2026年3月ARC-AGI-3の登場とAI評価のパラダイムシフト：対話型推論ベンチマークが暴くフロンティアモデルの限界とAGIへの道程

Gemini 3.1 pro

汎用人工知能（AGI）評価の新たな転換点

2026年3月25日、人工知能の真の推論能力と適応性を測定するための画期的なベンチマーク「ARC-AGI-3」が、ARC Prize Foundationによって正式にリリースされた[1]。大規模言語モデル（LLM）の急速な発展に伴い、AI業界全体が「AGI（汎用人工知能）の入り口に立っている」という熱狂に包まれる中、この新しい評価基準は、現在のフロンティアAIモデルが抱える根本的な限界を冷酷なまでに浮き彫りにした[2]。

ARC-AGI-3は、従来の静的なデータセットに基づく評価プロトコルから完全に脱却し、AIエージェントが未知の動的な環境にインタラクティブに関与し、自律的にルールを学習する能力を測定する、世界初の「対話型推論ベンチマーク（Interactive Reasoning Benchmark）」である[1]。このベンチマークの開発は、GoogleのAI研究者であるFrançois Cholletが2019年の論文「On the Measure of Intelligence」で提唱した知能の定義、すなわち「知能とは、事前知識や経験を、不確実性を伴う価値あるタスクにおける新しいスキルへと変換する効率（Skill-acquisition efficiency）である」という哲学に深く根ざしている[4]。

150以上の独自環境と1,000以上のレベルで構成されるARC-AGI-3では、エージェントに対して自然言語による指示や、タスクの目的を示すプロンプトは一切与えられない[1]。エージェントは、まるで人間がルール説明のない新しいビデオゲームを初めてプレイするかのように、環境を探索し、隠された物理法則や論理的力学を発見し、自ら目標を設定して計画を実行しなければならない[5]。人間のテスト参加者が事前の訓練なしに100%の環境を解決できるのに対し、2026年3月時点の最先端モデルは1%のスコアすら獲得できないという結果は、現在のAIアーキテクチャが真の「流動性知能（Fluid Intelligence）」を獲得していないことを明確に示している[2]。本稿では、ARC-AGI-3の設計思想、革新的な評価指標、そしてこのベンチマークがAI研究の未来に与える深い影響について包括的に分析する。

ARC-AGIの進化論：静的パターン認識から動的エージェントへ

ARC-AGI-3の意義を理解するためには、前身となるARC-AGI-1およびARC-AGI-2の歴史的文脈と、それらがどのように「攻略」されてきたかを検証する必要がある。

2019年に公開された初期のARC-AGI-1は、最大30x30のグリッド上で数個の入出力例（デモンスト

レーション）から変換ルールを推論し、新しい入力に対する出力を予測する静的なタスクであった[5]。当初、AIシステムにとってこの課題は極めて困難であり、2020年のKaggleコンペティションでの最高スコアは20%にとどまっていた[1]。しかし、テスト時計算（Test-time compute）の増大、合成データ（Synthetic data）を用いた大規模な事前学習、そして推論を反復的に最適化するリファインメントループ（Refinement loop）の導入により、2026年までにClaude Opus 4.6などのモデルは93.0%というスコアに到達し、ARC-AGI-1は実質的に飽和状態（Solved）となった[7]。

続くARC-AGI-2（2025年リリース）では、複数ステップの構成的推論（Compositional reasoning）や文脈依存のルール適用が求められ、難易度が大幅に引き上げられた[1]。リリース直後は最高スコアが20%台に低迷したものの、わずか1年足らずの間に、GoogleのGemini 3 Deep Thinkが84.6%、OpenAIのGPT-5.4 Proが83.3%を記録し、再び急速なスコアの上昇が見られた[7]。

しかし、ARC Prize Foundationの分析によれば、これらのスコア上昇は「AIが汎用的な推論能力を獲得した」ことを意味するものではなかった。むしろ、ドメイン特化型のハーネス（Harness）や、数十万件に及ぶ合成類似データへの過剰適合（Knowledge-dependent overfitting）によって、本来は「流動性知能」を測るはずのテストが、事実上「結晶性知能（事前知識）」のテストへと変質してしまった結果であった[9]。

この「パターンの暗記と検索（Memorization-and-retrieval）」によるショートカットを完全に遮断し、AGIの真の指標（North Star）を再構築するために開発されたのがARC-AGI-3である[5]。静的なデータセットに対するバッチ処理ではなく、環境との継続的な相互作用を要求することで、真の意味での「適応的行動（Adaptive behavior）」を評価するパラダイムシフトがここに完了した[7]。

