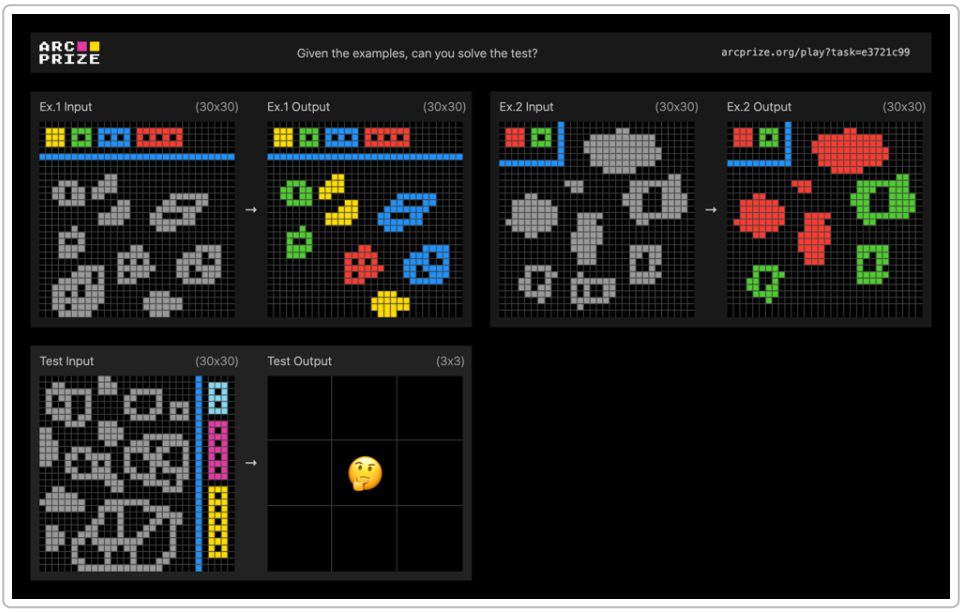
ChatGPT

ARC-AGI-2とAI汎用性評価：Johan Land氏の72.9%達成と多モデル反映的推論

ARC-AGI-2とは何か？汎用人工知能評価ベンチマークの概要

ARC-AGI-2（Abstraction and Reasoning Corpus for AGI 第2版）は、**汎用的な推論能力**を測るために設計された最新のベンチマークです [1]。もともとは2019年にFrancois Chollet氏が提唱した**ARC（抽象的推論コーパス）**に端を発し、少数の例から**新しいルールを学習して適用する能力**（流動性知能）をテストするものです [2] [3]。ARCの初代版（ARC-AGI-1）は2020年にKaggle競技会も開催されましたが、巨大言語モデル（LLM）が台頭した当初でも **純粋なLLMのスコアは0%** と極めて低く、2024年後半になるまでほとんど進展がありませんでした [4] [5]。ARC-AGI-2はその後継として2025年に公開され、**人間にとって容易だがAIにとって困難な課題**を新たに増やし、より詳細な評価指標を提供しています [6] [7]。

ARC-AGI-2のタスクは視覚パズルのような**グリッド変換問題**が中心で、各タスクでは少数の入出力例が提示され、それをもとに**未知のテスト入力に対する正しい出力を生成**する必要があります [2]。タスクは記号的な意味理解や複数ルールの同時適用、文脈に応じたルールの切り替えなど、人間には直感的に解けるが既存AIには難しい認知的チャレンジを含むよう設計されています [8] [9]。**「人間には易しくAIには難しい」**という設計原則のもと、各タスクは事前知識に依存しないよう工夫され、パターンの表面的な記憶ではなく**その場での推論と学習**が求められます [2] [3]。また、**ブルートフォース（総当り）では解けない**ように、解空間が大きすぎる問題や単純パターンは除外されています [10] [11]。



ARC-AGI-2のタスク例（「象徴的解釈」の課題）では、グリッド上のパターン（灰色の形）に対して、それらをただの図形ではなく**意味ある記号**とみなし特定の色で置き換えるルールを推論する必要がある [12]。このように**記号に意味を割り当てる抽象能力**は、人間にとって容易でも現行AIには難しく、ARC-AGI-2で強調される課題の一つである。

