ARC-AGIベンチマークにおける推論能力の進化とエージェント的知能の到達予測

Gemini3.1 pro

1. 序論：人工汎用知能（AGI）の測定とARCベンチマークの進化

人工汎用知能（AGI）の到達度を正確に測定することは、現代の人工知能研究および計算機科学における最大の課題の一つである。事前学習データの暗記やパターンマッチングに依存した従来の評価指標（標準的な推論ベンチマークや学力テストなど）が次々と飽和状態を迎える中、AI研究者のFrançois Cholletが2019年に発表した論文「On the Measure of Intelligence」において提唱された「抽象化と推論のコーパス（Abstraction and Reasoning Corpus: ARC-AGI）」は、真の「流動性知能（Fluid Intelligence）」を測定する極めて稀有かつ堅牢なベンチマークとして機能してきた[1]。流動性知能とは、過去に蓄積された膨大な知識データベース（結晶性知能）に依存するのではなく、全く未知のタスクや環境に直面した際に、その場で論理を構築し、適応する能力を指す[1]。

ARC-AGIの設計哲学は、人間の普遍的な認知基盤である「コア知識の事前条件（Core Knowledge Priors）」のみを前提としている点にある。これには、オブジェクトの永続性、基本的な幾何学的概念、トポロジーなどが含まれ、文化的な知識や言語への依存を完全に排除している[1]。AIモデルは、ごく少数の入出力例（デモンストレーション）から未知の変換ルールを推論し、新たな入力に対して正しい出力を生成することが求められる。2019年にリリースされた初代のARC-AGI-1は、数年にわたり最先端のAIモデルにとっての「推論の壁」として立ちはだかった。大規模言語モデル（LLM）のパラメーター規模を単に拡大させるだけではスコアは停滞し続けたが、2024年後半におけるテスト時計算（Test-time compute）およびテスト時適応（Test-time adaptation）技術の飛躍的な進化により、ついにモデルが人間のベースラインに肉薄する成果が記録された[5]。

この歴史的進展を受けて、推論モデルの真の限界と計算効率を再評価するため、2025年3月にはより複雑な構成的推論と文脈依存のルール適用を要求する「ARC-AGI-2」がリリースされた[2]。さらに、静的な推論から動的でインタラクティブな環境における継続学習へと評価の軸を完全に移した「ARC-AGI-3」のローンチが、2026年3月25日に控えている[2]。本報告書は、ユーザーの要求仕様に基づき、ARC-AGI-2のリリースからAIシステムが70%以上の性能を達成するまでにかかった時間を詳細なタイムラインと技術的背景とともに分析する。さらに、次世代ベンチマークであるARC-AGI-3において同等の70%以上の数値に達するまでの所要期間を、最新の実証的スケーリング則、エージェントシステムのアーキテクチャ上の制約、および専門家の見解を統合して理論的かつ実証的な観点から予測する。

2. ARC-AGI-2における70%到達までの軌跡と所要時間の分

析

2.1 ARC-AGI-2の設計思想と初期における性能の壁

2025年3月に公式発表されたARC-AGI-2は、AIの推論システムにおける計算効率と未知の環境への適応能力を極限までストレステストするために設計された次世代のベンチマークである[2]。初代ARC-AGI-1の評価データセットが長期にわたり研究コミュニティに広く公開された結果、一部の解法においてブルートフォース（総当たり）的なプログラム探索によるオーバーフィッティングの懸念が生じていた[9]。事実、ARC-AGI-1のタスクの一部は力技の探索によって解くことが可能であることが判明しており、知能の測定指標としての純度が低下しつつあった[10]。これを根本的に是正するため、ARC-AGI-2では、システムが暗記や単純なパターン認識に依存することを防ぐための厳格な設計が施された[11]。

具体的には、「構成的推論（Compositional Reasoning）」「文脈依存のルール適用（Contextual Rule Application）」、そして「意味論的シンボル解釈（Symbolic Interpretation）」という3つの高度な認知能力が不可欠なタスクが厳選された[2]。これらのタスクは、単一の大局的なルールを発見するだけでは不十分であり、複数のルールが相互に作用する状況や、視覚的なパターンを超えてシンボル自体に意味的意義を割り当てる能力を要求する[2]。さらに、ARC-AGI-2のタスクは、サンディエゴで実施された一般参加者400名以上による実証実験を経て、少なくとも2人の人間が2回以内の試行で解ける（pass@2）ことが確認されたもののみが採用されている[2]。

リリース当初の2025年3月におけるAIのパフォーマンスは、絶望的とも言える水準であった。純粋な大規模言語モデル（LLM）のゼロショット推論では0%のスコアしか記録できず、当時最先端とされたOpenAIのo3モデル（ARC-AGI-1において87.5%という驚異的なスコアを達成した高計算モデル）であっても、ARC-AGI-2の難度の前では低・中程度の推論設定で3%未満のスコアに沈んだ[5]。高推論設定（o3-high）においても、出力カバレッジが著しく低下し、120のタスク中15タスクしか回答を生成できず、その精度もわずか6%に留まった[12]。この「人間にとって容易だが、AIにとっては困難」というギャップの再構築により、ARC-AGI-2は真の流動性知能を測定する指標としての地位を再び確立したのである[7]。

