汎用人工知能への飛躍：Johan Land氏によるARC-AGI-2記録とMulti-Model Reflective Reasoningの包括的分析

Gemini 3 pro

1. 序論：AI推論能力のパラダイムシフト

2025年末から2026年初頭にかけて、人工知能（AI）研究の最前線において、汎用人工知能（AGI）の実現に向けた極めて重要なマイルストーンが達成された。AI研究者でありエンジニアであるJohan Land氏は、Francois Chollet氏らによって設計された、AIの「流動性知能（Fluid Intelligence）」と「一般化能力」を測定する最も過酷なベンチマークである**ARC-AGI-2（Abstraction and Reasoning Corpus for AGI - Version 2）**において、**72.9%**という驚異的な正答率を記録した [1]。

この記録は、直前のState-of-the-Art（SOTA）であったPoetiqシステムによるGPT-5.2を用いたスコア（約54%）を大幅に上回るものであり、AIが特定領域のスキルだけでなく、未知の問題に対する適応能力において人間レベル（平均的な人間は約60-66%とされる）を凌駕し始めたことを示唆している [3]。Land氏のアプローチの中核にあるのは、**「Multi-Model Reflective Reasoning（マルチモデル再帰的推論）」**と名付けられた手法であり、これは単一の巨大言語モデル（LLM）に依存するのではなく、OpenAIのGPT-5.2、GoogleのGemini 3、AnthropicのClaude Opus 4.5といった現行最強のフロンティアモデル群を動的に連携させる高度なアンサンブルシステムである [2]。

本レポートでは、Land氏の達成した記録の技術的詳細、採用されたアーキテクチャの特異性、そしてこの成果が示唆する「推論時計算（Inference-time Compute）」の重要性とコストのジレンマについて、包括的かつ詳細に分析を行う。

2. ARC-AGI-2：AIにとっての「最終障壁」

2.1 ベンチマークの哲学と進化

ARC-AGIは、2019年にKerasの開発者であるFrançois Chollet氏が発表した論文「On the Measure of Intelligence」に基づいて構築されたベンチマークである。Chollet氏は、特定のタスク（チェスや画像認識など）における「スキル」と、未知の環境に適応し新たなスキルを獲得する能力である「知能」を明確に区別した [5]。ARCは、膨大な訓練データの暗記（暗記による近似）を防ぎ、人間が先天的に持つ「コア知識（Core Knowledge Priors）」—物体認識、数概念、幾何学、トポロジー、エージェント性など—のみを前提とした純粋な推論能力を測定することを目的としている [5]。

2025年にリリースされた**ARC-AGI-2**は、初代ARCが当時のLLMや総当たり的なプログラム合成手法によって部分的に攻略されつつあったことを受け、その難易度と堅牢性を劇的に向上させたバージョンである。ARC-AGI-2は以下の点で強化されている：

