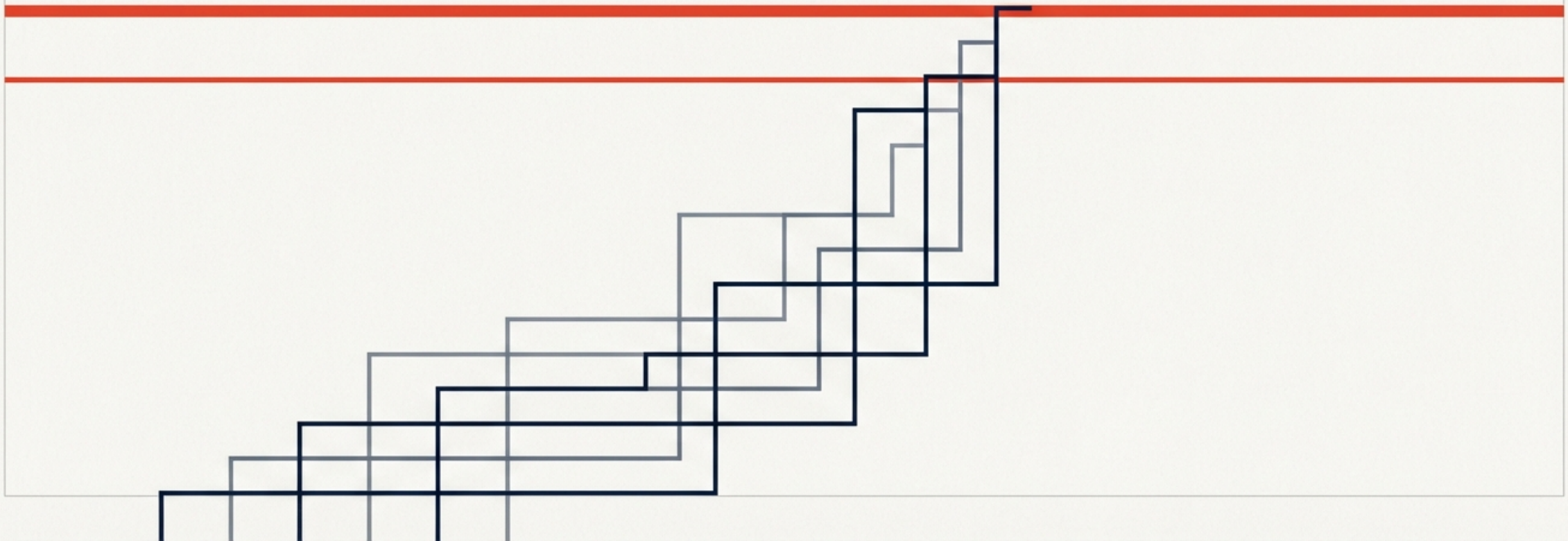
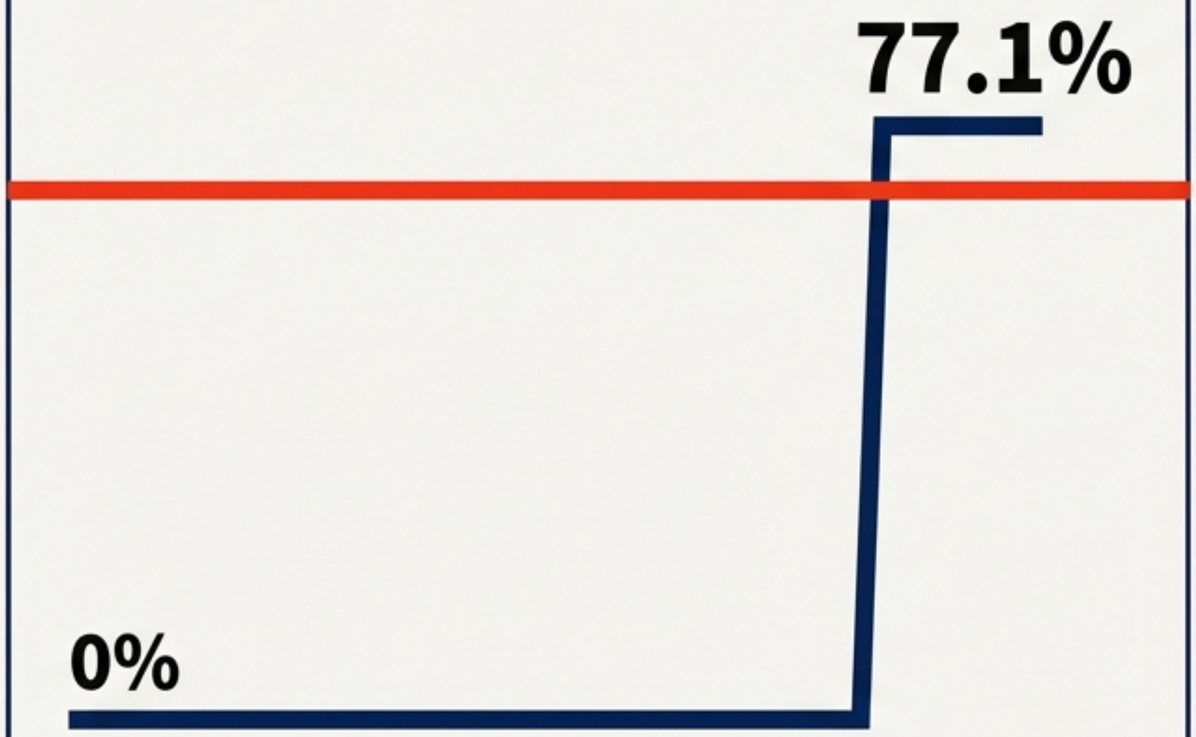
ARC-AGI-2の急伸とARC-AGI-3の到達予測