2.2 パフォーマンス向上の時系列分析：11ヶ月間の飛躍的進化

当初は完全に未解決とされたARC-AGI-2であったが、そのスコアが70%の閾値を超えるまでのプロセスは、モデルのパラメーター規模の単純な拡大（Scaling Up）ではなく、テスト時推論の抜本的なパラダイムシフトによって主導された。この技術的進展は、2025年3月から2026年2月にかけての約11ヶ月間で、以下の3つの明確なフェーズを経て進行した。

第一のフェーズは、2025年中旬から秋にかけての「オープンソースコミュニティと小規模モデルによる初期突破」である。ARC Prize 2025（2025年3月26日から11月3日まで開催）を通じて、独立系研究者やKaggleグランドマスターたちが、巨大なコンピュートに依存しない効率的な推論アーキテクチャを開拓した[7]。NVIDIAの研究者チーム「NVARC」は、わずか40億（4B）パラメーターの微調整モデルと、合成データ駆動のアンサンブルを組み合わせることで、タスクあたりわずか0.20ドルの超低コストで24.03%のスコアを達成し、Kaggleコンペティションで優勝を果たした[11]。これは、小規模モデルで

あってもテスト時適応（Test-time adaptation）の仕組みを高度に洗練させれば、巨大なベースモデルを凌駕できることを証明した歴史的なマイルストーンであった [8]。また同時期に、Tiny Recursive Models (TRM)やCompressARCといった、事前学習に依存せずテスト時にモデルの重みを直接最適化する手法が評価を高めた [8]。

第二のフェーズは、2025年末から2026年1月にかけて展開された「洗練ループ（Refinement Loops）と商用フロンティアモデルの適応」である。主要AIラボは、コンペティションで培われた知見を自社の巨大モデルに統合し始めた。AnthropicのClaude Opus 4.5は、64kの思考予算（Thinking budget）を割り当てることで37.6%を記録し [13]、PoetiqによるGemini 3 Proをベースとしたカスタマイズされた洗練ソリューションは、タスクあたり30ドルの高コストを投じることで54%という過半数の壁を突破した [13]。そして、2026年初頭には、OpenAIのGPT-5.2モデルが推論アーキテクチャの最適化により52.9%に到達し、さらにJohan Land氏によるGPT-5.2の洗練モデルが72.9%を達成するなど、70%の壁に迫る動きが加速した [15]。

第三のフェーズにして決定的な突破口は、2026年2月に開かれた「推論専用モードによる70%の突破と定着」である。AnthropicのClaude Opus 4.6は、120kの推論トークンと最高レベルの計算リソース（Max effort）を投じることで、68.8%から69.17%というほぼ70%に等しいスコアを叩き出した [15]。そして2026年2月12日、Google DeepMindがリリースした「Gemini 3 Deep Think」は、推論チェーンの大幅な拡張と並列仮説探索（Parallel Hypothesis Exploration）を実装し、ARC-AGI-2において84.6%という驚異的なスコアを記録した [15]。これに続く形で、2026年2月19日にプレビュー公開されたGemini 3.1 Proが77.1%を達成し [15]、OpenAIのGPT-5.4 Pro（xHigh設定）も83.3%を記録した [15]。さらに、同時期に発表された「Product of Experts」アプローチは、モデル自身を生成器および評価器として機能させることで、タスクあたりわずか0.02ドルの低コストで71.6%を達成し、効率性の観点でも歴史的なブレイクスルーを果たした [24]。

以下の表は、ARC-AGI-2におけるAIパフォーマンスの推移を時系列で整理したものである。データは、2025年3月のリリース時から2026年2月のブレイクスルーまでのスコアの上昇トレンドを明確に示している。

達成時期	モデル / システム名	スコア (%)	タスクあたりのコスト ($)	備考・技術的特徴
2025年3月	OpenAI o3 (Medium)	2.9%	2.52	ARC-AGI-2リリース初期の最高水準 [12]
2025年11月	NVARC (Kaggle優勝)	24.0%	0.20	4Bモデルと合成データアンサンブル [11]

2025年12月	Claude Opus 4.5 (64k)	37.6%	2.20	商用モデルによる洗練ループの導入 [13]
2026年1月	GPT-5.2	52.9%	0.75	エージェント的コーディングに最適化 [16]
2026年1月	Poetiq (Gemini 3 Refinement)	54.0%	30.00	Gemini 3 Proベースの高度な洗練ソリューション [13]
2026年2月	Claude Opus 4.6 (120k, Max)	68.8% - 69.17%	3.64	120k思考トークンの投入 [15]
2026年2月	Product of Experts	71.6%	0.02	モデルの自己生成・自己評価による超高効率化 [24]
2026年2月	Gemini 3.1 Pro (Preview)	77.1%	0.96	コア推論能力のステップアップ [15]
2026年2月	GPT-5.4 Pro (xHigh)	83.3%	16.41	大規模な推論時計算の適用 [15]
2026年2月	**Gemini 3 Deep Think**	**84.6%**	13.62	並列仮説探索と推論チェーンの極大化 [15]

2.3 結論：ARC-AGI-2が70%に達するまでの所要時間

上記のデータと歴史的経緯の分析から、一つの明確な結論が導き出される。ARC-AGI-2が2025年3月に公式にリリースされてから、AIモデルが継続的かつ公式に70%以上のスコアを達成する（2026年1月下旬から2月中旬にかけて）までにかかった時間は、およそ**10**ヶ月から**11**ヶ月であった [2]。