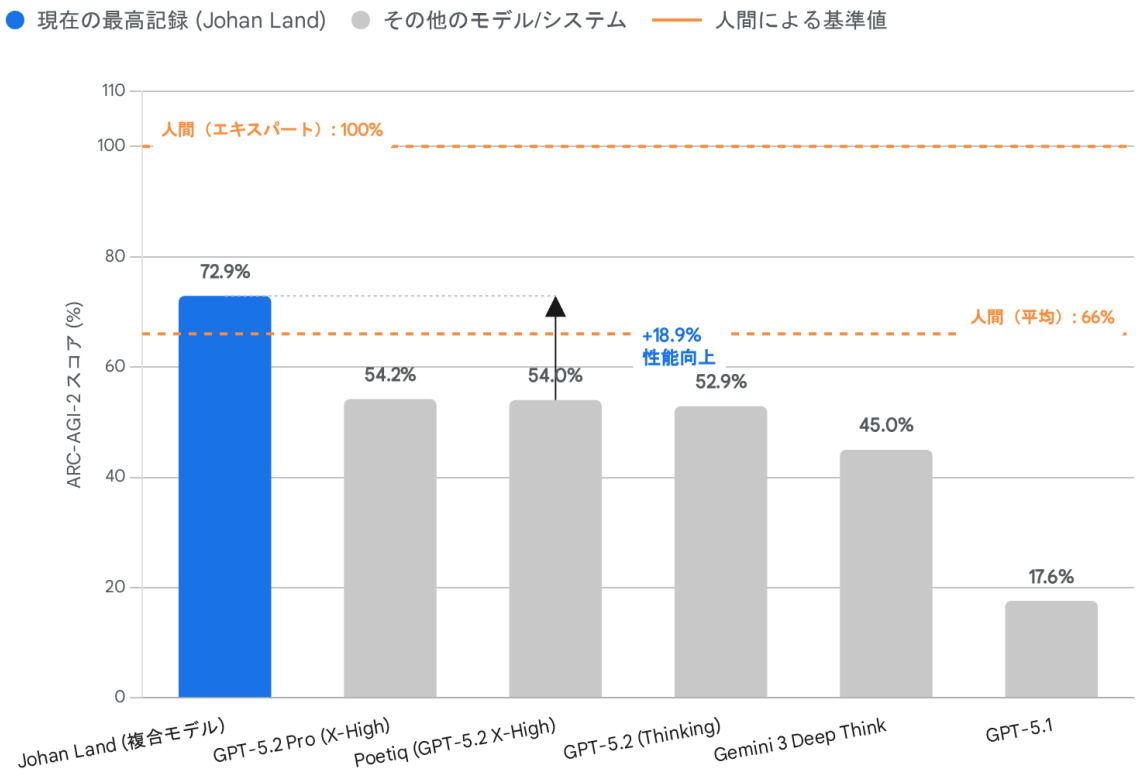
1. 記号的解釈（**Symbolic Interpretation**）：記号が持つ視覚的パターン以上の意味（意味論的意義）を文脈から解釈する必要があるタスクの増加。AIは対称性や変形といった表面的なパターンには強いが、記号自体に割り当てられた意味的役割を理解するのに苦労する [8]。
2. 構成的推論（**Compositional Reasoning**）：複数のルールを同時に、あるいは文脈に応じて順序立てて適用する能力が要求される。単一のルール適用ではなく、ルールの相互作用を理解する必要がある [8]。
3. 文脈的ルール適用（**Contextual Rule Application**）：周囲の状況に応じて適用すべきルールが変化するタスク。これにより、単純なパターンマッチングを無効化する [8]。

2.2 人間とAIのギャップ

ARC-AGI-2の特筆すべき点は、これらのタスクが「人間にとっては依然として容易である」ように設計されていることである。400人以上の人間による検証の結果、すべてのタスクが「2人以上の人間によって2回以下の試行で解決可能」であることが確認されている [3]。人間の平均的な正答率は60%〜66%（実験設定に依存）であり、専門家パネルや時間をかけた検証では100%に達する [3]。

対照的に、2025年後半時点での最先端モデル（GPT-4oやGemini 1.5 Proなど）のベースラインスコアは一桁台（<5%）に留まっていた [9]。これは、大規模言語モデルが依然として「記憶と補間」に強く依存しており、真の意味での「未知への適応（Extrapolation）」を苦手としていることを浮き彫りにしていた。この背景において、Johan Land氏が達成した72.9%という数字は、AIがこの「一般化の壁」を突破しつつあることを示す歴史的なデータポイントである。

ARC-AGI-2 ベンチマークにおけるAIモデル性能の進化と人間のベースライン



Johan Land氏のシステムは、GPT-5.2単体や従来のSOTAであるPoetiqシステムを大きく上回り、人間の平均的なパフォーマンス（約60-66%）を超過する72.9%を記録した。これは、単なるモデルの大規模化ではなく、推論プロセスの高度化（マルチモデル連携）が壁を突破したことを示している。

Data sources: Reddit (LocalLLaMA), LessWrong, Johan Land (via Reddit), Poetiq, Intuition Labs