コア知識（Core Knowledge）とゲーム力学の設計思想

ARC-AGI-3の環境設計において最も重要な制約は、Elizabeth Spelkeらの発達心理学研究に基づく「コア知識の事前条件（Core Knowledge priors）」の厳格な適用である[5]。真の知能を測定するためには、AIと人間が持つ事前知識の差を排除し、両者を公平な基盤に置く必要がある[4]。

排除された文化的バイアスと事前知識

ARC-AGI-3のすべての環境（ゲーム）は、言語、数字、現実世界のオブジェクト（鍵や花など）の画像、あるいは「緑は進行、赤は停止」といった人間の文化的な慣習に依存する要素を完全に排除している[5]。その代わりに、以下の生得的な認知モジュールのみを前提として構築されている[5]。

- 物体性（**Objectness**）：物体が独立した実体として存在し、まとまりを保つこと。
- 幾何学と位相幾何学（**Basic geometry/topology**）：対称性、接続性、包含関係。
- 基本的な物理法則（**Basic physics**）：重力、運動量、反発などの直感的な力学。
- エージェント性（**Agentness**）：特定の存在が意図を持ち、目標に向かって行動するという認識。

これらのコア知識のみを用いて、過去のいかなるビデオゲームやベンチマークとも異なる、完全に新規（Novel）で手作りの150以上の環境が生成された[1]。

部分観測マルコフ決定過程（**POMDP**）と未知の報酬関数

技術的には、ARC-AGI-3の環境は16色の離散的なパレットを持つ64×64ピクセルの2Dグリッド空間として表現される[13]。AIエージェントは各ターンでフレーム（またはアニメーションの連続フレーム）を観測し、行動を決定する[5]。環境エンジンはPythonで構築され、秒間1,000フレーム以上の高速な実行環境を提供する[5]。

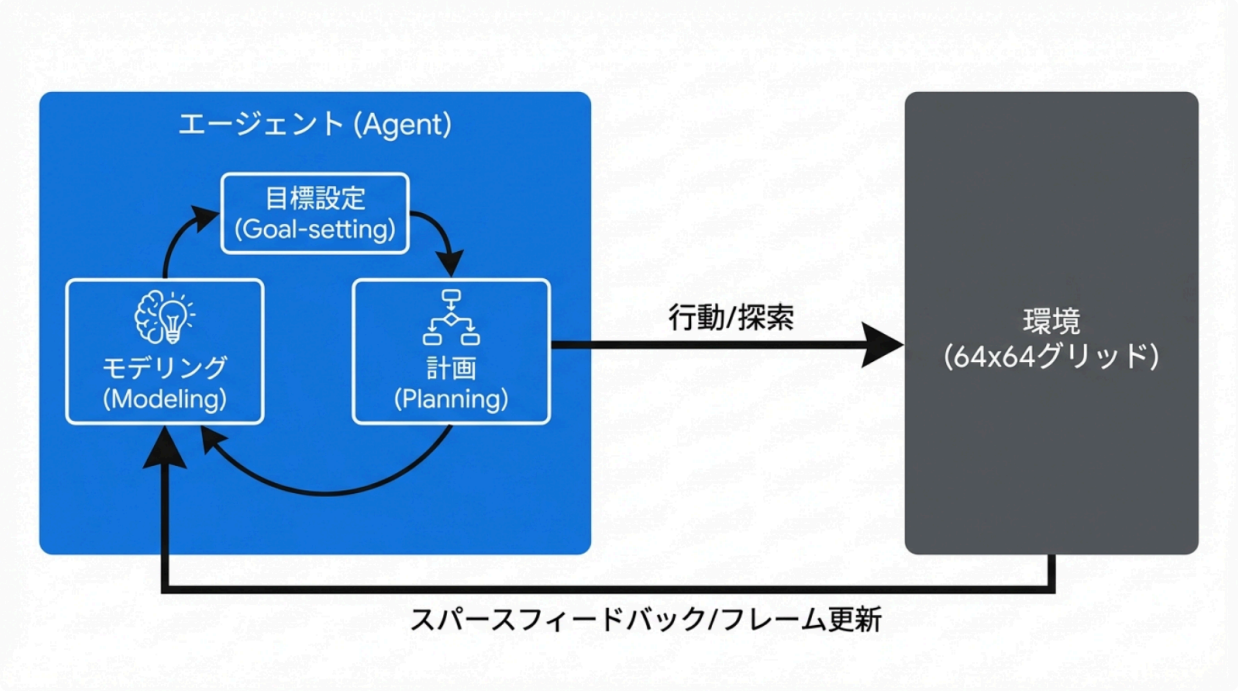
この課題は、強化学習における「部分観測マルコフ決定過程（POMDP）」として定式化できるが、決定的な違いが存在する。それは、エージェントが状態遷移の力学を学習するだけでなく、「報酬関数そのもの（目標やクリア条件）」をも自ら発見しなければならない点である[14]。唯一のフィードバックシグナルは、レベルが完了したか、あるいはステップ制限を超えて初期状態にリセットされたかという、極めて希薄（Sparse）なものに限られる[15]。