この所要時間は、AIの歴史において極めて重要な意味を持つ。初代ARC-AGI-1において、AIのスコアがゼロから人間に近い水準（o3による87.5%など）に達するまでには、2019年から2024年末までの約5年という長い歳月を要した [5]。それに比べ、推論の複雑さにおいて格段に難易度が高いとされるARC-AGI-2が、わずか1年足らずで攻略された事実は、AI業界における学習パラダイムが「事前学

習による知識の圧縮とパターンの暗記」から「テスト時における適応とプログラム合成（Test-time adaptation and program synthesis）」へと完全に移行し、その最適化の速度が劇的に加速していることを実証している[8]。この11ヶ月間は、単なる計算リソースの暴力的な投入ではなく、推論アーキテクチャそのものの効率化と洗練が牽引した期間であったと言える。

3. ARC-AGI-2を攻略した技術的パラダイムシフトとコスト効率の分析

わずか11ヶ月で70%の壁を突破した背景には、従来の「事前学習スケーリング則（Pretraining Scaling Laws）」に依存しない、アーキテクチャ上の決定的な革新が存在した。本章では、スコアを押し上げた技術的要因と、ARC-AGIが重視する「知能の効率性」について詳述する。

3.1 テスト時計算（Test-Time Compute）と推論スケーリングの成熟

2024年後半から2026年初頭にかけての最大の技術的ブレイクスルーは、推論時（Inference）により多くの計算資源を動的に割り当てることで、モデルの正答率を飛躍的に高める「テスト時計算（Test-time compute）」の確立と成熟である[5]。Gemini 3 Deep ThinkやGPT-5.4 Proなどの最先端モデルは、単一のフォワードパスで直感的な回答を生成するSystem 1思考ではなく、内部的に長大な推論チェーン（Chain-of-Thought）を展開し、解決策を多段階に分解して自己検証を行うSystem 2思考を模倣している[13]。

Gemini 3 Deep Thinkのケースでは、複雑なタスクに対して最大138,000トークンもの推論プロセスを生成し、並行して複数の仮説を探索（Best-of-N selection）し、一貫性チェックを行うアプローチが採用された[13]。これにより、標準モデルであるGemini 3 Proが31%に留まっていたのに対し、Deep Thinkモードは84.6%という驚異的なスコアを叩き出した[13]。このアプローチは、モデルが自らの推論の誤りをリアルタイムで検知し、軌道修正する能力を付与するものであり、ARC-AGI-2が求める多段階の構成的推論の要件と完全に合致していた。

3.2 進化型洗練ループ（Refinement Loops）とプログラム合成の融合

単なる自然言語による思考チェーンを超えてスコアを決定的に押し上げたのが、「洗練ループ（Refinement Loops）」と「プログラム合成（Program Synthesis）」の融合である[8]。ARC-AGIのタスクは、本質的に「入力グリッドを出力グリッドに変換する未知のアルゴリズムを少数例から推論し、プログラムとして表現する」作業と同義である。最新のシステムは、LLMに直接グリッドの答えを出力させるのではなく、Pythonなどのプログラムコードを生成させ、それを実行環境でテストし、発生したエラーや出力のズレ（フィードバック信号）に基づいてコードを反復的に修正・進化させるアプローチを標準化させた[13]。

ARC Prize 2025の優秀論文で示された「Evolutionary Test-Time Compute（進化的テスト時計算）」や「Evolutionary Program Synthesis」は、この反復プロセスを極限まで最適化し、事前学習に依存せずにテスト環境そのものでモデルの振る舞いや重みを調整する手法を確立した[13]。例えば、Alexia Jolicoeur-Martineauらによる論文賞1位の「CompressARC」や、Tiny Recursive Models (TRM)のアーキテクチャは、パラメーター数がわずか数百万から数千万規模でありながら、再帰的な推論と最

小記述長（MDL）に基づく圧縮原理を利用することで、巨大モデルに匹敵する構成的推論能力を発揮した[8]。

3.3 知能の定義としてのコスト効率と「Product of Experts」

François Cholletは、知能の本質を単なるタスクの達成度ではなく、「リソースを最小限に抑えながらスキルを獲得する効率性」と厳密に定義している[1]。ARC-AGIのリーダーボードにおいて、正答率のスコアだけでなく「タスクあたりの推論コスト（Cost per Task）」が極めて重要な指標として扱われているのはこのためである[15]。真の知能とは、無限の計算資源を用いたブルートフォース探索ではなく、限られた資源の中で最適解を導き出す能力に他ならない[15]。

初期のテスト時計算は極めて非効率的であり、一部の推論モデルはタスクあたり数千ドルの計算コストを消費していた（例：2024年末時点のo3モデルではタスクあたり約4500ドルと試算されていた）[8]。しかし、2025年末から2026年初頭にかけて、このコストは劇的に低下した。GPT-5.2の段階でコストは12ドル付近まで下落し、1年で390倍のコスト削減が実現された[8]。

以下の表は、ARC-AGI-2における主要システムのアプローチと、スコアに対するコスト効率の相関関係を示したものである。データが示す通り、巨大なフロンティアモデルが高コストをかけて力技でスコアを伸ばす一方で、アーキテクチャを最適化した小規模・中規模のシステムが極めて低いコストで高い知能効率を示していることがわかる。