332日のスプリントから、次世代ベンチマークの長期戦へ

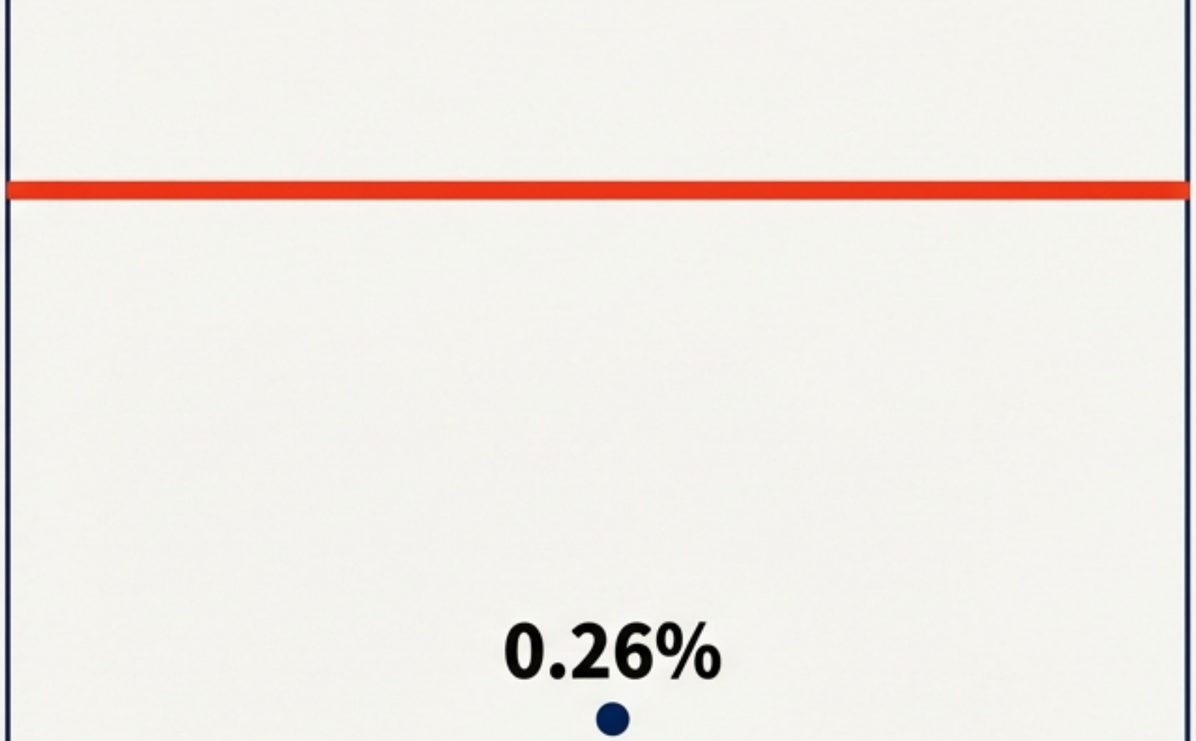


ARC-AGI-2: 332日の急伸



純粋なLLMの0%から、公式検証済みスコア77.1%までわずか約11か月で到達。

ARC-AGI-3: 270倍のキャズム

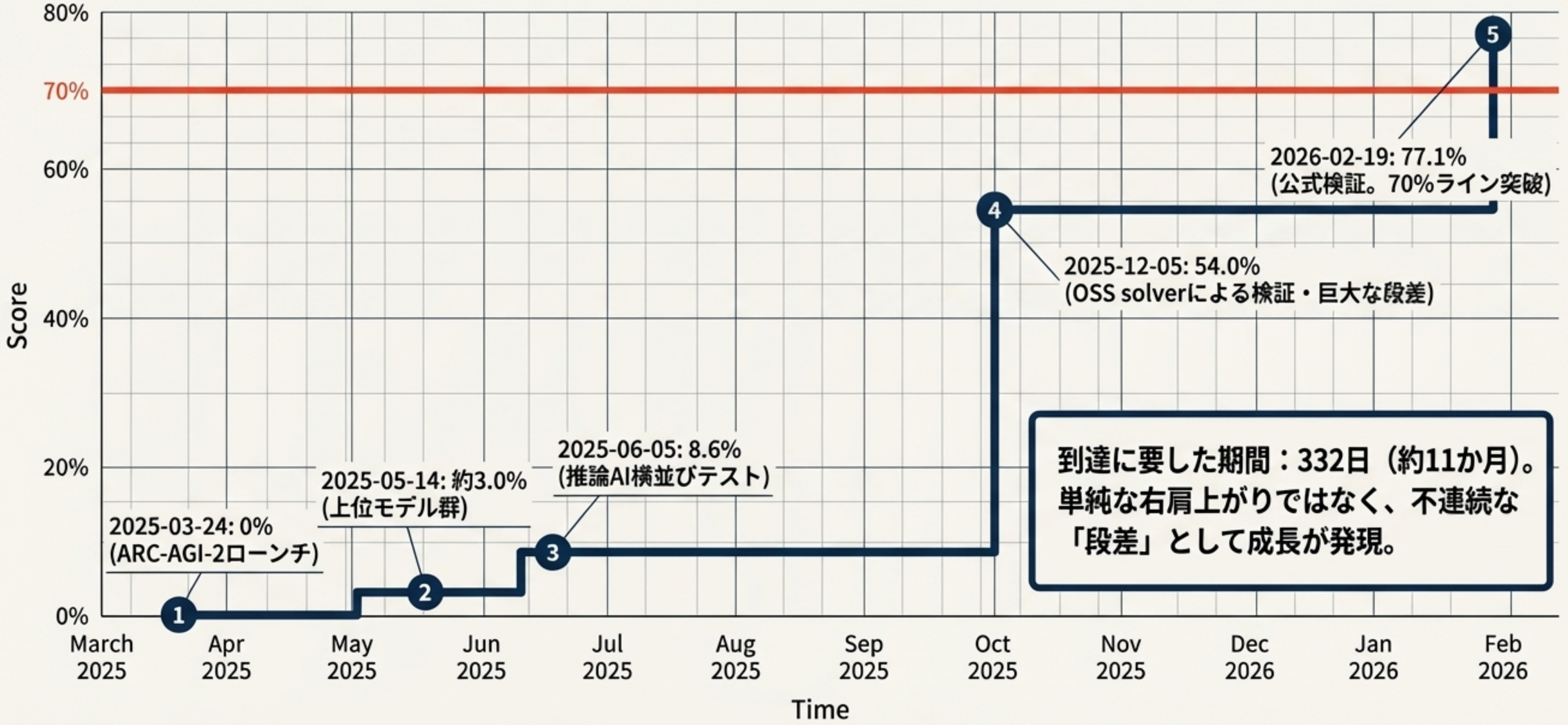


フロンティアAIの初期スコアはわずか0.26%。
中央値予測で到達まで3〜4年を要する長期戦へ。

曖昧な「70%」の定義を解体する

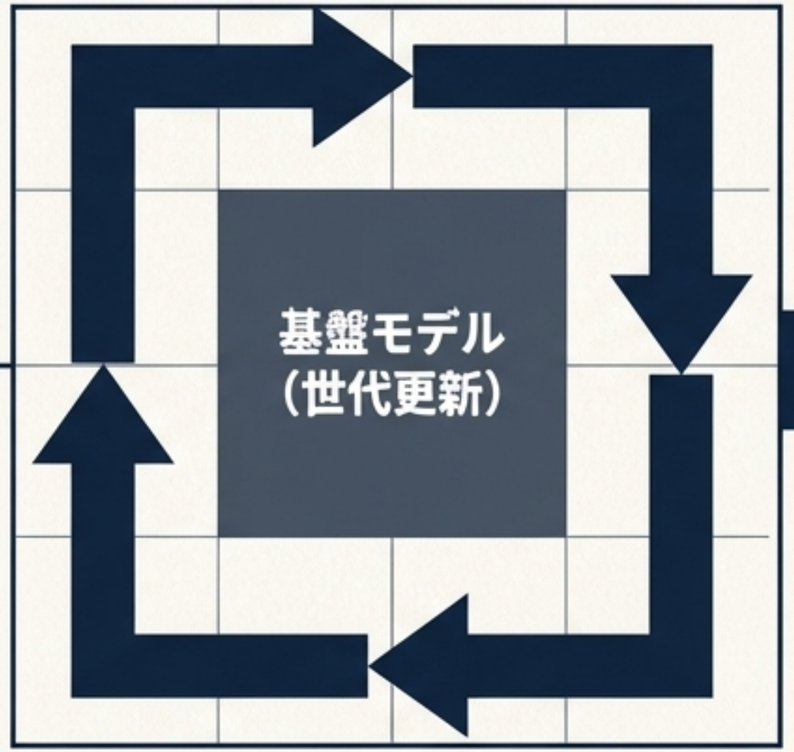
	定義A (Semi-Private / Verified) 本レポートの主軸 🎯	定義B (Fully-Private)	定義C (Public Eval)
特徴	公式・比較可能。商用APIモデル等を低リークリスクで遠隔評価。	コンペ最終順位決定用。リークリスク最小。	公開評価セット。反復検証による過適合リスクが高い。
到達状況	✅ 到達済み (2026-02-19にGoogle Gemini 3.1 Proが77.1%%を公表）。	❌ 未到達 (2025年大会の最終SOTAは約24%）。	⚠️ 慎重な扱い（優位の主張はあるが、Semi-Privateへの性能低下を想定）。