評価される**4**つのコア能力

このような過酷な環境下で、ARC-AGI-3はAIエージェントの以下の4つのコア能力を評価する[16]。

1. 探索（**Exploration**）：受動的に情報を与えられるのではなく、周囲と相互作用することで能動的に情報を獲得する能力。
2. モデリング（**Modeling**）：生の観測結果から、未来の状態や結果を予測できる汎用的な「世界モデル（World model）」を構築する能力。
3. 目標設定（**Goal-setting**）：明示的な指示なしに、環境のヒントと内発的動機づけから、望ましい未来の状態（勝利条件）を独自に見つけ出す能力。
4. 計画と実行（**Planning and Execution**）：現在の状態から目標に至るまでの行動経路を戦略的にマッピングし、予期せぬ結果に応じてリアルタイムで軌道修正する能力。

ARC-AGI-3における対話型推論ループ



エージェントは指示のない環境において、探索を通じてルールと目標を推測し、内部の適応的世界モデルを継続的に更新しながら行動を計画する。

ゲーム環境の多様性とアクション空間の非対称性

ARC-AGI-3の環境は、操作方法によって劇的に異なるアクション空間の複雑さを持つ。例えば、公開環境の一つであるls20は、見えない内部状態（Latent state）によって支配される環境内での記憶とナビゲーションをテストする[1]。このゲームは方向キーのみを使用するため、アクション空間のサイズは4（上下左右）である[13]。

対照的に、vc33（閾値と予算の管理）やft09（抽象的なパターンマッチング）といった環境では、クリックベースの操作が求められる[1]。ここでは、エージェントは64×64のグリッド上の任意のピクセルをクリックできるため、各ステップにおける可能なアクションの数は4,096に膨れ上がる[13]。このアクション空間の巨大な非対称性は、ランダムな探索行動や総当たり攻撃（Brute-force search）による攻略を事実上不可能にしている[13]。エージェントは、視覚情報から意味のある領域を抽出し、意図を持ってアクションを選択しなければならないのである。

スコアリングの革新：相対的人間行動効率（RHAE）の徹底解剖

ARC-AGI-3がAI業界に投げかけた最も議論を呼ぶ、しかし最も重要な革新が、その独自のスコアリ

ング手法である。従来のAIベンチマークは「タスクを解決できたか否か」という最終結果のみを評価してきた[6]。しかし、ARC-AGI-3は、知能を「情報をいかに効率的に戦略へと変換できるか」というプロセスとして測定する[6]。

知能の「コスト」を測定する

この目的のために導入されたのが、「相対的人間行動効率（RHAE: Relative Human Action Efficiency）」と呼ばれる指標である[17]。RHAEは、あるレベルをクリアするためにAIが消費した「行動数（アクション数）」をカウントし、それを「人間のベースライン行動数」と比較する[18]。外れ値の影響を排除するため、人間のベースラインは「初見プレイで2番目に少ない行動数でクリアした人間」のデータが採用される[18]。

AIエージェントのレベルごとのスコアは、人間のベースライン行動数をAIの行動数で割り、それを二乗する（二乗効率ペナルティ）ことで算出される[5]。なお、AIが人間よりも少ない行動数でクリアした場合でも、スコアは最大1.0（100%）に制限される[18]。また、モデル内部での推論ステップやツール呼び出しなど、環境を変化させない内部処理はアクションとしてカウントされない[5]。

RHAEの二乗ペナルティの厳しさは、具体的な数値を見れば明らかである。以下の表は、人間のベースライン行動数が10手であった場合の、AIの行動数と獲得スコア（RHAE）の減衰を示している。

人間のベースライン行動数	**AI**の消費行動数	効率比 **(Human / AI)**	**RHAE**スコア (効率比の二乗)
10	10	1.00	**100.0%**
10	20	0.50	**25.0%**
10	30	0.33	**11.1%**
10	50	0.20	**4.0%**
10	100	0.10	**1.0%**

データが示すように、行動数が人間の2倍（20手）になった時点でスコアは25%に急落し、10倍（100手）の手数を費やした場合はわずか1%の評価しか得られない。さらに、ある環境下で人間の5倍以上のアクションを要した場合、その時点でハードカットオフ（打ち切り）となり、スコアは0として扱われる厳しい制約も設けられている[18]。

なぜ二乗ペナルティが必要なのか

この厳格なペナルティ構造に対しては、一部の批判者から「意図的に低いスコアを捏造するための不公平な設計である」との指摘も上がった[20]。中央値の人間がトップクラスの人間より1.5倍の手数をかけただけでスコアが26.7%に落ちるような計算式は、AIの能力を不当に低く見せているという主張である[19]。