システムカテゴリ	モデル / 手法	スコア (%)	タスクコスト ($)	効率性の評価
Kaggle最適化	NVARC (4B)	24.0%	0.20	小規模モデルによる極めて高いコスト効率[13]
ベース**LLM**	Claude Opus 4.5 (64k)	37.6%	2.20	標準的な商用モデルのベースライン効率[15]
洗練システム	Poetiq Refinement	54.0%	30.00	高スコアだがコスト効率は低い（力技の洗練）[13]
ベース**LLM**	Claude Opus 4.6 (Max)	68.8%	3.64	商用ベースモデルとしての最高峰のバランス[15]

効率化システム	**Product of Experts**	**71.6%**	**0.02**	データ拡張と自己評価による究極のコスト効率 [24]
推論システム	GPT-5.4 Pro (xHigh)	83.3%	16.41	大規模な計算資源への依存による高スコア獲得 [15]
推論システム	Gemini 3 Deep Think	84.6%	13.62	同上、計算リソースの暴力による壁の突破 [15]
ベースライン	Human (人間)	100.0%	17.00	汎用知能としての絶対的基準（人件費ベース） [15]

このデータの分析から特筆すべき点は、「Product of Experts」と呼ばれるアプローチの台頭である。この手法は、オープンソースモデルに高度なデータ拡張（Data augmentation）とテスト時学習を組み合わせ、モデル自身を生成器とスコアラーの双方として機能させることで、人間を超える71.6%の精度をタスクあたりわずか0.02ドルで達成した [24]。このような超高効率な推論アーキテクチャの確立は、フロンティアラボが追求する「スケーリングの力学」とは異なるアプローチが、AIにAGIの要件である「学習効率」をもたらしつつあることを証明している [8]。コストの観点から見れば、AIはすでに人間の数百分の一のコストで同等の推論能力を獲得しつつあると言える。

4. ARC-AGI-3の導入：静的推論からインタラクティブなエージェント的知能への飛躍

ARC-AGI-2が商用の推論モデルや洗練されたオープンソース手法によって攻略されつつある中、ベンチマークの設計者であるFrançois CholletおよびARC Prize Foundationは、AGIへの次なる北極星（North Star）として「ARC-AGI-3」を2026年3月25日にローンチする [2]。ARC-AGI-3は、これまでのバージョンとは根本的に異なるパラダイムを採用しており、知能の定義を「静的なパターン認識の効率性」から「動的なエージェント的知能（Agentic Intelligence）の適応性」へと昇華させている [2]。この変革は、AI研究における新たな「冬」あるいは強固な「壁」をもたらす可能性が高い。

4.1 静的タスクからインタラクティブな継続学習環境への完全移行

初代ARC-AGI-1およびARC-AGI-2は、固定されたグリッドベースの入出力画像ペア（通常30x30以下のセルで構成される）を観察し、背後にある変換ルールを推論する「静的な（Static）」テストであった [2]。AIモデルは推論時にどれほど膨大な時間を費やそうとも、環境自体に介入して能動的にフィー

ドバックを得ることはできなかった。つまり、与えられた情報から演繹的あるいは帰納的に答えを導き出す「受動的な流動性知能」の測定に特化していた。

これに対し、ARC-AGI-3は、全く新しい「インタラクティブなターン制の環境（Interactive turn-based environments）」を採用している[2]。テストのフォーマットは静止画からゲームライクな動的環境へと変化した。さらに決定的な違いとして、AIエージェントには自然言語による事前の指示や目標（ゴール）は一切与えられない。エージェントは未知の環境に放り込まれ、自ら環境内を探索（Exploration）し、行動を起こし、その結果から環境の法則性（メカニクス）と勝利条件（Win conditions）をその場で推論・学習し、目標を達成しなければならない[2]。

4.2 ARC-AGI-3が測定する4つのエージェント能力の柱

ARC-AGI-3のテクニカルペーパーによれば、このベンチマークは単なる強化学習アルゴリズムのテストではなく、人間の認知発達に近い以下の4つの柱を通じてエージェント的知能を統合的に評価する[2]。

1. 探索（**Exploration**）：未知の環境において、ランダムに行動するのではなく、有益な情報を積極的に取得するための戦略的な行動選択能力。環境の境界やオブジェクトの性質をテストする能力が含まれる。
2. モデリング（**Modeling**）：生の観察データ（Perception）の断片から、環境全体の法則を説明する汎用的な「内部世界モデル（World Model）」を構築し、短期記憶を効率的に圧縮する能力。
3. 目標設定（**Goal-Setting**）：明示的な指示やプロンプトなしに、環境内での望ましい未来の状態や勝利条件（何がクリア条件なのか）を自律的に特定し、設定する能力。
4. 計画と実行（**Planning and Execution**）：スパースな（まばらな）フィードバックの下で、長期的な行動計画を立案し、環境からの新たな証拠や予期せぬ結果に基づいて即座に戦略を修正（コース補正）する俊敏性。

この複雑なインタラクションプロセスを高速かつスケーラブルに評価するため、ARC-AGI-3の開発チームはUnityのような重い商用ゲームエンジンを避け、1秒間に1000フレームの反復処理が可能な独自のPythonベースの環境エンジンを内製している[2]。