タイムライン：0%から77.1%への332日



ステップ関数の解剖：なぜ急伸したのか？

推論ループ / ハーネス
（探索→検証→修正）



+45.4pt の飛躍

要因1: ハーネスの導入

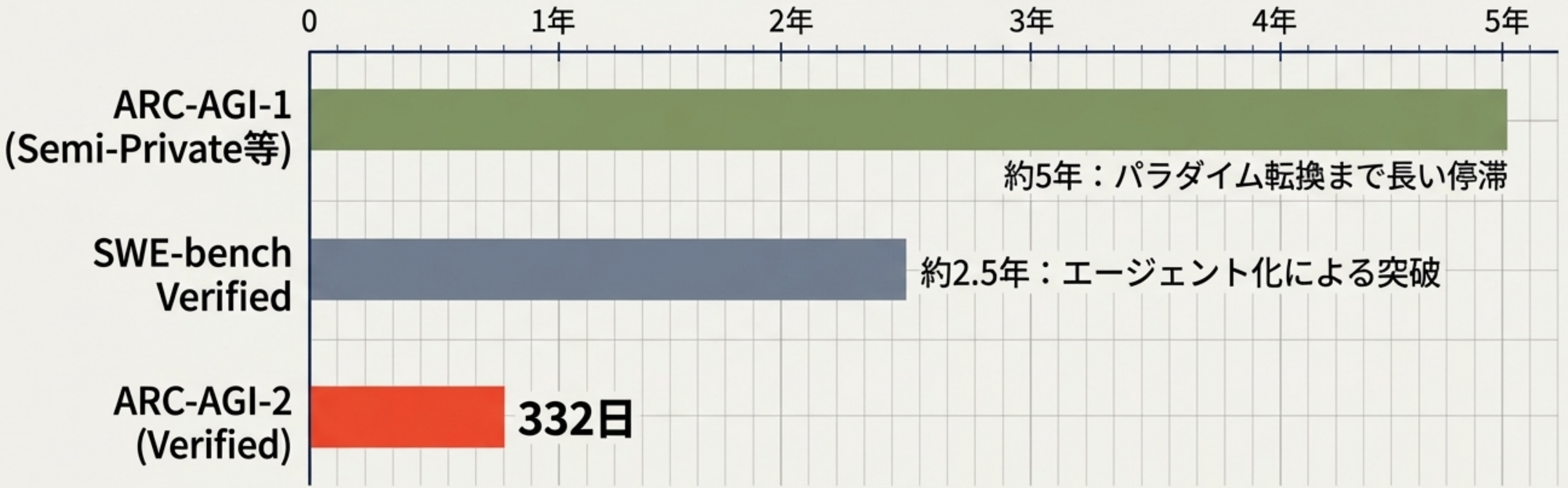
単純な出力ではなく「探索→検証→修正」の反復ループの適用。

要因2: モデルの世代更新

推論能力に特化した次世代基盤モデルの投入。

結論: 約6か月での+45.4ptは純粋な学習スケーリングではなく「評価手法のパラダイムシフト」によるもの。直線的外挿は成立しない。

類似ベンチマークの進捗曲線との比較



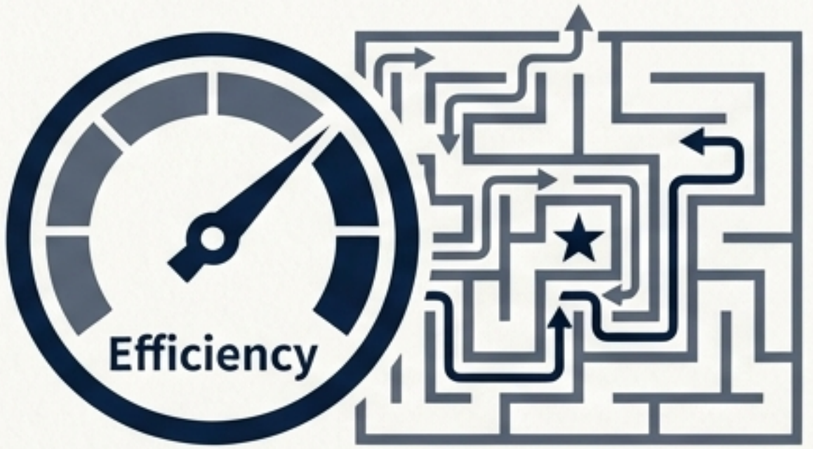
ARC-AGI-2の332日は「**速い部類**」に属するが、過去の「**長期停滞からの急伸**」という前例を考慮すると、次世代**ベンチマークに対する急伸前提の短期外挿は極めて危険**である。