ARC-AGI-2では評価手法も強化され、人間とAIの性能差を公平に測るために**難易度のキャリブレーション**が行われています [13] [14]。まず、公開・非公開含め全ての評価タスク（計360問）は**少なくとも2人の人間が2回**

以内の試行で解けたことが事前検証されており [15] [16]、一般人400名以上による実証実験で平均正解率は**約60%**でした [17] [18]。一方で、従来の大規模モデルをそのまま適用した場合のAI正解率は**数%以下**に留まります [19]。例えば2025年時点での代表的なモデルはARC-AGI-2においてGPT-4.5で0%台、GPT-5.1で17.6%、GoogleのGemini 2.5で4.9%と散々な結果でした [20]。ARC-AGI-2はこの**人間とAIのギャップ**を明確化し、AIが汎用的な問題解決能力（いわゆるAGIに必要な流動知能）を獲得するにはスケール以外の新原理が不可欠であることを示す指標となっています [21] [22]。

さらに重要な点として、ARC-AGI-2は**効率（コスト）**の観点を初めて評価に組み込みました [23] [24]。単に「解けるか」だけでなく**どれだけ少ない試行や計算資源で解けるか**を Intelligence の一側面と捉え、AIと人間のコスト比較を行っています。人間は平均2.3分で1問を解き、報酬換算で1問あたり約17ドル程度の労力と見積もられています [25]。対してAIシステムは、初期の手法では**1問あたり数百ドル**分の計算資源（API費用やGPU時間など）を投じても人間に遠く及ばない状態でした [24]。ARC-AGI-2は**効率と正確さの両立**こそが真の知能の証であるという立場をとっており、リーダーボードでは各手法の正解率とともに**タスクあたりコスト**が公開されています [24] [26]。

Johan Land氏の最新成果：ARC-AGI-2での72.9%達成

2025年末、研究者の**Johan Land**氏がARC-AGI-2で驚異的な**72.9%の正解率**を記録し、AI分野の話題となりました [27]。これは当時の公開ベンチマーク上で**新たなSOTA（state-of-the-art）**となる快挙であり、OpenAIのGPT-5.2やGoogleのGemini 3 Proといった最新モデル単独の成績（約54%と31%）を大きく上回るものでした [28] [29]。特にGPT-5.2（Thinkingモード）のスコア52.9% [30] やGPT-5.2 Proの54.2% [30] を20ポイント近く上回っており、**平均的な人間参加者の60%前後** [18] さえ凌駕する水準です。実際、この72.9%はARC-AGI-2の**人間基準（平均53~60%程度）を突破**したとも評されました [31] [32]。Land氏自身もその後2026年1月初頭に**76.11%**までスコアを更新したことを報告しており、現時点で公表されている中では**最高性能**とされています [33]。

驚くべきは、この成果が**一つのモデルを改良して得られたのではなく**、Land氏が提唱する**「Multi-Model Reflective Reasoning（多モデル反映的推論）」**と呼ばれる**複合システム**によって達成された点です [34]。ARC-AGI-2のような難題に対し、Land氏は単一のLLMに頼るのではなく**複数の最先端AIモデルを協調させる**アプローチを取っています [34]。具体的には、OpenAIの**GPT-5.2**を基盤に、Google DeepMindの**Gemini-3**、Anthropicの**Claude Opus 4.5**といった**異なる強みを持つモデル**を組み合わせ、それぞれの出力を活用しながら解答を生成・精査しました [35]。Land氏の報告によれば、このシステムは**長時間にわたる多段階の推論**（1問あたり約6時間！）を行い、エージェント的に**Pythonコードを生成・実行**しながら試行錯誤することで解を探索しています [35]。実際に**10万回以上ものコード実行**を行い、画像的な推論（視覚パターンの分析）も取り入れ、最終的な回答候補を複数モデルによる**「審議会（council of judges）」**で評価して決定するという、手の込んだ戦略でした [35]。

Land氏の手法は一種の**メタAI**とも言えるものです。興味深いことに、このARC-AGI解答システム自体のコードは、人手ではなく**Gemini-3の対話型CLI**を用いてAIが自動生成したものだといいます [36]。つまり**AIが新たなAIソルバーを産み出した**とも言えるプロセスであり、Land氏は「これは以前のSOTAを打ち破るAIをAI自身が作り出したと言えるだろうか？」と示唆的なコメントを残しています [37]。