4.3 新たな評価基準「RHAE（相対的人間行動効率）」によるブルートフォースの無効化

ARC-AGI-3の難易度を劇的に引き上げ、現在のAIシステムを無力化している最大の要因は、そのスコアリング手法にある。従来のARC-AGIでは「最終的に正しい出力グリッドを生成できたか否か（Accuracy）」というタスク特化型の評価が用いられていた[2]。しかし、ARC-AGI-3では「RHAE（Relative Human Action Efficiency：相対的人間行動効率）」という全く新しいフレームワークへと移行した[2]。

RHAEのメカニズムは極めて厳格である。この指標は、AIエージェントが環境のルールを理解し、実際に目標を達成するまでに消費した「行動の数（ステップ数）」を計測し、それを人間のベースライン（具体的には、テストされた人間の中で2番目に効率の良かった人間の行動数）と相対的に比較する[2]。人間は初めてプレイするパズルゲームであっても、総当たりで行動するのではなく、過去の経験

（コア知識）に基づいて素早くメンタルモデルを構築し、少数の仮説を的確に検証し、戦略を瞬時に洗練させるため、極めて少ないステップ数でクリアすることが可能である[28]。

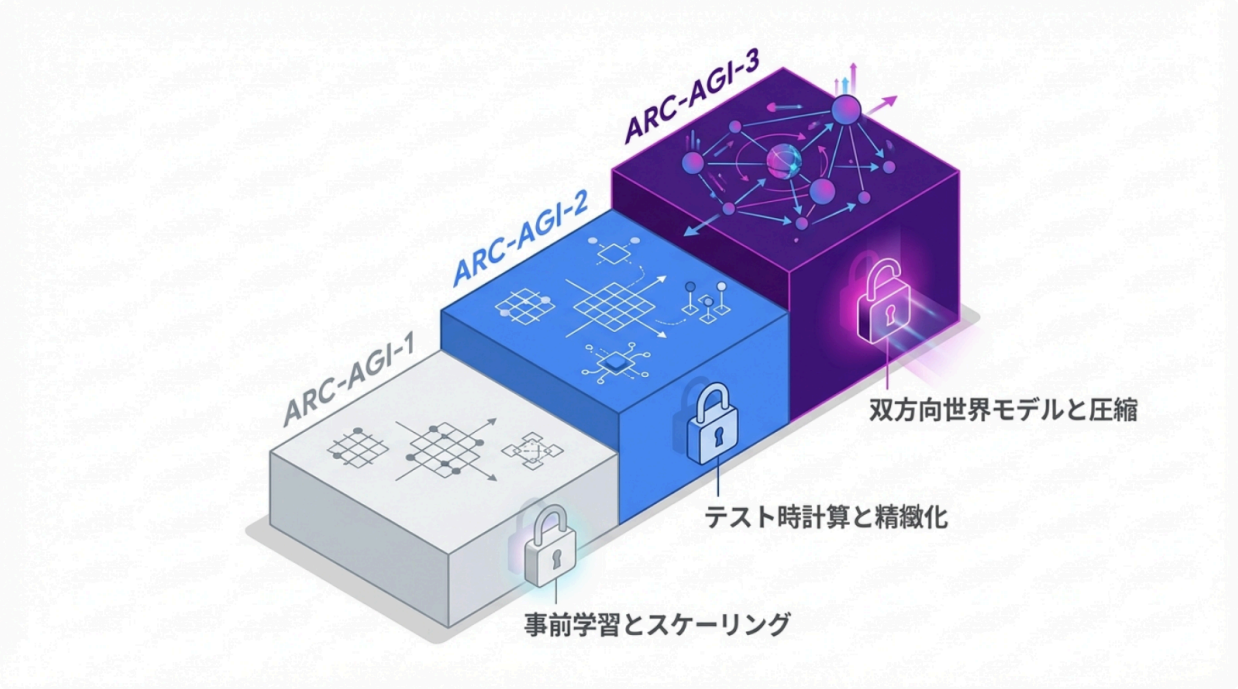
RHAEの評価体系において、AIが最終的に環境をクリアできたとしても、人間より10倍多いステップ数を費やして試行錯誤した場合、スコアは劇的に低下（例えば1%未満など）するように設計されている。採点基準は線形ではなく、「べき乗則の効率項（Power-law efficiency term）」に従うため、無駄な行動は致命的なペナルティとなる[2]。これにより、現在のAIが得意とする「強化学習やテスト時計算を用いた試行錯誤によるブルートフォース探索」を完全に無効化し、「未知の概念に対する学習の効率性」そのものを厳密に測定することが可能となったのである[2]。

以下の表は、ARC-AGIシリーズの進化の歴史を、評価対象、フォーマット、およびAIの攻略パラダイムの観点から比較したものである。

ベンチマーク特徴	**ARC-AGI-1 (2019年リリース)**	**ARC-AGI-2 (2025年リリース)**	**ARC-AGI-3 (2026年リリース)**
評価の対象	静的パターンの抽象化と推論	構成的・文脈依存の静的推論	エージェント的知能・継続学習の効率
フォーマット	静的なグリッド画像（2D配列）	静的なグリッド画像（難度上昇）	インタラクティブなターン制動的環境
ゴールの提示	入出力の例示から明示的に類推	入出力の例示から明示的に類推	非明示（環境からの自律的発見が必須）
人間のベースライン	ほぼ100%（平均正答率 77%〜）	ほぼ100%（Pass@2による校正）	**100% (極めて少ないステップ数でクリア)**
評価スコアの基準	最終的な出力の正確性 (Accuracy)	最終的な出力の正確性 (Accuracy)	行動効率性 **(RHAE:** 人間との相対比較**)**
AIの攻略パラダイム	テスト時推論スケーリ	洗練ループ / プログラ	未解決（リアルタイム世