3. 技術的深層：Multi-Model Reflective Reasoningの解剖

Land氏のシステムは、単一の「天才的」なモデルに依存するのではなく、異なる特性を持つ複数のモデルを高度に組織化し、反復的な検証プロセスを経ることで解に到達する。このアプローチは「Multi-Model Reflective Reasoning（マルチモデル再帰的推論）」と呼ばれ、以下の主要なコンポーネントとプロセスから構成されている。

3.1 モデル・アンサンブルの戦略的配置

システムは、各タスクに対して最適な認知能力を提供するために、以下のフロンティアモデル群を適材適所で起用している [2]。

- **GPT-5.2 (OpenAI):** システムの論理的バックボーンとして機能する。特に「Thinking」モードや

APIの「X-High」推論レベルを活用することで、複雑な文脈保持とマルチステップの論理構築を担当する。GPT-5.2はARC-AGI-2単体でも52.9%を記録する強力なベースライン能力を持っており、全体のオーケストレーションを担う[4]。

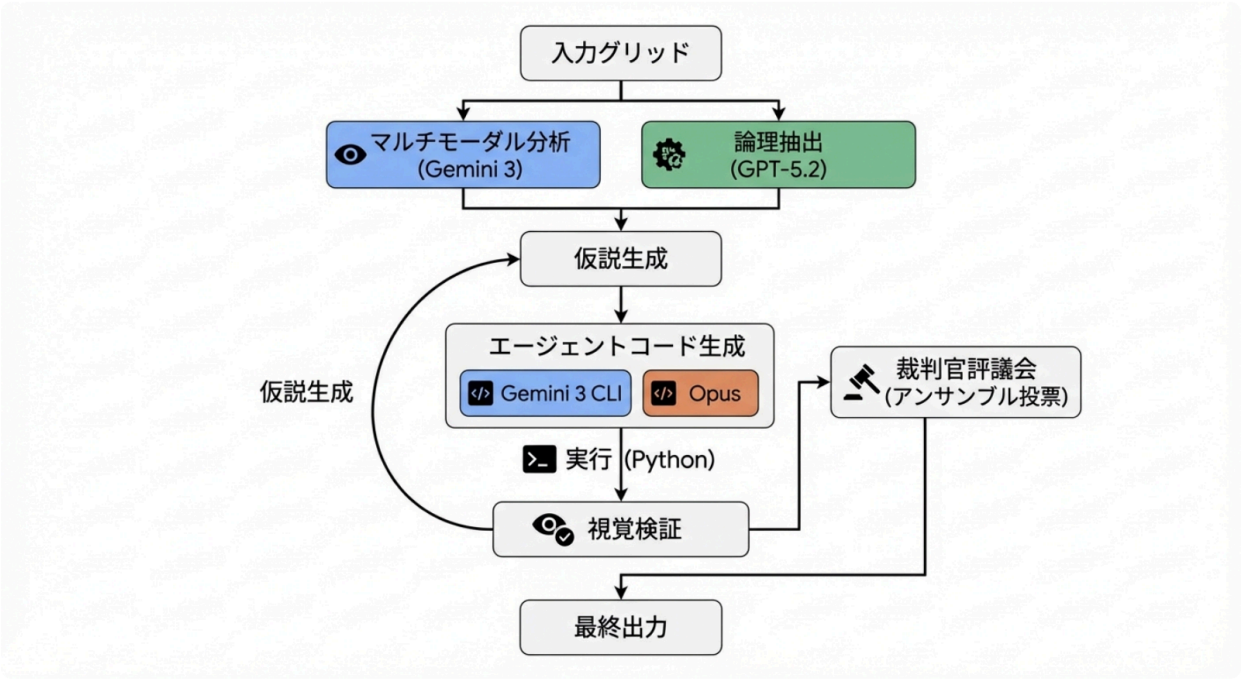
- **Gemini 3 (Google):** ネイティブ・マルチモーダルアーキテクチャを持つGemini 3は、ARCのような視覚的パズルにおいて極めて重要な役割を果たす。テキスト化されたグリッド情報だけでなく、視覚情報を直接処理することで、人間が「直感」と呼ぶようなパターン認識（対称性、包含関係、接続性の即時把握など）を提供する。また、Land氏の報告によれば、ソルバー自体のコード生成（メタプログラミング）においてもGemini-3-CLIが活用されており、AIがAIの思考ツールを作成するという自己言及的な側面も持つ[2]。
- **Claude Opus 4.5 (Anthropic):** 長文脈における安定性と高いコーディング能力により、生成された仮説の検証や、他のモデルが見落としたエッジケースの指摘、あるいは異なる視点からの解法提案（Lateral Thinking）に寄与している[12]。