もっとも、この手法には**大きな計算資源とコスト**が伴います。ARC Prizeの発表によれば、Land氏のシステムは**ARC-AGI-2で1タスクあたり約39ドル**の計算コストを要した一方、ARC-AGI-1では約11ドルで94.5%の正解率を達成しています [27]。後述するようにARC-AGI-1は現行モデルにとってほぼ解ける水準まで来ているため、ARC-AGI-2では精度と引き換えにコストが大幅に増大している状況です。この**効率面の課題**もあり、コミュニティでは「この解法は知性というよりブルートフォース（力ずく）ではないか」との議論も起きています [38] [39]。しかしながら、**汎用AIの萌芽**として「未知の問題に対して自律的にプログラムを書き、試行錯誤で答えを見つけ出す」アプローチは大きな注目を集めています。ARC Prizeの公式も本手法を**「カスタムリ**

ファインメントによる提出 (bespoke refinement submission)」と位置づけ、高い評価と共にさらなる汎用性・効率性への課題を指摘しています [40] [41]。

次世代モデルの技術動向：GPT-5.2、Gemini-3、Claude 4.5の特徴と進化

Land氏の多モデル解法に用いられた各モデルは、それ自体が現行最高峰のAI技術を体現しています。それぞれの**技術的特徴**と近年の進化を概観します。

- **GPT-5.2（OpenAI）**： 2025年12月に公開されたOpenAIの最新フラッグシップLLMです [42]。GPT-5.2は「**プロフェッショナルな知的作業のための最先端モデル**」と位置づけられ [43]、従来モデルより**事実性や精度が向上し幻覚（誤回答）の減少**が謳われています [44]。最大の特徴は**推論能力の飛躍的強化**で、ChatGPTなどのエンタープライズ利用者から「コーディングや長文解析が格段に良くなった」と評価されています [45]。GPT-5.2は**マルチモーダル対応**（画像入力の理解）や**長大なコンテキスト**を扱う能力を持ち、**ツール使用やマルチステップのプロジェクト実行**にも優れるとされています [46] [47]。内部的には**Instantモード**（応答速度重視）と**Thinkingモード**（深い推論重視）の2形態があり、さらに高精度な**GPT-5.2 Pro**ではより長い思考時間・計算量を許容することで難問解決力を高めています [42] [48]。実際、ARC-AGI-2ベンチマークでもGPT-5.2 (Thinking)は**52.9%**をマークし、前バージョンGPT-5.1の17.6%から大幅なジャンプを示しました [30]。Pro版では**54.2%**まで到達し [30]、**チェイン・オブ・ソート（逐次思考）系モデルとして当時の新記録**を樹立しています。「より長く考えさせるほど難問に強くなる」傾向が確認されており、OpenAIはこのモデルで**連続的な論理展開やツール呼び出し**を組み合わせた**エージェント的タスク**に本格対応したとしています [46] [47]。

- **Gemini-3（Google DeepMind）**： 2025年11月にリリースされたGoogleの次世代モデルで、**マルチモーダルとエージェント能力の融合**が特徴です [49] [50]。Geminiシリーズは2023年末のGemini 1で高精度な**ネイティブ多モーダル対応**を実現し、Gemini 2で**エージェント的振る舞い**と高度な推論力を拡張、そして**Gemini 3でそれらを統合**した「最も知的なモデル」と位置づけられています [51] [52]。Gemini-3 Proは**テキスト・画像・音声・動画・コード**までシームレスに扱い、ユーザーの意図を深く汲み取って**プランニングやツール操作**ができるとされています [53] [54]。各種ベンチマークでも当時最高性能を記録し、例えば学術推論テストの「Humanity's Last Exam」ではツール未使用で37.5%（GPT-5.1は26.5%） [55]、マルチモーダル理解指標のMMMUでは81%と突出したスコアを出しました [56]。ARC-AGI-2に関して言えば、Gemini-3 Proの標準モードで**31.1%**（GPT-5.1は17.6%）を達成し [20]、さらに**Deep Thinkモード**（推論拡張モード＋コード実行あり）では**45.1%**に達したと報告されています [57] [58]。これはGPT-5.2登場までは最高クラスの成績であり、Googleは「Gemini 3は複雑な問題を解決する能力で新たな水準を打ち立てた」と述べています [56] [59]。GeminiはGoogleのサービス群（検索AIモードやクラウドVertexへの組み込み等）にも即日統合され、**大規模プロダクトでのデプロイ**を念頭に進化が続けられています [60] [61]。