	ング / o3	ム合成	界モデル構築が必要）

ARC-AGIの進化と要求される知能のパラダイムシフト



ARC-AGI-1と2は「テスト時計算」の拡張によって突破されたが、ARC-AGI-3が求める「エージェント的知能」を攻略するには、リアルタイムでの環境モデリングと目標推論を行う全く新しいアーキテクチャ（第3の波）が必要となる。

5. ARC-AGI-3においてAIが70%に到達するまでの期間予測

ARC-AGI-2が約11ヶ月という短期間で70%を突破したのに対し、ARC-AGI-3で同等の成果を収めるまでにはどれだけの時間を要するのか。現在の性能データ、エージェントシステムに関する最新の実証的スケーリング則、および業界のトップ専門家による予測を総合的に分析すると、**ARC-AGI-3**の攻略（**70%**以上のスコア達成）には、リリース（**2026**年**3**月）から**3**年～**5**年、すなわち**2029**年から**2031**年頃までの長期的なタイムラインが見込まれるという結論に至る。本章では、その予測を裏付ける3つの技術的・歴史的根拠を論証する。

5.1 初期スコアの絶望的なギャップと飽和していないベンチマーク

2026年3月のリリース直前、および2025年7月に実施されたプレビューの段階において、ARC-AGI-3の難易度は圧倒的であることが確認されている。人間のプレイヤー1,200人以上が参加したテストで

は、人間は短いプレイセッション（約20分以内）で100%の環境をクリアし、かつ極めて少ないステップ数で目標に到達する高い行動効率を示した [2]。

これと対照的に、最先端のフロンティアAIモデルであっても、RHAEに基づくスコアはプライベートテストセットにおいて「1%未満」に留まっており、比較的容易なプレビュー版のゲームにおいても、最も優れたAIシステム（StochasticGooseなど）で最大「約12.58%〜13%」の行動効率しか記録できていない [3]。この90ポイント近い人間とのギャップは、ARC-AGI-2の初期段階（0〜3%程度）と似ているように見えるかもしれない。しかし、後述するスケーリング則の観点から見ると、この壁の厚さと性質は根本的に異なるものである。François Cholletは、「現在、ARC-AGI-3は飽和していない唯一のエージェント型AIベンチマークである」と明言しており、このリーダーボードでの突発的なスコア上昇は、AIの能力に関する重大なパラダイムシフト（AGIへのブレイクスルー）を意味すると指摘している [30]。

5.2 エージェントシステムにおける既存のスケーリング則の崩壊

ARC-AGI-2をわずか11ヶ月で攻略できた最大の要因は、テスト時推論（System 2）に膨大な計算リソースを投じることで性能が対数直線的に向上するという「テスト時スケーリング則（Test-time scaling laws）」の恩恵をフルに受けたことである [25]。AIは、静的な問題に対して並列に数万の仮説を生成し、自己検証を繰り返すことで正答に辿り着くことができた。しかし、ARC-AGI-3のような連続的でインタラクティブな環境においては、このアプローチがそのまま機能しないことが実証されている。

Googleが発表した最新の研究論文「Towards a Science of Scaling Agent Systems」によると、単一タスクの正答を求める従来モデルとは異なり、マルチステップで環境と相互作用するエージェントシステムにおいては、「単にエージェントの数を増やす（推論リソースを増やす）」アプローチはすぐに天井に達することが示されている。同研究における180のエージェント構成の評価では、タスクが直列的（Sequential）な性質を持つ場合、不要な推論ステップがエラーを連鎖的に増幅させる「破滅的故障モード（Catastrophic Failure Modes）」を引き起こし、かえって性能を低下させることが判明した [32]。

ARC-AGI-3では、エージェントは一つ前の自身の行動の結果に基づいて次の行動を決定する必要があり、かつ環境からのフィードバックは極めてスパース（まばら）である [2]。このため、静的なタスクで有効だった並列探索（Parallel Hypothesis Exploration）の利点が大きく削がれる。現在のLLMやVLM（視覚言語モデル）は、事前の大規模データセットに基づく次トークンの予測（Prediction）に最適化されており、「未知の環境における目標の自律的発見」や「少ないサンプルからの世界モデルの構築」に向けた学習メカニズムを根本的に欠いているのである [8]。

5.3 突破に不可欠な「第3のパラダイムシフト」と専門家のコンセンサス

ARC-AGI-3においてRHAE 70%以上のスコアを達成するためには、LLMのスケールアップ（Pretraining Scaling）や推論時の計算増大（Test-time Scaling）の延長線上にない、**「第3のパラダイムシフト」**が不可欠である。学術界やARC Prizeのトップ研究者たちは、その突破口を以下の二つのアーキテクチャ的革新に見出している。

第一に、**MDL（Minimum Description Length:** 最小記述長）に基づく情報の圧縮である。現在の自己回帰的（Auto-regressive）な言語モデルは、パターンの暗記に傾倒しやすい。真の人間に近い知

