

AGI評価の革命：ARC-AGI-3の限界とARC-AGI-4への進化

GPT-6 Astraの検証結果を通じて、AI評価軸が「パズル解決」から「自律的イノベーション」へと転換する流れを伝える。

ARC-AGI-3の衝撃と「ハーネス論争」

環境依存の智能評価



評価環境による37.2ポイントのスコア差が物議。
標準では62.7%だが、特化環境では99.9%に到達。

記号的世界モデルによる高速化



人間のベースラインを上回る効率

環境からDSLを自律構築し、人間より少ない手数で課題を解決。

相対的人間行動効率 (RHAЕ)

手数に基づく厳格な採点。
クリアの成否だけでなく、
人間の探索手数との比較で
知能を定量化。



ARC-AGI-4：未知の領域を測る新基準

閉鎖系パズルから「開放系」への転換



自律的な課題発見とイノベーション

あらかじめ定義された正解がない環境で、自ら問いを立てる能力を評価。

多元的シミュレーション環境での科学的発見



物理法則が変容する空間で、新たな物理モデルを定式化する力を測定。

ARCベンチマークの世代進化と評価対象の変遷

世代	主な評価対象	成功の定義 (メトリクス)
ARC-1/2	静的な2Dグリッド変換	面素の完全一致 (正答率)
ARC-3	能動的探索・多段計画	人間との行動効率比較 (RHAЕ)
ARC-4	自律的イノベーション	新規性・有用性・情報理論的驚き