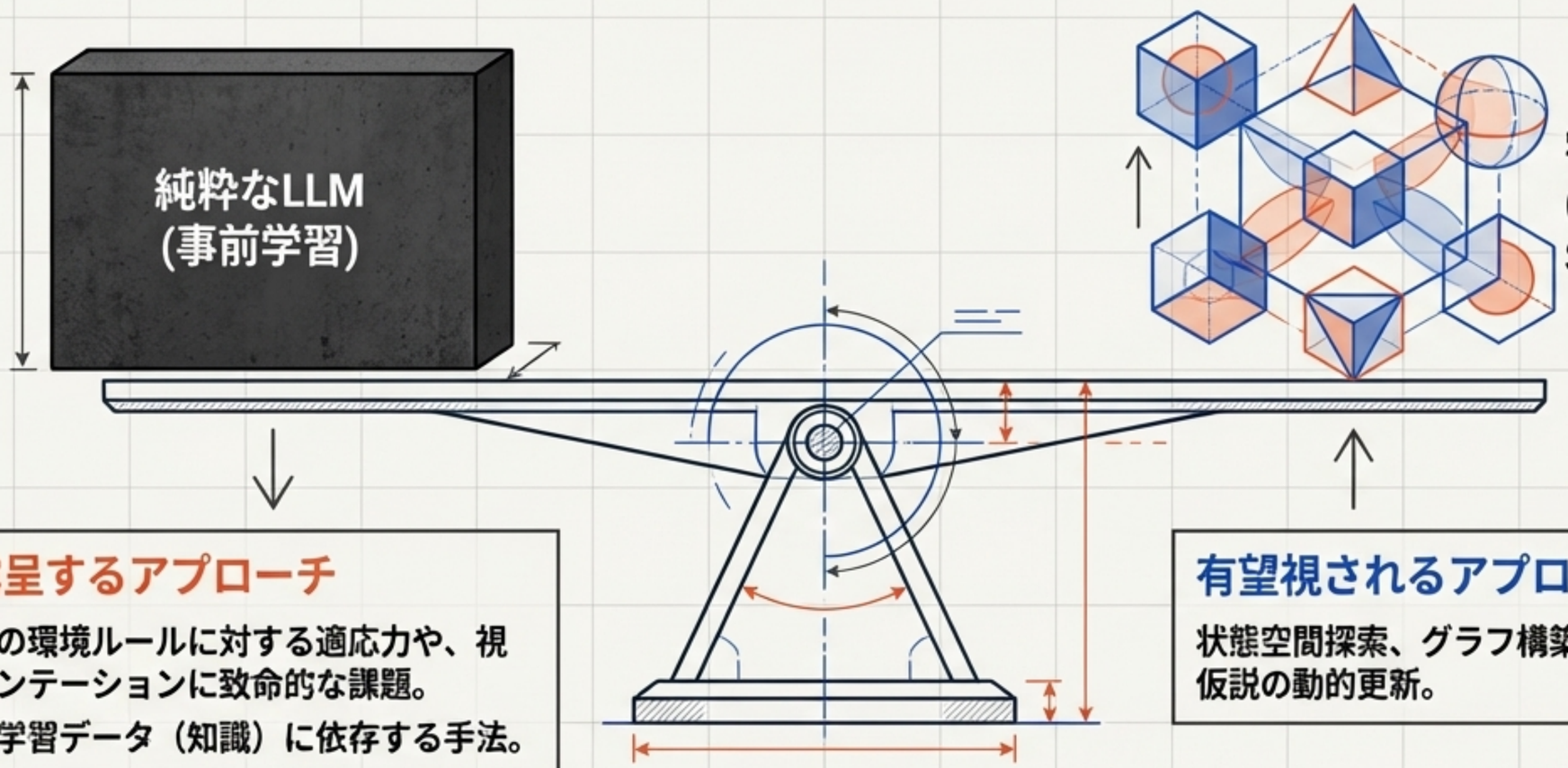
目的: LLM以外の「探索アーキテクチャ」の有効性を測る基準点を作成。

Step 3: ONLINE (API検証とログ取得)

行動: REST APIを使用した本番類似環境への接続。

制約: レート制限下での実行。スコアカードとリプレイ録画の生成。

学術的文脈:
「事前学習」から「探索」への回帰



限界を露呈するアプローチ

弱点：未知の環境ルールに対する適応力や、視覚的セグメンテーションに致命的な課題。

膨大な事前学習データ（知識）に依存する手法。

有望視されるアプローチ

状態空間探索、グラフ構築、メモリ管理、仮説の動的更新。

洞察: 短期的には、モデルのスケーリングよりも、未知環境における「記憶と仮説の管理」がAI進化の最大のボトルネックとなる。

ARC-AGI-3がもたらす戦略とガバナンスへの含意

研究開発 (Research)

- **焦点**：汎用的な「記憶圧縮」と「仮説管理」のアーキテクチャ構築。
- **注意**：単一のベンチマーク特化を避け、他ドメインへ移植可能な「真の学習効率」を追求する。

企業・実務 (Enterprise)

- **指標: RHAE**（無駄な試行を抑える効率）は、実業務におけるエージェントAIの「**運用コスト**」や「**安全性**」の評価に直結する。
- **アクション**：社内ツール制約を加えた独自ベンチとの組み合わせ。

政策とガバナンス (Policy)

- **指針:** ARC-AGI-3は**AGI達成のリトマス試験紙ではない**。複数評価軸での検証を必須とする。
- **国際規格**：**EU AI Act**等の監査可能性に合致するよう、リプレイログによる透明性を確保。

AI評価の未来：透明性、多角的な検証、および実用的な監査可能性の確立