能を獲得するには、情報を確率的に予測するのではなく、観察された現象を最も簡潔なルールとして「圧縮（Compression）」することに明示的な報酬を与える最適化目標が必要となる[8]。ARC Prize 2025で成果を上げたCompressARCやTiny Recursive Models (TRM)のような、圧縮原理に基づくアプローチの発展系が、エージェントの内部世界モデル構築に採用される可能性が高い[8]。

第二に、継続学習（**Continual Learning**）と動的なエピソード記憶の統合である。ARC-AGI-3では、事前のプロンプトが存在しないため、エージェントはプレイを通じて学んだ環境ルールをリアルタイムで自身の内部表現に組み込む必要がある[2]。重みが固定された現在のニューラルネットワークは、このオンザフライの学習（Few-shot adaptation）に極めて弱いため、外部メモリモジュールと連動しながら動的に推論パスを再構築する神経記号的（Neuro-symbolic）なアーキテクチャの実用化が待たれる。

これらの根本的なアーキテクチャの刷新にかかる時間を考慮すると、過去の技術的マイルストーンとの比較が有用である。

- 第**1**波（静的パターン認識）: ARC-AGI-1のリリース（2019年）から、テスト時計算の導入によるブレイクスルー（2024年末）まで、約**5**年の歳月を要した[5]。
- 第**2**波（構成的推論の最適化）: OpenAIのo1/o3系など、新たな推論アーキテクチャの土台がすでに完成していたため、難度を上げたARC-AGI-2（2025年3月）の突破には約**11**ヶ月という短期間しかかからなかった[21]。
- 第**3**波（動的エージェント知能）: ARC-AGI-3（2026年3月）は、AIが要求されるアーキテクチャの前提が根本的に異なるため、第2波のような「既存のテスト時推論技術の延長（スケーリング）」では解決できない[8]。したがって、パラダイムの移行期間として、第1波と同等かそれに近い開発サイクルが必要となる。

さらに、François Chollet自身は、ARC-AGIでの継続的なベンチマーク更新を通じてAIと人間のギャップをゼロにするプロセスこそがAGIへの道程であると述べており、真のAGIの到来を2030年頃と予測している[13]。ARC-AGI-3において人間と同等の行動効率（70%以上のRHAEスコア）を獲得することは、事実上のAGI要件を満たすことに等しい[30]。コミュニティの研究開発サイクルと、自己改善型メタ学習システムの実用化にかかる期間を統合すると、ARC-AGI-3でAIが70%を超えるまでの期間は、**リリースから少なくとも3年～5年（すなわち2029年から2031年の間）**を要するという予測が、最も蓋然性の高いシナリオとして導出される。

6. 結論：AGIへの道程と知能評価の未来

ARC-AGIベンチマークの進化の歴史は、人工知能が単なる「大量の知識を暗記した確率的な辞書」から、「その場で論理を構築し適応する思考エンジン」へと進化していく過程を極めて正確にトレースしてきた。本報告書の分析が示す通り、ARC-AGI-2のリリースからわずか11ヶ月で70%の壁が突破されたという事実は、テスト時計算（Test-time compute）と洗練ループ（Refinement loops）という技術がもたらした破壊的な進歩を雄弁に物語っている。タスクあたり数千ドルかかっていた推論コストが、わずか数セントの最適化手法へと落とし込まれたことは、知能の効率化という観点で歴史的な偉業であると言える。

しかし、2026年3月に登場するARC-AGI-3は、AI研究業界に再び強固な「アーキテクチャの壁」を突きつける。環境の能動的な探索、ルールの自律的発見、そして内部世界モデルのリアルタイムな構築という「エージェント的知能」の要件は、現在のスケーリング則の単純な延長線上には存在しない。AIがARC-AGI-3において人間と同等の行動効率（70%以上のスコア）を獲得するには、最適化の目標を「予測」から「圧縮」へと転換し、動的な継続学習を可能にする抜本的なパラダイムシフトが必要であり、その達成には2029年から2031年頃までの期間を要すると予測される。ARC-AGI-3のリーダーボードにおいて、人間と同水準の行動効率を示すスコアが観測されたその時こそ、人類が真の人工汎用知能（AGI）の入り口に明確に足を踏み入れた瞬間となるだろう。