3.2 6時間の熟考：推論プロセスの外部化

Land氏のシステムは、1つの問題を解くために平均して約**6**時間もの計算時間を費やす[2]。これは、入力に対して即座に確率的なトークンを出力する従来のLLMの挙動とは決定的に異なる。この6時間は、人間の脳における「システム2（熟慮的思考）」を、外部ツールとマルチエージェントループによって明示的に実装したプロセスである。

1. 多角的な仮説生成（**Hypothesis Generation**）：入力グリッドと出力グリッドのペアから、変換ルールに関する多数の仮説を生成する。ここではGeminiの視覚的直感とGPT-5.2の言語的論理が並行して稼働し、多様な角度からの「解法候補」を洗い出す[14]。
2. エージェンティック・コード生成（**Agentic Codegen**）：生成された仮説を検証するためのPythonコードを記述する。Land氏のシステムでは、1つのタスクにつき**10**万回以上の**Python**呼び出しが行われるとされる[2]。AIは「頭の中」だけで考えるのではなく、実際にプログラムを書き、実行し、その結果（生成されたグリッド）を見てコードを修正するという「試行錯誤」のループを超高速で回転させる。これにより、モデルのハルシネーション（幻覚）を排除し、実行可能な真実のみを抽出する。
3. 視覚的検証（**Visual Reasoning**）：コード実行によって生成された出力グリッドを、再度視覚モデルが評価し、期待される出力と視覚的に一致しているか、あるいは新たなパターンが出現していないかを確認する。

Multi-Model Reflective Reasoning システムアーキテクチャ概要



Johan Land氏のシステムは、GPT-5.2、Gemini 3、Claude Opus 4.5を連携させ、仮説生成からコード実行、検証、そして「裁判官評議会（Council of Judges）」による最終決定に至るまで、人間のような熟考プロセスを模倣している。1問あたり数千回の反復と検証が行われる。

3.3 Council of Judges（裁判官評議会）：合議による品質保証

Land氏のシステムのもう一つの重要な特徴は、「Council of Judges（裁判官評議会）」と呼ばれるメカニズムの採用である[2]。これは、複数のエージェントが生成した解法や推論プロセスに対し、別のエージェント群が「審査員」として振る舞い、その妥当性を評価・投票するシステムである。

- ハルシネーションの抑制: 単一のモデルが自信満々に誤った推論（もっともらしい嘘）を出力した場合でも、異なるアーキテクチャやプロンプトを持つ他のモデルがそれを批判的に検証することで、誤謬を検出する確率が高まる。これは「LLM Council」や「Debate with Judge」といったマルチエージェントデザインパターンの一種であり、合意形成を通じて出力の信頼性を担保する[16]。
- 多様性の確保: 異なるモデル（審査員）は異なるバイアスを持つため、全会一致で採択された解法は、特定のモデルの癖に依存しない普遍的なロジックに基づいている可能性が高い。

3.4 AIによるAIの開発

Land氏が公開した興味深い事実として、「すべてのソルバーコードはGemini-3-CLIによって記述された」という点がある[2]。これは、ARC-AGI-2を解くためのプログラム自体を人間が手書きしたのではなく、AI（Gemini 3）がコーディングエージェントとして機能し、問題を解くためのシステムを構築したこ

とを意味する。これは「AIがより優れたAI（あるいは問題解決システム）を生み出す」という再帰的な自己改善（Recursive Self-Improvement）の初期的な事例とも解釈でき、SOTA達成のプロセス自体がAGI的な性質を帯びている。