しかし、François Chollet をはじめとする設計者たちは、このペナルティこそが「真の知能」を偽装する力任せの探索（Brute-force）を無効化するための核心であると反論している[20]。数万回に及ぶ無作為なクリックや、あらゆる可能性のしらみつぶし探索によって偶然正解にたどり着くことは、強大な計算資源（コンピュート）の証明であっても、環境のルールを理解し適応したこと（知能）の証明にはならない[2]。RHAEは、「知能とは限られた資源（手数）で最適解に到達する能力である」という哲学を数学的に実装したものである[4]。

フロンティアモデルの完全な崩壊：「パフォーマンスの崖」の実態

2026年3月のARC-AGI-3公開に合わせて発表されたリーダーボードは、AI業界が長年依存してきた「スケーリング則（Scaling Laws）」と、現在のアーキテクチャに対する過信を打ち砕くものとなった[2]。人間が一切の事前指示なしに100%のタスクを解決できる一方で、世界最高峰の計算資源を誇るフロンティアモデル群は、軒並み1%の壁すら超えられなかったのである[2]。

AIモデル / テスト参加者	**ARC-AGI-1** スコア	**ARC-AGI-2** スコア	**ARC-AGI-3** スコア **(RHAE)**
人間のパネル **(Human Panel)**	98.0%	100.0%	**100.0%**
Gemini 3.1 Pro (Preview)	98.0%	77.1%	**0.37%**

GPT-5.4 (High)	92.7%	67.5%	**0.26%**
Claude Opus 4.6 (Max)	93.0%	68.8%	**0.25%**
Grok-4.20 (Reasoning)	N/A	N/A	**0.00%**

※データソース：ARC-AGI公式リーダーボード（2026年3月）[2]

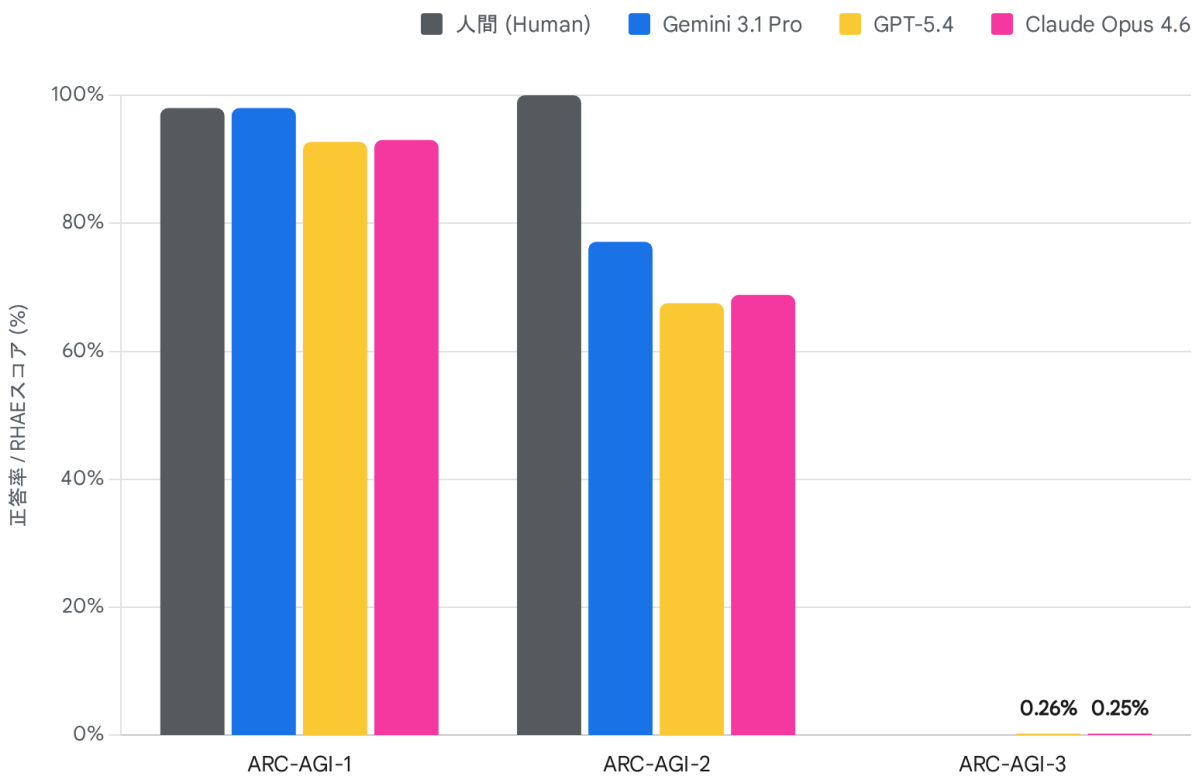
限界を露呈した「次トークン予測」パラダイム

ARC-AGI-1および2からARC-AGI-3へと移行した際に生じる、この極端な「パフォーマンスの崖（Performance Cliff）」は、現在のAIシステムが抱える根源的な脆弱性を証明している[7]。GPT-5.4やGemini 3.1 Proのようなモデルは、高度なCoT（Chain-of-Thought）推論やリファインメントループを用いることで、静的なパターン推論（ARC-AGI-2）においては人間の平均（約60%）を超える成績を収めることができる[7]。