- **Claude 4.5（Anthropic）**：Anthropic社のClaudeシリーズも2025年に大きく刷新され、特に**Claude "Opus" 4.5**は長時間のコーディングやエージェントタスクにおける飛躍が報告されています。Opus 4.5は従来のClaude 2やClaude "Sonnet" 4.5からのアップグレード版で、**曖昧さへの対処や自律的な問題解決**能力が著しく向上したとテスターらが評価しています [62] [63]。例えば複雑なバグの原因究明や、大規模なコードベースのリファクタリングといった長い思考の連続を要する作業で「Opus 4.5は途中で行き詰まらずに全体像を捉えて解決策を見出す」とされ [62] [64]、**長大な対話や30分以上に及ぶ自律コード生成**でも安定した性能を示すといいます [65] [66]。特に**コーディング能力とエージェント的ワークフロー**に優れ、GitHub Copilot等との統合下で既存モデルを上回るパス率・効率性を発揮したとの報告があります [67] [68]。実際、Opus 4.5は社内の難解コード評価で前版Sonnet 4.5比で15%改善し、使用トークンも半減するなど**効率と品質の両立**を達成したとのことです [65] [68]。ARC-AGI-2において公式発表された単独スコアは不明確ですが、コミュニティ推計では**約38%**前後に達し

ているとの分析があります [69]。Anthropic社はOpus 4.5を「フロンティアの推論力を備えたモデル」と位置づけ [70]、複数ステップの推論・プランニングや自己改善的なエージェント挙動において新たな地平を開いたと評しています [71] [72]。

以上のように、GPT-5.2・Gemini-3・Claude 4.5はいずれも**汎用AIへの里程標**ともいうべき高度な機能を備えており、それぞれ得意領域が若干異なります。GPT-5.2は**論理推論と対話統合**、Gemini-3は**多モーダル統合とプランニング**、Claude 4.5は**コード生成や長時間エージェント**といった強みを持ち、今回Land氏はそれらを**相補的に活用することで性能を極限まで引き出した**わけです [35]。

Multi-Model Reflective Reasoningの仕組みと意義

Multi-Model Reflective Reasoning（多モデル反映的推論）とは、文字通り**複数のAIモデルを組み合わせ、反復的な推論と自己改善を行う**手法です。Land氏のARC-AGI-2ソルバーはまさにこの理念を体現しており、その動作は以下のように整理できます。

1. **モデル間役割分担と協調**: GPT-5.2、Gemini-3、Claude 4.5といったモデルにそれぞれ**役割を割り振り**、協調させています [34]。例えば、GPT-5.2は与えられた課題の高レベルな分析・思考（チェイン・オブ・ソートの骨子作成）に用い、Claude 4.5は具体的な**コード（プログラム）生成とデバッグ**を担当し、Gemini-3は**画像パターンの認識**や対話的プロンプト作成に寄与する、といった具合です。これにより**各モデルの強みを相乗的に活かす**ことが可能になります。

2. **反復的推論（Reflective Reasoning）**: 解法は一度の推論で決まるのではなく、モデルが出力した仮説やプログラムを**逐次評価しフィードバック**するループを回します [35]。具体的には、生成したPythonコードを実行してタスクの入出力例に適用し、**結果が期待に合致しなければ原因を分析**します。そのフィードバックを元にモデルへ再プロンプトし、コードや推論を修正・改善させるというサイクルを繰り返します。まさに**試行→評価→改善**の**自己反省的プロセス**であり、人間がパズルを解く際に「ああでもないこうでもない」と考え試す様子に近い手順です。

3. **大規模探索と長期的思考**: 上記の反復を**膨大なスケール**で行う点も特徴的です。Land氏の報告では、1問あたり平均して**6時間**程度もかけ、多数の仮説コードを生成・検証したとされています [35]。通常のモデル推論では数秒〜数分で得られる解答を、人間のように**何時間も粘り強く考えさせる**ことで、浅いパターンマッチでは見つからない解決策に到達できるわけです。この長時間推論を実現するため、外部ツール（Python実行環境や記憶メモリ）を駆使する**AIエージェント的手法**が用いられています [35]。100,000回以上のコード実行はその極端な例ですが、計算機に**疑似的な忍耐力と探索能力**を与えることで、極めて難度の高い問題にも解を発見できることを示しました。

