未知領域の開拓能力を測るARC-AGI-4の設計思想とAGI評価構造の転換

Gemini Fable 5.1

2026年9月3日、OpenAIが発表した最新フラッグシップモデル「GPT-6 Astra」は、対話型・能動的知能ベンチマーク「ARC-AGI-3」において99.9%というスコアを記録し、同社は「AGI時代への突入」を公言した[1]。しかし、この記録はOpenAI独自の推論状態保持機構を用いた「Provider Adapter」環境で達成されたものであり、全モデル共通の中立な「Standard harness」における独立検証スコアは62.7%にとどまる[2]。ベンチマーク創設者のFrançois CholletおよびARC Prize Foundationは、「ルールと正解があらかじめ規定された閉鎖的パズル環境の攻略は、真のAGIの証明にはならない」と反論し、次期ベンチマーク「ARC-AGI-4」の策定を電撃的に発表した[2]。

ARC-AGI-4は、2027年第1四半期の公開が予定されている新世代の知能評価体系である[2]。その中核となる評価対象は「自律的で終わりなきイノベーション（autonomous open-ended innovation）と科学的発見」であり、従来のパズル型評価から抜本的に脱却している[2]。あらかじめ設計者が定義した正解（Ground Truth）を当てる閉鎖系探索から、未定義の物理・記号空間において自ら課題を発見し、仮説検証サイクルを回し、新たな概念体系を創造し続ける「開放系（open-ended）における創造的メタ知能」の検証へと、AGI判定の主戦場が完全に移行したことを意味する[2]。

本論では、GPT-6 AstraによるARC-AGI-3攻略の内部構造とハーネス論争を詳細に検証した上で、ARC-1からARC-4に至る評価軸の変遷、ARC-AGI-4が測定対象とする未定義課題のアーキテクチャ、そしてフロンティア研究所間の協調路線とオープンソース陣営との知能判定権を巡る対立構造を包括的に解明する[4]。

GPT-6 AstraによるARC-AGI-3攻略の技術解剖とハーネス論争

ARC-AGI-3におけるGPT-6 Astraの評価結果は、基盤モデル単体の推論力とそれを取り巻くエージェント環境（テストハーネス）の役割分担について、AI研究界に本質的な問いを突きつけた[2]。ARC Prize Foundationによる公式検証では、同一のGPT-6 Astraでありながら、テストハーネスの構造差異によってスコアが37.2ポイントも変動する結果が示された[3]。

評価条件	使用テストハーネス	推論リソース設定	達成スコア	総検証コスト	状態管理および文脈制御方式
公式共通基準	Standard harness	reasoning = max	62.71%	約26,098ドル	手番ごとに推論文脈をリセット。モデルが自発的に残したテキストメモ

					のみを次手へ引き継ぐ中立環境[3]。
プロバイダ特化	Provider Adapter harness	reasoning = high	99.95%	約18,817ドル	OpenAIネイティブAPIを介し、不可視の推論状態（opaque reasoning state）を保持。文脈長超過時は要約圧縮（compaction）を実行[3]。

Standard harnessは、各アクションの実行後に推論トレースを破棄し、モデルが自発的に外部出力したスクラッチパッド上のテキストのみを次ターンに引き渡す最小構成の中立インターフェースである[4]。この環境下では、手番が進むごとに過去の思考過程をテキストから再解釈し直す必要が生じ、手数の増加を招く[3]。これに対しProvider Adapter harnessは、直前のレスポンスIDを保持して不可視の推論状態（opaque reasoning state）をセッション間で永続化させ、トークン数が上限に近づいた際も単純な切り捨てを行わずに要約圧縮（compaction）を実施する[4]。これにより、過去の思考の蓄積を直接活用することが可能となり、Provider Adapterを用いたAstraは、最小限の推論設定（low）でも96.7%を記録し、Standard harnessにおける最大推論設定（max）の62.7%を大差で上回る結果となった[2]。

記号的世界モデルの自律構築メカニズム

ハーネスによる支援構造が存在する一方で、GPT-6 Astraが前世代モデル（GPT-5.6 Solの7.78%やClaude Opus 5の30.2%）に対して圧倒的な差をつけた核心的要因は、環境の観測から自律的に簡潔な記号体系（DSL）を立ち上げる「オンザフライの記号的世界モデリング（on-the-fly symbolic world modeling）」能力にある[1]。