しかし、彼らのアーキテクチャの根本は「次トークン予測（Next-token prediction）」に基づく自己回帰モデルである[23]。彼らはインターネット上の膨大なテキストやコードを記憶し、高次元のパターンを補間することはできても、動的な物理空間における「グラウンディング（Grounding）」を持たない[25]。

ARC-AGI-3のように、テキストベースの指示がなく、アクションの結果（フィードバック）に基づいて自身の内部状態（信念）をリアルタイムに更新し続けなければならない環境では、彼らの推論はたちまち破綻する[6]。Yann LeCun（Meta）やGary Marcusといった著名な研究者らが長年指摘してきた通り、LLMは世界に対する真の「世界モデル（World model）」や階層的な計画能力を欠いており、巨大な記憶の海に依存した「確率的なオウム（Stochastic parrots）」の域を出ていないことが、このインタラクティブなテストによって実証されたのである[25]。

フロンティアモデルの「パフォーマンスの崖」：ARC-AGI世代別スコア比較



ARC-AGI-1および2では静的なパターン推論により高いスコアを記録するモデルも、対話的な探索が求められるARC-AGI-3では1%未満へと劇的にスコアを落としている。

データソース: ARC Prize Leaderboard, ARC-AGI-3 Technical Report, arXiv (2603.13372v1)

ハーネス（足場）への依存と「真の汎化」の分離

ARC-AGI-3の低いスコアを解釈する上で、もう一つ重要な要素が「ハーネス（Harness）」や「足場（Scaffolding）」と呼ばれる外部ツールの排除である。

現在、多くのフロンティアモデルが複雑なコーディングタスク（SWE-benchなど）や高度な推論テストで優れた成績を収めているが、その背後にはTerminus-2や、Symbolica AIが開発したArcgenticaのような高度に専門化されたエージェント・フレームワークが存在している[28]。これらのハーネスは、LLMが思考を整理するためのワークフロー（例えば、Pythonコードの実行環境、エラー時の自動リトライ、マルチエージェントによる監視システムなど）を人間が手作業で構築したものである[30]。

ARC Prizeのテクニカルレポートによれば、特定の環境（例えばls20やft09）に特化したハーネス（Duke harnessなど）を手動で構築し、そこにClaude Opus 4.6を組み込んだ場合、特定のタスクにお

いて97.1%という驚異的なスコアを達成できることが確認されている[5]。しかし、同じモデルとハーネスの組み合わせを別の未知の環境に適用すると、スコアは再び0.0%に急落する[5]。

これは何を意味するのか。François Cholletが明確に指摘するように、精巧なハーネスを用いた際の高スコアは、AIシステム自身の汎用知能を測定しているのではなく、「そのシステムを特定のドメインに適応させるために注ぎ込まれた、人間の開発者の知能と巧妙さ」を測定しているに過ぎない[4]。

真のAGIは、人間による事前の手助け（ハードコーディングされた探索アルゴリズムやルールベースの誘導）なしに、未知の環境に適応できなければならない[20]。そのため、ARC-AGI-3の公式評価では、汎用的なAPIアクセスを通じたゼロショットに近い評価を重視し、モデル固有の推論能力そのものをテストするアプローチが採用されているのである[19]。

プレビューコンペティションが示す代替アーキテクチャの可能性

巨大な言語モデルが挫折を味わう中、ARC-AGI-3に向けたアプローチとして、全く異なるパラダイムから有望な兆しが見え始めている。2025年7月に開催された「ARC-AGI-3 Preview Agent Competition」では、LLMに依存しない小規模で特化型のアーキテクチャが上位を独占した[5]。

エージェント名	開発者 / 組織	アプローチの概要	**RHAE** スコア	クリアしたレベル数	消費アクション数
StochasticGoose	Tufa Labs	CNN + 変化予測に基づく強化学習	**12.58%**	18 / 20	255,964
Blind Squirrel	個人 (Will Dick)	グラフベースの有向状態空間マッピング	**6.71%**	13 / 20	109,108

※データソース：ARC-AGI-3 プレビューリーダーボード（2025年8月）[31]