4. **評議会（Council of Judges）による検証**: 最終解答の決定に際しては、**複数モデルが合議**する仕組みも導入されています [35]。例えば複数の候補プログラム解答を用意し、それぞれを別のモデル（GPT-5.2やClaude等）が**審査員**として評価・スコア付けすることで、信頼度の高い解答を選び出すアプローチです。これにより一つのモデルのバイアスや誤謬に依存せず、**合議制による安定解**を得ることができます。いわば**多数決型の自己検証**であり、AI同士が相互にチェックし合うことで解答精度を高めています。

以上のステップにより、Land氏のシステムは**極めて高い正解率**を叩き出しました。しかし同時に指摘されるのは、その**汎用性と効率性**です。現在のところ、この解法はARC-AGI-2という特定ベンチマークで威力を発揮しましたが、他分野の問題へどこまで一般化できるかは未知数です。また人間が2分で解く問題に6時間・数万回の試行で挑むのは、知能というより計算力に物を言わせた「解空間の探索」に見えるかもしれません [73] [39]。ARC Prize運営も「リファインメント（試行適応）の一般化可能性」と「プログラム生成にかかる高コスト」が課題であると述べています [41]。一方で、この手法は**現在のモデル群の持つポテンシャルを最大限引き出す**ものであり、裏を返せば「適切なメタ推論アルゴリズム」を与えれば既存モデルを組み合わせて**人間**

に迫る知的能力を発揮し得ることを示しました [37]。これは今後、より効率の良い自己反省アルゴリズムや軽量なメタ制御モデルを開発することで、汎用AIに一歩近づける可能性を示唆しています。

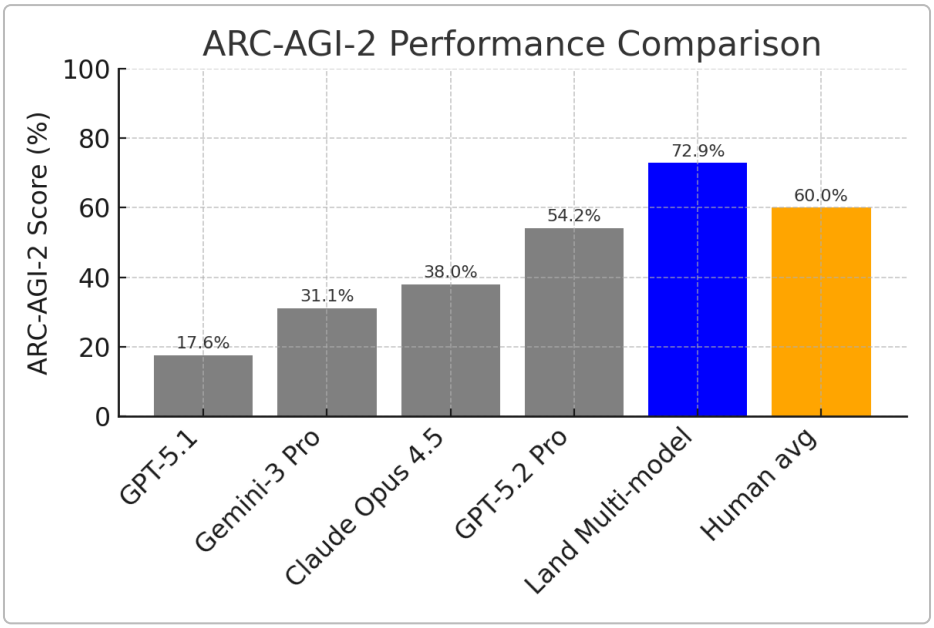


図: ARC-AGI-2における主要モデルの性能比較（正解率）。OpenAIのGPT-5.2登場以前、トップクラスのモデル（GPT-5.1、Gemini-3 Pro、Claude 4.5など）のスコアは**50%以下**であったが、GPT-5.2 (Pro) は約**54%**を達成し [28]、Johan Land氏のMulti-Model Reflective Reasoningシステムは**72.9%** [27] とさらに大幅に上回った。人間参加者の平均は約**60%** [18] と報告されており、Land氏の結果は**人間平均をも凌駕**している。