Astraは、未説明の64×64ピクセル環境に投入された際、画素データを単なる視覚配列として受動的に処理し続けるのではなく、自ら環境の状態変数を定義し、操作系との対応関係を極小の代数的記号列へと圧縮していた[4]。例えば、多段機構パズル「s5i5」の実行ログでは、機構の局所回転角や長さを構造化した状態把握ログを生成し、色番号8と10の連動機構に対する制御シーケンスを短いコード形式で記述した上で、画面上の制御座標と操作コマンドを動的にバインドして計画を実行していた[4]。

このように、未踏環境の挙動から即座に内部世界モデルを構築し、それを記号表現へと抽象化して推論空間内でシミュレーションを行うプロセスが成立したことで、探索空間の爆発を防ぎ、人間と同等以上の効率で解法を導出することに成功した[4]。

行動効率指標RHAEと人間のベースライン突破

ARC-AGI-3の採点基準は、従来のタスク成否（pass@k）ではなく、人間の探索効率と比較した「相対的人間行動効率（RHAE: Relative Human Action Efficiency）」を採用している[3]。クリア達成時におけるレベルスコアの計算式は以下のように定義されている[3]。

$$\mathrm{RHAE} = \min\left(1.15, \left(\frac{h}{a}\right)^2\right)$$

ここで h は一般被験者約500名の検証データから得られた当該ゲームの中央値操作手数であり、a はAIエージェントが要した実際の操作手数である[3]。AIが人間の2倍の手数を費やしてクリアした場合、得点は4分の1（25%）に激減し、10倍の手数を要した場合は100分の1（1%）にまでペナルティが科される二次減衰構造となっている[3]。

Provider Adapter環境におけるGPT-6 Astraは、全レベルの96.0%において人間の中央値を下回る手数でゴールに到達し、全環境を通じた平均手数は人間のベースラインより51.7%少なかった[4]。探索、モデリング、目標設定、計画・実行というエージェント的推論プロセスが極めて高い精度で機能した結果、手数のペナルティを回避し、99.95%という飽和スコアを叩き出した[4]。

ARC-AGIベンチマークの世代進化史

François Cholletが提唱したARC（Abstraction and Reasoning Corpus）は、知識の記憶量を競う従来ベンチマークと一線を画し、「未知の概念に対する流動性知能（Fluid Intelligence）と学習のサンプル効率」を定量化することを目的に進化してきた[9]。ARC-1からARC-4への軌跡は、AIの推論フロンティアが直面してきた限界とパラダイムシフトの歴史を如実に反映している[9]。

ベンチマーク世代	公開時期	環境・タスク構造	評価対象となる認知機能	正解の定義と評価環境	採点メトリクス	直面した限界と次世代への移行要因
ARC-AGI-1	2019年	静的な2Dグリッド変換（数問の入出力例示ペア）[9]	少数例からの帰納推論、生得的知識（Core Knowledge）の適用[9]	単一の正解グリッド（グラウンドトゥルース固定）[9]	正答率（画素の完全一致、pass@k）[9]	テスト時計算量スケーリングや大規模プログラム合成による総当たり解法に脆弱[9]。

ARC-AGI-2	2024年	概念の合成度を高めた複雑な静的2Dグリッド変換[9]	合成的汎化能力、計算リソース制約下での推論適応[9]	単一の正解グリッド（固定）[9]	計算コスト上限（例: 1提出50ドル）付き正答率[9]	静的画像変換の枠組みにとどまり、動的な環境介入や試行錯誤ループを評価不能[14]。
ARC-AGI-3	2026年3月	未知の対話型2Dゲーム環境（64×64ピクセル、離散アクション）[9]	能動的探索、環境の力学同定、目標設定、多段計画（Agentic Intelligence）[4]	環境内部に埋め込まれた勝利状態（決定論的ルール）[2]	相対的人間行動効率（RHAE）[3]	パズルが決定論的閉鎖系であるため、外部ハーネスの文脈保持機能により探索が飽和[2]。
ARC-AGI-4	2027年Q1（予定）[2]	開放系シミュレーション、可変公理空間、科学的発見タスク[2]	自律的オープンエンド・イノベーション、長期仮説検証、自己主導的知識獲得[6]	単一の正解が存在しない未定義課題（問題自体の定式化）[2]	新規性、有用性、情報理論的サプライズ、改善の持続性[7]	客観的な正解がない状況下での厳密な自動採点およびベンチマーク漏洩防止の設計難度[6]。

ARC-1およびARC-2は静的な入出力関係の推論を問い、高度な探索手法やプログラム合成によってスコアが押し上げられた[9]。ARC-3は動的な試行錯誤ループを導入したが、ゲームのルールや終了条件は環境内に決定論的に記述されていた[2]。GPT-6 Astraによる飽和は、AIが「設計者が定めたルールを逆算する能力」において人間の水準に到達したことを示したものの、未解決領域を切り拓く真の創造性には至っていないという課題を突きつけ、ARC-AGI-4の策定へとつながった[2]。