StochasticGoose：視覚的優先度と能動的サンプリング

1位を獲得した「StochasticGoose」は、わずか4層の畳み込みニューラルネットワーク（CNN）を使用し、入力された64×64のフレームから「どのアクションが画面に変化を引き起こすか」を予測する強化学習アプローチを採用した[33]。このエージェントは、完全にランダムに動くのではなく、環境内の潜在

的なメカニズム（例えば、クリック可能なタイルや動くオブジェクト）を確率的にサンプリングし、変化をトリガーする行動に探索のバイアスをかけることで、効率的な情報収集を実現した[33]。

Blind Squirrelとグラフベース探索：システマティックな適応

2位の「Blind Squirrel」や、その後の関連研究（arXiv: 2512.24156）で示されたアプローチは、「グラフベースの探索（Graph-Based Exploration）」である[15]。この手法では、エージェントは観察した各フレーム（状態）をハッシュ化して記憶し、実行したアクションと結果をノードとエッジとして有向グラフ上にマッピングしていく[36]。

このシステマティックな状態空間の追跡により、エージェントは同じ状態をループする無駄な行動を回避し、未検証のアクションへ至る最短経路を自律的に計算できる[5]。LLMが数ターンの操作で文脈を失い幻覚（ハルシネーション）に陥るのに対し、グラフベースの探索は、希薄（Sparse）なフィードバック環境下において極めて強力なベースラインとして機能することが実証された[15]。

しかし、これらのアプローチも完璧ではない。StochasticGooseでさえ、人間が数手でクリアできる水量調整のレベル（vc33）において、序盤に約350回もの無駄なクリックを消費している[34]。Karl Fristonらが提唱する「能動的推論（Active Inference）」—エージェントが自らの内部世界モデルと現実とのズレ（不確実性）を最小化するように行動を選択するメカニズム—の実装こそが、これらの方策をさらに洗練させ、人間に匹敵するアクション効率（RHAE）を達成するための鍵となるだろう[38]。

ARC Prize 2026とオープンソースAGI研究の推進

ARC-AGI-3の正式リリースに伴い、Kaggleプラットフォーム上では賞金総額200万ドル（約3億円）を懸けた「ARC Prize 2026」コンペティションが開幕した[11]。この取り組みは、単なる技術的なコンテストではなく、AI開発の主導権を少数の巨大テック企業による「密室の開発」から取り戻し、オープンソースコミュニティによる透明性の高い科学的進歩へと回帰させることを目的としている[40]。

コンペティションの要件とマイルストーン

ARC-AGI-3トラックには総額85万ドルが割り当てられており、人間のベースライン（100%）に到達した最初のチームには70万ドルのグランドプライズが授与される[16]。このコンペティションの最も重要なルールは以下の2点である。

1. インターネットアクセスの禁止： 評価プロセス中、エージェントは外部ネットワークに接続できない[16]。これにより、GPT-5.4やClaude 4.6といったAPIベースの巨大商用モデルの利用が物理的に排除され、参加者は効率的で独立した自己完結型モデルの開発を迫られる[40]。
2. 完全なオープンソース化の義務： 賞金を受け取るためには、評価セットのスコアを受け取る前に、すべてのコード、手法、モデルウェイトをMIT-0やCC0などの寛容なオープンソースライセンスで公開しなければならない[16]。

2026年6月30日および9月30日に設定されたマイルストーン賞（各75,000ドル）は、コンペティション期間中であっても画期的な手法を早期に公開した研究者を称えるものであり、コミュニティ全体での知識の共有と再帰的な進歩（Refinement loop）を促すための巧みなインセンティブ設計となっている

[16]。

結論と次なる展望：真のAGIに向けた北極星

2026年3月のARC-AGI-3の登場は、AIの歴史において「計算資源の暴力によるパターンの暗記」から「未知の環境における適応的なスキル獲得」へと、評価のパラダイムが決定的に移行した瞬間として記憶されるだろう。

現在のフロンティアAIモデルが記録した0%台というスコアは、LLMが「世界を理解している」のではなく、単に「人間の思考プロセスが記録された広大なテキストの海を、確率的に航海しているに過ぎない」という事実を、冷徹なデータとして証明した[26]。静的なベンチマークが次々と飽和していく中で、自律的な探索、連続的な学習（Continuous learning）、および動的な世界モデルの構築を要求するARC-AGI-3は、現在「唯一の未飽和なエージェントAIベンチマーク」としてそびえ立っている[42]。

我々は今、AI開発の岐路に立っている。既存の自己回帰トランスフォーマーのパラメータを数兆規模に拡大し続けるだけでは、AGIには到達できない[11]。神経記号的AI（Neurosymbolic AI）、能動的推論（Active inference）、あるいは強化学習と探索アルゴリズムの新たな融合など、抜本的なアーキテクチャの革新が必要とされている[43]。