ARC-AGI-3：完全に異なるパラダイムの幕開け

ARC-AGI-2

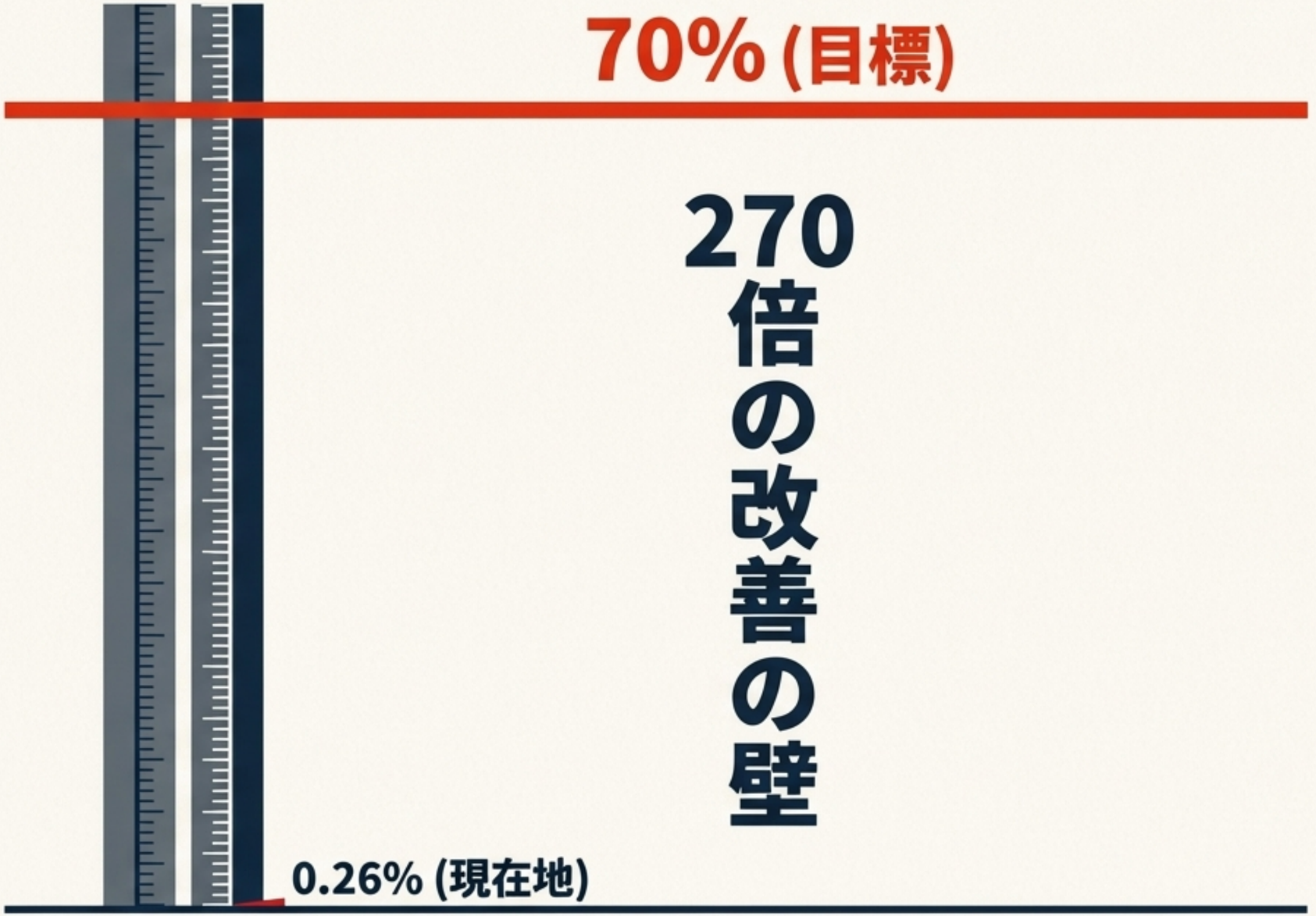
- **構造:** 静的（Static）
- **要件:** 初見テスト入力に対して2回の試行（pass@2相当）。
- **評価軸:** 結果の「正確性（Accuracy）」

ARC-AGI-3



- **構造:** インタラクティブ環境（ゲーム型）
- **要件:** ルールも目標も明示されない中での探索、勝利条件の発見。
- **評価軸:** 人間ベースラインに対する「行動効率（Action Efficiency）」の正規化スコア

