

ARC-AGI-4：自律的な「オープンエンド・イノベーション」への進化

ARC-AGI-3の限界とARC-4への進化

「閉じたパズル」から
「終わりのない発見」へ
固定された問題の正答率ではなく、
時間とともに新しい成果が増え続ける
能力を測る。



GPT-6 Astraによる
「ARC-3飽和」の教訓
ハーネス設定次第でスコアが急変し、エージェント
能力と環境管理の分離が
困難になった。

Meta-Skill

科学的イノベーションを
「メタスキル」と定義
物理・化学等の特定知識ではなく、未知の課題から法則を見出す
一般学習効率を重視する。

ARC-AGI-4の評価指標と安全性の統合



3つの核となる指標：
妥当性・新規性・有用性
正解が既知でない成果に対し、科学的正しさと新しさ、
価値を多角的に検証する。

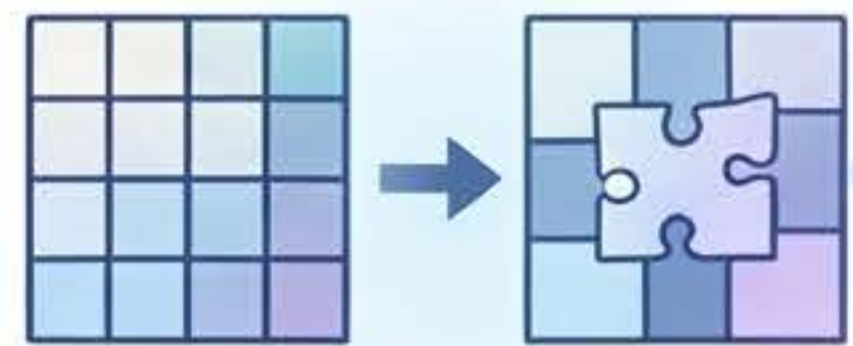


「検証済み発見率」に
よる長期評価
1時間から1週間という長期
スパンで、AIの改善曲線と
資源効率を人間と比較する。



能力評価と「安全監査」
の同時実行
科学発見能力はサイバー攻撃
や生物兵器等のリスクと
直結するため、安全門を必須とする。

ARC-AGIシリーズの世代による焦点の変化



ARC-AGI-1/2

評価対象：静的な抽象推論パズル
形式：2Dグリッド変換
状態：公開済み・飽和中



ARC-AGI-3

評価対象：相互作用するエージェント知能
形式：未知の対話型環境(ゲーム)
状態：公開済み・Astraが突破



ARC-AGI-4

評価対象：自律的なオープンエンド・イノベーション
形式：未指定(科学発見、発明、コード)
状態：2026年9月に構想発表