ARC Prize 2026が提供するオープンな競争環境を通じて、世界中の研究者たちがこの「パフォーマンスの崖」をどのように乗り越えていくのか。ARC-AGI-3のリーダーボード上で突然のスコア上昇が観測されたとき、我々は初めて、人類が真の汎用人工知能（AGI）の創造に成功したという、歴史的な証明を目撃することになるだろう[42]。

引用文献

1. ARC-AGI-3, 3月 26, 2026にアクセス、https://arcprize.org/arc-agi/3/
2. ARC-AGI-3 resets frontier AI scoreboard, 3月 26, 2026にアクセス、https://www.therundown.ai/p/arc-agi-3-resets-frontier-ai-scoreboard
3. ARC-AGI-3, 3月 26, 2026にアクセス、https://arcprize.org/arc-agi/3
4. What is ARC-AGI? - ARC Prize, 3月 26, 2026にアクセス、https://arcprize.org/arc-agi
5. arXiv:submit/7403127 [cs.AI] 24 Mar 2026 - ARC Prize, 3月 26, 2026にアクセス、https://arcprize.org/media/ARC_AGI_3_Technical_Report.pdf
6. Because ARC-AGI-3 reliably measures high IQ (145+) in both humans and AIs, we can finally know how super intelligent our AIs are becoming. - Reddit, 3月 26, 2026にアクセス、https://www.reddit.com/r/agi/comments/1raaig8/because_arcagi3_reliably_measures_high_iq_145_in/
7. The ARC of Progress towards AGI: A Living Survey of Abstraction and Reasoning - arXiv.org, 3月 26, 2026にアクセス、https://arxiv.org/html/2603.13372v1
8. Leaderboard - ARC Prize, 3月 26, 2026にアクセス、https://arcprize.org/leaderboard
9. ARC Prize 2025: Technical Report - arXiv, 3月 26, 2026にアクセス、

https://arxiv.org/html/2601.10904v1
10. The ARC of Progress towards AGI: A Living Survey of Abstraction and Reasoning | Request PDF - ResearchGate, 3月 26, 2026にアクセス、https://www.researchgate.net/publication/402479990_The_ARC_of_Progress_towards_AGI_A_Living_Survey_of_Abstraction_and_Reasoning
11. ARC Prize, 3月 26, 2026にアクセス、https://arcprize.org/
12. ARC Prize 2024: Technical Report, 3月 26, 2026にアクセス、https://arcprize.org/media/arc-prize-2024-technical-report.pdf
13. Graph-Based Exploration for ARC-AGI-3 Interactive Reasoning Tasks - OpenReview, 3月 26, 2026にアクセス、https://openreview.net/pdf?id=YGTxOepY49
14. Sensi: Learn One Thing at a Time—Curriculum-Based Test-Time Learning for LLM Game Agents - arXiv, 3月 26, 2026にアクセス、https://arxiv.org/html/2603.17683v1
15. Graph-Based Exploration for ARC-AGI-3 Interactive Reasoning Tasks - arXiv, 3月 26, 2026にアクセス、https://arxiv.org/html/2512.24156v1
16. ARC Prize 2026 - ARC-AGI-3 Competition, 3月 26, 2026にアクセス、https://arcprize.org/competitions/2026/arc-agi-3
17. ARC-AGI-3 Scoring Methodology, 3月 26, 2026にアクセス、https://docs.arcprize.org/methodology
18. ARC AGI 3 scores are not calculated the same way as ARC AGI 1 or 2 : r/singularity - Reddit, 3月 26, 2026にアクセス、https://www.reddit.com/r/singularity/comments/1s3ihv3/arc_agi_3_scores_are_not_calculated_the_same_way/
19. ARC-AGI-3 | Hacker News, 3月 26, 2026にアクセス、https://news.ycombinator.com/item?id=47521150
20. Every model failed this benchmark - Mastermind, 3月 26, 2026にアクセス、https://mastermindnewsletter.substack.com/p/every-model-failed-this-benchmark
21. ARC-AGI-3 Benchmark Unveiled to Evaluate AI Agents' Intelligence, 3月 26, 2026にアクセス、https://phemex.com/news/article/arcagi3-benchmark-unveiled-to-evaluate-ai-agents-intelligence-69185
22. ARC-AGI-3 offers $2M to any AI that matches untrained humans, yet every frontier model scores below 1% - The Decoder, 3月 26, 2026にアクセス、https://the-decoder.com/arc-agi-3-offers-2m-to-any-ai-that-matches-untrained-humans-yet-every-frontier-model-scores-below-1/
23. The 2026 Architecture Leap: A Blueprint for the Massive Acceleration of Synthetic Minds : r/Realms_of_Omnarai - Reddit, 3月 26, 2026にアクセス、https://www.reddit.com/r/Realms_of_Omnarai/comments/1s0lfai/the_2026_architecture_leap_a_blueprint_for_the/
24. The changing goalposts of AGI and timelines - Hacker News, 3月 26, 2026にアクセス、https://news.ycombinator.com/item?id=47299009
25. The Threshold of Superintelligence: A Comprehensive Analysis of AGI Development, Benchmarks, and Global Governance in 2026 | by sendy ardiansyah, 3月 26, 2026にアクセス、https://sendyardiansyah.medium.com/the-threshold-of-superintelligence-a-compre