引用文献

1. What is ARC-AGI? - ARC Prize, 3月 26, 2026にアクセス、https://arcprize.org/arc-agi
2. ARC-AGI-3 - ARC Prize, 3月 26, 2026にアクセス、https://arcprize.org/arc-agi/3/
3. 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
4. slides.pdf - The Science of Benchmarking, 3月 26, 2026にアクセス、https://benchmarking.science/slides.pdf
5. ARC-AGI-2, 3月 26, 2026にアクセス、https://arcprize.org/arc-agi/2/
6. OpenAI o3 Breakthrough High Score on ARC-AGI-Pub, 3月 26, 2026にアクセス、https://arcprize.org/blog/oai-o3-pub-breakthrough
7. Announcing ARC-AGI-2 and ARC Prize 2025, 3月 26, 2026にアクセス、https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025
8. 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
9. Multimodal Reasoning to Solve the ARC-AGI Challenge | Maximilian Seeth, 3月 26, 2026にアクセス、https://omseeth.github.io/blog/2025/MLLM_for_ARC/
10. why is arc-agi-v2 so much harder for AIs than v1? is it contamination? : r/singularity - Reddit, 3月 26, 2026にアクセス、https://www.reddit.com/r/singularity/comments/1kx1h3b/why_is_arcagiv2_so_much_harder_for_ais_than_v1_is/
11. NVIDIA Kaggle Grandmasters Win Artificial General Intelligence Competition, 3月 26, 2026にアクセス、https://developer.nvidia.com/blog/nvidia-kaggle-grandmasters-win-artificial-general-intelligence-competition/
12. Analyzing o3 and o4-mini with ARC-AGI, 3月 26, 2026にアクセス、https://arcprize.org/blog/analyzing-o3-with-arc-agi
13. ARC Prize 2025 Results and Analysis, 3月 26, 2026にアクセス、https://arcprize.org/blog/arc-prize-2025-results-analysis
14. Claude Opus 4.5 System Card - Anthropic, 3月 26, 2026にアクセス、https://www.anthropic.com/claude-opus-4-5-system-card
15. Leaderboard - ARC Prize, 3月 26, 2026にアクセス、https://arcprize.org/leaderboard
16. LLM Leaderboard - Vellum, 3月 26, 2026にアクセス、https://vellum.ai/llm-leaderboard

17. Claude Opus 4.6 vs GPT-5.3 Codex - UniFuncs, 3月 26, 2026にアクセス、https://unifuncs.com/s/Dw7uxWVs
18. GPT-5.2 Complete Guide: Features, Benchmarks & API - Digital Applied, 3月 26, 2026にアクセス、https://www.digitalapplied.com/blog/gpt-5-2-complete-guide
19. Claude Opus 4.6 System Card - Anthropic, 3月 26, 2026にアクセス、https://www-cdn.anthropic.com/0dd865075ad3132672ee0ab40b05a53f14cf5288.pdf
20. Gemini 3 Deep Think: Reasoning Benchmarks & Complete Guide - Digital Applied, 3月 26, 2026にアクセス、https://www.digitalapplied.com/blog/gemini-3-deep-think-reasoning-benchmarks-guide
21. Google Gemini - Wikipedia, 3月 26, 2026にアクセス、https://en.wikipedia.org/wiki/Google_Gemini
22. Gemini 3 Deep Think - ARC-AGI 2 score of 84.6% : r/accelerate - Reddit, 3月 26, 2026にアクセス、https://www.reddit.com/r/accelerate/comments/1r331kx/gemini_3_deep_think_arcagi_2_score_of_846/
23. A new era of intelligence with Gemini 3 - Google Blog, 3月 26, 2026にアクセス、https://blog.google/products-and-platforms/products/gemini/gemini-3/
24. Product of Experts approach achieves 71.6% on ARC-AGI (beats human baseline) at $0.02/task : r/learnmachinelearning - Reddit, 3月 26, 2026にアクセス、https://www.reddit.com/r/learnmachinelearning/comments/1p7dayf/product_of_experts_approach_achieves_716_on/
25. How Scaling Laws Drive Smarter, More Powerful AI - NVIDIA Blog, 3月 26, 2026にアクセス、https://blogs.nvidia.com/blog/ai-scaling-laws/
26. OpenAI o3 is leading the newly announced ARC-AGI-2, but no AI is getting above 4%, 3月 26, 2026にアクセス、https://www.reddit.com/r/OpenAI/comments/1jjdhki/openai_o3_is_leading_the_newly_announced_arcagi2/
27. ARC Prize, 3月 26, 2026にアクセス、https://arcprize.org/
28. Introducing ARC-AGI-3 : r/LocalLLaMA - Reddit, 3月 26, 2026にアクセス、https://www.reddit.com/r/LocalLLaMA/comments/1s3ll4i/introducing_arcagi3/
29. ARC-AGI-3 - Hacker News, 3月 26, 2026にアクセス、https://news.ycombinator.com/item?id=47521150
30. 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
31. The State of AI Models: Scaling, Reasoning, and Agentic Intelligence | by Shashank Guda, 3月 26, 2026にアクセス、https://shashankguda.medium.com/the-state-of-ai-models-scaling-reasoning-and-agentic-intelligence-ee53a29722a6
32. Towards a science of scaling agent systems: When and why agent systems work, 3月 26, 2026にアクセス、https://research.google/blog/towards-a-science-of-scaling-agent-systems-when-and-why-agent-systems-work/
33. Towards a Science of Scaling Agent Systems - arXiv, 3月 26, 2026にアクセス、

https://arxiv.org/html/2512.08296v1
34. If Sutskover is right about a scaling wall, we have no choice but pivot to stronger and more extensive logic and reasoning algorithms : r/agi - Reddit, 3月 26, 2026にアクセス、https://www.reddit.com/r/agi/comments/1p81lnz/if_sutskover_is_right_about_a_scaling_wall_we/
35. Poetiq's ARC-AGI scores now in the works of getting verified : r/accelerate - Reddit, 3月 26, 2026にアクセス、https://www.reddit.com/r/accelerate/comments/1p8ay6e/poetiqs_arcagi_scores_now_in_the_works_of_getting/
36. François Chollet: How We Get To AGI - YouTube, 3月 26, 2026にアクセス、https://www.youtube.com/watch?v=5QcCeSsNRks