ARC-AGI-4の設計思想と未定義領域を測定する新アーキテクチャ

ARC-AGI-4が標榜する「自律的で終わりなきイノベーション（autonomous open-ended innovation）」は、あらかじめ勝利条件や正解が与えられていない動的な開放系環境において、システムが自律的に探索を続け、価値ある新規概念や技術体系を創出し続ける能力を定義の核に据えている[6]。

閉鎖系パズルから開放系イノベーションへの転換

ARC-AGI-3までの評価系は、どれほど複雑であっても設計者が想定した「正解（Ground Truth）」が存在する閉鎖系ゲームであった[2]。AIは報酬シグナルやクリア判定を頼りに逆問題を解くだけで達成可能であったが、現実の科学的発見や産業イノベーションにはあらかじめ定められた正解が存在しない[2]。

ARC-AGI-4では、問題定義そのものをモデル自身に行わせるパラダイムを採用している[7]。外部から明確なゴールを与えられずとも、環境内の矛盾や不連続性を自発的に検知し、自律的に仮説を立てて実験を設計・実行し、得られた観測データから理論モデルを再構築する長期的な仮説検証ループの持続性が問われる[7]。さらに、単一の狭いタスクで得た経験を別の異質なドメインへメタ原理として抽象化・転移させる能力、そして既存データの統計的補間を超えた外挿的な解法を創出できるかが中核要件として組み込まれている[8]。

想定される検証アーキテクチャと新評価指標

ARC Prize Foundationは、客観的な正解が存在しない課題を厳密にスコアリングするため、新たな検証アーキテクチャを模索している[2]。

その一つが、物理法則や公理系が動的に変容する多元的シミュレーション環境（Poly-physical Virtual Worlds）である[8]。エージェントが特定の介入を行うことで力学法則そのものが相転移を起こす環境において、新たな物理モデルを自律的に定式化し、その法則に適合した構造物や機能的システムを設計・実証できるかを測定する[8]。また、形式検証系を用いた高度数学・アルゴリズム設計タスクの導入も検討されている[8]。提示された定理の証明のみならず、未解決の数学空間において「非自明かつ検証可能な新規予想」を自発的に立て、その証明体系を構築できるかが評価軸となる[8]。

客観的グラウンドトゥルースの不在という課題に対しては、情報理論的指標に基づく定量的採点モデルの導入が進められている[2]。生成された解法や理論モデルがどれだけ環境の記述長を圧縮できたかを示すコルモゴロフ複雑性の低減度、他モデルや人間にとっての予測困難性を測るサプライズ度、そしてその発見が後続の探索空間をどれだけ有効に拡張したかを示す分岐エントロピーを統合し、新規性（Novelty）と有用性（Utility）の双方を数理的に担保する設計が研究されている[8]。

コミュニティの論争と知能判定の主権争い

GPT-6 Astraの登場とARC-AGI-4の発表は、知能の本質を巡る概念的対立のみならず、AI産業全体の安全保障体制とガバナンスを揺るがす深刻な主権争いを誘発した[2]。

François CholletらによるAGI宣言への反論

OpenAIがAstraをもって「AGI時代への突入」を主張したのに対し、François CholletおよびARC Prize共同創設者のMike Knoopは明確に異議を唱えた[2]。Knoopは「誰も答えを知らない未知の問題を自力で切り拓く力は未だ実現しておらず、AstraをAGIと呼ぶ証拠はない」と指摘した[2]。

Cholletの知能観の根幹にあるのは、「知能とは特定の静的スキル（性能）そのものではなく、未知の環境において極少のサンプルから新たなスキルを獲得する変換効率（流動性知能）」であるという原則である[9]。Provider Adapterによる99.9%という数値は、外部のAPIラッパーが思考状態を保持・要約することで得られたシステム全体の成果であり、モデル自体の基礎的な適応力や重みの自己更

新能力が達成されたわけではない[2]。AI研究者コミュニティ内でも、ARC-3がモデル本来の知能ではなく、ハーネス設計の巧妙さを競うテストに変質してしまったのではないかという懸念が議論された[2]。