hensive-analysis-of-agi-development-benchmarks-and-b84b13b85faa
26. The ARC-AGI leaderboard made me realize something terrifying (but weirdly comforting) about LLMs vs human brains : r/singularity - Reddit, 3月 26, 2026にアクセス、https://www.reddit.com/r/singularity/comments/1s3w9ls/the_arcagi_leaderboard_made_me_realize_something/
27. LLMs won't take us to AGI and this paper explains why : r/ArtificialInteligence - Reddit, 3月 26, 2026にアクセス、https://www.reddit.com/r/ArtificialInteligence/comments/1s2z5y0/llms_wont_take_us_to_agi_and_this_paper_explains/
28. SotA ARC-AGI-2 Results with REPL Agents | Symbolica Blog, 3月 26, 2026にアクセス、https://www.symbolica.ai/blog/arcgentica
29. Gemini 3.1 Pro - Model Card - Google DeepMind, 3月 26, 2026にアクセス、https://deepmind.google/models/model-cards/gemini-3-1-pro/
30. LangGraph templates to solve ARC-AGI-3 benchmark games | Plank, 3月 26, 2026にアクセス、https://www.joinplank.com/articles/arc-prize-langgraph
31. ARC-AGI-3 Preview Agent Competition, 3月 26, 2026にアクセス、https://arcprize.org/competitions/arc-agi-3-preview-agents/
32. ARC-AGI-3 Leaderboard, 3月 26, 2026にアクセス、https://three.arcprize.org/leaderboard
33. DriesSmit/ARC3-solution: My submission to the ARC-AGI-3 Developer Preview Agent Compitition. - GitHub, 3月 26, 2026にアクセス、https://github.com/DriesSmit/ARC3-solution
34. World's Top Large Models Suffer Severe Blow Overnight: Humans Score Full Marks, Top-Ranked AI Gets Just 0.2% in Toughest Test - 36氪, 3月 26, 2026にアクセス、https://eu.36kr.com/en/p/3739700584644867
35. Graph-Based Exploration for ARC-AGI-3 Interactive Reasoning Tasks - ResearchGate, 3月 26, 2026にアクセス、https://www.researchgate.net/publication/399276330_Graph-Based_Exploration_for_ARC-AGI-3_Interactive_Reasoning_Tasks
36. Graph-Based Exploration for ARC-AGI-3 Interactive Reasoning Tasks - arXiv, 3月 26, 2026にアクセス、https://arxiv.org/pdf/2512.24156
37. Sensi: Learn One Thing at a Time -- Curriculum-Based Test-Time Learning for LLM Game Agents - ResearchGate, 3月 26, 2026にアクセス、https://www.researchgate.net/publication/402739672_Sensi_Learn_One_Thing_at_a_Time_--_Curriculum-Based_Test-Time_Learning_for_LLM_Game_Agents
38. From artificial intelligence to active inference: the key to true AI and the 6G world brain [Invited] - Optica Publishing Group, 3月 26, 2026にアクセス、https://opg.optica.org/jocn/abstract.cfm?uri=jocn-18-1-A28
39. VERSES® “Digital Brain” Featured in WIRED and Popular Mechanics, 3月 26, 2026にアクセス、https://www.verses.ai/news/verses-digital-brain-featured-in-wired-and-popular-mechanics
40. ARC Prize 2026, 3月 26, 2026にアクセス、https://arcprize.org/competitions/2026
41. ARC Prize 2026 - ARC-AGI-3 | Kaggle, 3月 26, 2026にアクセス、

https://www.kaggle.com/competitions/arc-prize-2026-arc-agi-3
42. AI #161 Part 1: 80,000 Interviews - by Zvi Mowshowitz - Substack, 3月 26, 2026にアクセス、https://thezvi.substack.com/p/ai-161-part-1-80000-interviews
43. Why I don't think AGI is imminent - dlants.me, 3月 26, 2026にアクセス、https://dlants.me/agi-not-imminent.html
44. Future of AI Research - Association for the Advancement of Artificial Intelligence (AAAI), 3月 26, 2026にアクセス、https://aaai.org/wp-content/uploads/2025/03/AAAI-2025-PresPanel-Report-FINAL.pdf