おわりに

ARC-AGI-2は現在、**「真の汎用人工知能」に近づくために克服すべき壁**を端的に示すベンチマークとなっています [74]。単に知識を詰め込んだモデルでは太刀打ちできず、**未知の問題への適応力**こそが評価されるこのテストで、AI研究者たちは様々な創意工夫を凝らしています。Johan Land氏の成果は、巨額の賞金（ARC Prizeでは85%正解で100万ドルの賞が掲げられています [75] [76]）を狙ったものでもありますが、それ以上に**新しいAIアーキテクチャの可能性**を示した点で学術的にも価値が高いと言えます。今後は、2026年に予定されているARC-AGI-3 [77] や他の汎用知能ベンチマークにおいて、どこまでこの多モデル・自己反映アプローチが通用し、あるいは新たな手法が登場するのか注目されます。人間と同等、あるいは超える汎用AIの実現に向け、**ベンチマーク駆動の研究**はこれからも加速していくでしょう。

参考文献・情報源：（論文・記事・公式発表より） - ARC Prize公式: ARC-AGI-2 技術報告 [6] [78] [19] - ARC Prize公式ブログ: ARC-AGI-2 発表記事 [4] [13] [17] - BracAIテックブログ: ARC-AGI-2ベンチマーク解説 [74] [24] - OpenAI: GPT-5.2 発表ページ [46] [30] - Google DeepMind: Gemini 3 発表ブログ [56] [79] - Anthropic: Claude Opus 4.5 発表記事 [68] - Johan Land氏 LinkedIn投稿 [34] - Reddit (r/singularity) SOTA達成報告スレッド [33] [37]

---

[1] [2] [3] [24] [25] [28] [29] [69] [74] Best reasoning LLM in 2026: ARC-AGI-2 benchmark leaderboard
https://www.bracai.eu/post/arc-agi-2-benchmark

[4] [5] [13] [14] [15] [17] [18] [21] [22] [23] ARC-AGI-2
https://arcprize.org/arc-agi/2/

6 7 8 9 10 11 12 16 19 75 76 78 ARC-AGI-2 A New Challenge for Frontier AI Reasoning Systems
https://arcprize.org/blog/arc-agi-2-technical-report

20 53 54 55 Gemini 3 Pro — Google DeepMind
https://deepmind.google/models/gemini/pro/

26 ARC Prize - Leaderboard
https://arcprize.org/leaderboard

27 33 37 38 39 40 41 73 77 New SOTA achieved on ARC-AGI : r/singularity
https://www.reddit.com/r/singularity/comments/1quzgg5/new_sota_achieved_on_arcagi/

30 43 46 47 Introducing GPT-5.2 | OpenAI
https://openai.com/index/introducing-gpt-5-2/

31 32 ARC-AGI-2 human baseline surpassed (updated) — LessWrong
https://www.lesswrong.com/posts/DX3EmhmwZjTYp9PBf/arc-agi-2-human-baseline-surpassed-updated

34 35 36 The holidays were filled with coding this year, and I spent time experimenting with ARC-AGI. The result surprised me: 76% on ARC-AGI 2, beating public GPT-5.2 and Gemini 3 Pro baselines by >20%, and… | Johan Land
https://www.linkedin.com/posts/johanland_the-holidays-were-filled-with-coding-this-activity-7413966228060508160-1omU

42 45 48 GPT-5.2 - Wikipedia
https://en.wikipedia.org/wiki/GPT-5.2

44 GPT-5.2 is OpenAI's latest move in the agentic AI battle | The Verge
https://www.theverge.com/ai-artificial-intelligence/842529/openai-gpt-5-2-new-model-chatgpt

49 50 51 52 56 57 58 59 60 61 79 Gemini 3: Introducing the latest Gemini AI model from Google
https://blog.google/products-and-platforms/products/gemini/gemini-3/

62 63 64 65 66 67 68 70 71 72 Introducing Claude Opus 4.5 \ Anthropic
https://www.anthropic.com/news/claude-opus-4-5