「Pace the Frontier」協調路線とオープンソース原則の対立

ARC-AGI-4の発表直後、業界構造を二分する重大な政治的対立が発生した[7]。2026年9月12日、AnthropicのCEOであるDario Amodeiはエッセイ「We Must Pace the Frontier」を発表し、再帰的自己改善（RSI）の加速が安全検証の閾値を超えつつあると警告した[10]。Amodeiは、業界全体でフロンティアモデルの能力開発速度を意図的に抑制（Pacing）し、第三者評価機関による事前審査を義務付ける協調規制を呼びかけた[10]。この呼びかけに対し、OpenAIのSam AltmanやGoogle DeepMindのDemis Hassabisら主要ラボの首脳陣が次々と賛同の意を表明した[10]。

これに対し、ARC Prize Foundationは公式声明を通じて即座に強い懸念と対抗姿勢を表明した[7]。ARC-AGI-4を完全なオープンソース基準として運用する方針を再確認した上で、以下のように痛烈な牽制を行っている[7]。

フロンティアAI技術に関する知見は、一握りの巨大組織によって寡占されるべきではなく、研究者やアカデミア全体に広く分散して共有されなければならない[7]。安全性の名の下に開発速度の抑制や業界主導の事前調整を行い、フロンティアAIへのアクセスを特定の大手企業群に集中させようとするいかなる協調体制も、人類全体のポジティブサムな未来を損なうものである[7]。

大手クローズド研究所が安全性を根拠に評価と開発の主導権を握ろうとする動きに対し、ARC Prizeは外部から独立したオープンな審判基準としての役割を堅持する姿勢を明確にした[7]。

実施スケジュール、資金規模、参画要件

ARC-AGI-4に向けた開発およびコンペティションの立ち上げは、すでに具体的な資金的裏付けとロードマップを伴って進行している[2]。

公開スケジュールと開発ロードマップ

公式発表および予測市場の動向によると、ARC-AGI-4の正式公開とコンペティションの開幕は2027年第1四半期（Q1）に設定されている[2]。2026年後半を通じて、開放系シミュレーション環境のストレステスト、非公開評価データセットの作成、および人間によるベースライン測定が実施される計画となっている[8]。

資金基盤と賞金規模

ARC Prize Foundationは、2024年の賞金総額100万ドル規模のコンペティション運営以来、オープンソースコミュニティや篤志家からの支援を拡大させてきた[23]。2026年6月には、ARC-AGI-4の開発および検証インフラの構築を直接支援するため、**1,001,337**ドルの追加資金提供が受領されたことが記録されている[30]。この資金的基盤に基づき、ARC-AGI-4のコンペティション賞金プールは過去の大会規模を上回る規模で設定される見込みである[23]。

計算リソース制約（**Compute Cap**）の設計方針

ARC Prizeが重視してきた公平性の原則は、ARC-AGI-4の参加要件にも厳格に適用される[9]。

ARC-AGI-2では1回の提出につき計算コスト上限が50ドルに制限されていたが、GPT-6 AstraによるARC-AGI-3の検証ではSemi-Privateセット完走に1万8000ドルから2万6000ドル規模のAPIコストが投じられていた[4]。

ARC-AGI-4では、巨大テック企業による計算資源の力任せの投入を排除するため、推論時計算量（FLOPsおよびAPIコスト換算）に上限を設ける「リソース制約部門」と、自由な計算資源を用いた「フロ

ンティア部門」の二重構造が検討されている[9]。これにより、パラメータ規模を抑えつつ高いサンプル効率を達成する先鋭的なアーキテクチャ（TRMやCompressARCの系列など）を公平に評価し、真のアルゴリズム的革新を可視化することを目指している[9]。
ARC-AGI-4が提示する本質的な転換点は、AI研究における「事前に定義されたゴール」の終焉である。GPT-6 AstraがARC-AGI-3を99.9%で飽和させたことは、人間がルールと評価関数を記述できる閉じた問題空間が、適切なコンテキスト管理と推論の保持によって機械的に解き尽くせる領域に到達したことを証明した[2]。知能の評価基準が「与えられた迷路を最短で脱出する効率」から「正解の存在しないカオスの中で自ら意味のある問いを発見し、新たな理論を創出できるか」へと移行した現在、ARC-AGI-4は真の汎用知能（AGI）の到達を峻別する最後の砦として、人間知性と機械知能の間に横たわる決定的な創造的断絶を検証することになる[2]。