0.26%からの出発：約270倍のキャズム



現在地

公開時点（2026-03-25）のフロンティアAIのスコアは0.26%前後。
※人間を100%とした場合

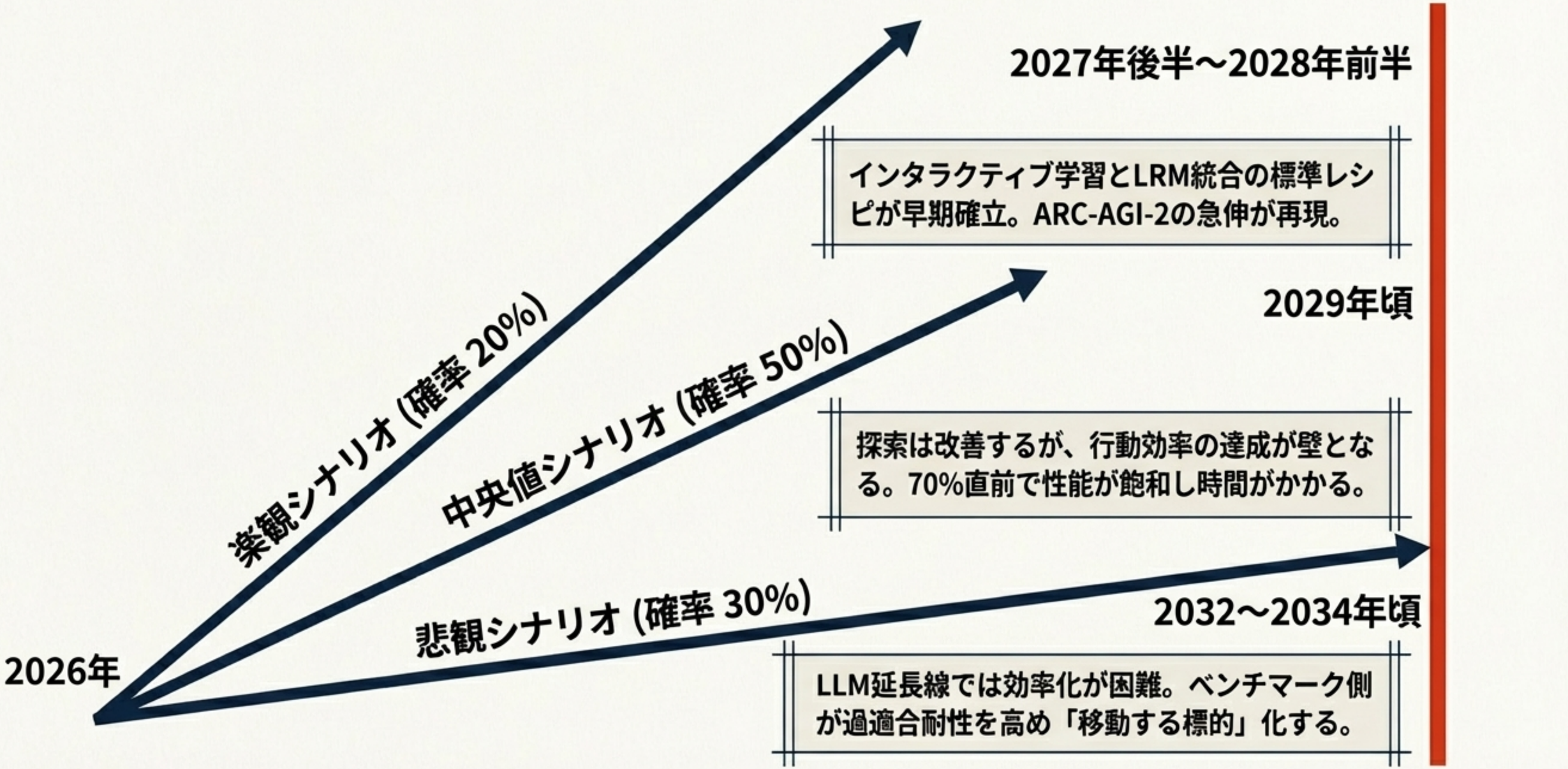
課題の性質

単純に「正解に辿り着く」だけでは不十分。無駄手を減らし、人間並みの行動効率で環境を探索しなければならない。

結論

ARC-AGI-2で機能した「外部ループ＋基盤モデル」の構図をそのまま適用しても、この壁は越えられない可能性が高い。

ARC-AGI-3 70%到達のシナリオ予測



ETAを左右する4つの感度要因

Dial 1: データ（学習環境）の供給

多様なインタラクティブ環境で「探索→学習→転移」を鍛えるための高品質な合成データ・シミュレータが供給されるか。

Dial 2: 計算資源とコスト構造

「効率」が評価指標に含まれるため、推論時計算量（Test-time compute）の単純な買い増しがスコアに直結しにくい構造。

Dial 3: アーキテクチャの変化

世界モデル、階層型方策、長期メモリなど、現行LLM推論の外部にある中核要素の成熟度合い。

Dial 4: 評価の変動（ゴールポスト）

更新頻度、反復提出の扱い、コミュニティ発のハーネスの公式認定など、ベンチマーク側のルール調整。

結論と構造的な不確実性

01

「70%」の定義と観測の限界

ARC-AGI-2の332日は「Googleによる検証済み公表日」を基準としており、未公表で早く到達していた可能性を含む。 Public Evalでの過適合議論は依然として残る。

02

スケール則から「効率」のパラダイムへ

ARC-AGI-2は「静的推論ループ」の追加で劇的な段差を形成した。しかし、ARC-AGI-3は人間ベースラインの「行動効率」を要求するゲームへ変化した。

03

最終見解

初期値0.26%からの270倍の改善を要する。過去の急伸曲線を外挿することはできず、到達目標は「計算量の暴力」から「エージェント的探索の効率化」へと完全に移行した。