引用文献

1. AGI能力テストで99.9%、ChatGPTの新モデル「GPT-6 Astra」登場, https://pc.watch.impress.co.jp/docs/news/2138206.html
2. 「AstraはまだAGIではない」評価団体が反論 - AI新聞, https://exawizards.com/column/ai-trend/news09-07-2026/
3. GPT-6 Astraは99.9%か62.7%か | ARC-AGI-3を統計で読む, https://wakara.co.jp/mathlog/20260905-3
4. OpenAI's GPT-6 Astra on ARC-AGI-3 - ARC Prize, https://arcprize.org/blog/astra
5. GPT-6 Astra released Has OpenAI reached the age of AGI? What, https://xpert.digital/en/gpt-6-astra/
6. "ARC-AGI-4 will be a benchmark for autonomous open-ended, https://www.reddit.com/r/accelerate/comments/1weltgt/arcagi4_will_be_a_benchmark_for_autonomous/
7. r/singularity - Reddit, https://www.reddit.com/r/singularity/rising/
8. ARC-AGI-4 Will Target Autonomous Invention: “the Meta-skill That, https://www.reddit.com/r/singularity/comments/1wenztq/arcagi4_will_target_autonomous_invention_the/
9. The ARC of Progress towards AGI: A Living Survey of Abstraction, https://arxiv.org/html/2603.13372v1
10. Dario Amodei calls to pace the frontier, and Altman and Hassabis, https://aisocratic.org/news/dario-amodei-calls-to-pace-the-frontier-and-altman-and-hassabis-sign-on
11. GPT-6 Astra が ARC-AGI-3 で 62.7% にも 99.9% にもなる ... - Qiita, https://qiita.com/sukimaengineer/items/d74337e14a5c5815fa9c
12. OpenAI「GPT-6 Astra」の評判・評価に関する深層調査報告書, https://yorozuipsc.com/uploads/1/3/2/5/132566344/a01ef068016e9dc8a5a8.pdf
13. ARG-AGI-3が99.9パーセント。特化ハーネスとは何か調べた - Zenn, https://zenn.dev/acntechjp/articles/a302d301eb9cec
14. GPT-6「Astra」がARC-AGI-3で99.9%——AIはついに「未知の世界, https://note.com/meta_meta7/n/n21298f917747
15. ARC-AGI-3登場に関する検証・技術分析レポート, https://yorozuipsc.com/uploads/1/3/2/5/132566344/06d91c14cc7d6f157cc8.pdf

16. ARC-AGI-3入門 — フロンティアAI全モデルが1%未満の ... - Qiita, https://qiita.com/kai_kou/items/3c7eeaac89e85d9dece2
17. ARC-AGI-3: A New Challenge for Frontier Agentic Intelligence - arXiv, https://arxiv.org/html/2603.24621v2
18. OpenAI's GPT-6 Astra on ARC-AGI-3 Scored 99.95% with Provider, https://news.ycombinator.com/item?id=49591755
19. AGI Benchmark. If ARC-AGI-3 is solved, do we have AGI? - Habr, https://habr.com/en/articles/1081200/
20. Chapter 19 - Deep Learning With Python, https://deeplearningwithpython.io/chapters/chapter19_future_of_ai
21. Below the Ice — The Agent That Aced the Test Nobody Could Pass, https://penguinalley.com/en/blog/2026-08-21-below-the-ice-nvidia-avo-arc-agi
22. Breaking: ARC-AGI-4 will be a benchmark for autonomous ... - Reddit, https://www.reddit.com/r/accelerate/comments/1weo9t9/breaking_arcagi4_will_be_a_benchmark_for/
23. ARC Prize: Posts, https://arcprize.kit.com/
24. ARC-AGI-3を完全攻略したAIは、汎用AIと呼べるか - note, https://note.com/pharma_i_cist/n/n281f2b98453a
25. ARC Prize Announces Next AI Benchmark Focused on Open, https://www.kucoin.com/news/flash/arc-prize-announces-next-ai-benchmark-focused-on-open-innovation
26. 「arc-agi-4」のYahoo!リアルタイム検索 - X（旧Twitter）を, https://search.yahoo.co.jp/realtime/search?md=t&ei=UTF-8&rkf=1&ifr=tl_unit&p=arc-agi-4
27. GPT-6とは？Astraの性能・料金・GPT-5.6との違い【2026年9月最新】, https://japan-ai.co.jp/media/10406/
28. The prevalent problem of misleading benchmark reporting (re: Astra), https://www.reddit.com/r/singularity/comments/1w6knep/the_prevalent_problem_of_misleading_benchmark/
29. [참고] ARC, 차세대 벤치마크 'ARC-AGI-4' 공개... 오픈소스 기조 유지, https://www.promppy.com/item/1627490
30. The New Forces After OpenAI Have Begun to Move - note, https://note.com/kagawatomo/n/nfa97baccac78?hl=en
31. [François Chollet] ARC-4 est en préparation, et sortira début ... - Reddit, https://www.reddit.com/r/singularity/comments/1r3a6l7/fran%C3%A7ois_chollet_arc4_is_in_the_works_to_be/?tl=fr