4. コストと効率性のトレードオフ：パレートフロンティアの移動

Johan Land氏のアプローチは性能において圧倒的であるが、その代償として計算コストとエネルギー効率において大きな課題を突きつけている。ARC-AGI-2の運営チームは、スコアだけでなく「効率性」も重要な指標として追跡している [3]。

以下の表は、ARC-AGI-2における主要なアプローチの精度とコストの比較を示したものである。

システム / モデル	**ARC-AGI-2** スコア	推定コスト **(per task)**	備考
Johan Land (v2)	**72.9%**	**$38.9**	Multi-Model, 6時間の推論, 10万回のコード実行
Poetiq (GPT-5.2 X-High)	54.0%	$8.0 - $30.0	メタシステムによる最適化
GPT-5.2 Pro (Official)	54.2%	~$15.27	OpenAI公式ベンチマーク (X-High)
GPT-5.2 Thinking	52.9%	~$1.90	標準的なThinkingモード
Gemini 3 Deep Think	45.0%	不明 (高コスト)	GoogleのDeep Think技術
NVIDIA (NVARC)	27.6%	**$0.20**	Kaggleチームによる高効率アプローチ
人間 **(平均)**	~66.0%	~$5.0 + α	報酬ベース。エネルギー効率は極めて高い

データソース: [2]

このデータは、現在のAIが「精度」を購入するために「計算リソース」を指数関数的に投入している現状を浮き彫りにしている。Land氏のシステムは、NVIDIAの軽量なアプローチと比較して約200倍のコストをかけている。この「富豪的推論（Inference Scaling）」のアプローチは、コスト度外視で正解を探索するものであり、Chollet氏が定義する「知能とはスキル獲得の効率性である」という理想からは乖離しているように見えるかもしれない [18]。

しかし、別の見方をすれば、Land氏の成果は**「計算リソースさえ十分に投入すれば、AIは未知の抽象推論タスクにおいて人間を超えられる」**という事実を実証した点で革命的である。これまでは、どんなに計算量を増やしてもARCのような新規性の高いタスクは解けない（汎化できない）と考えられていたが、推論時計算のスケーリング則（Inference Scaling Laws）が有効であることが証明されたのである [19]。

5. GPT-5.2：推論特化型モデルの到達点

Johan Land氏のシステムの成功は、その基盤となるOpenAIの**GPT-5.2**の性能向上なくしては語れない。2025年12月にリリースされたGPT-5.2は、前世代のGPT-5.1と比較して、推論能力において質的な飛躍を遂げている。

- **ARC-AGIスコアの激増:** GPT-5.1のARC-AGI-2スコアがわずか17.6%であったのに対し、GPT-5.2（Thinking）は52.9%を記録した。約3倍のスコア向上は、通常のモデル更新では考えられない幅であり、モデルが「推論の手順」をより深く理解し始めたことを示している [4]。
- **Thinking Modeの実装:** GPT-5.2には、ユーザーが推論の深さ（Effort Level）を調整できる機能が搭載されている。API経由で利用可能な「X-High」設定では、モデルは回答を出力する前に内部で膨大な「思考トークン」を消費し、自己検証と論理の組み立てを行う。Land氏のシステムは、この機能を外部ループでさらに拡張したものと言える [4]。
- **エージェント機能の強化:** SWE-bench Verifiedにおいて80.0%を記録するなど、コーディングとツール利用の能力が大幅に向上した。これにより、Land氏のシステム内での「Pythonコードを書いて検証する」というプロセスの成功率が担保された [11]。

6. 結論と展望：AGIへの道程

Johan Land氏によるARC-AGI-2での72.9%達成は、AI研究における一つの転換点である。それは以下の3つの重要な示唆を我々に与えている。

1. 「システム」としての**AI**の勝利: 最先端のAI性能は、単一の巨大モデルではなく、複数のモデルとツール（Python実行環境など）を高度に連携させた「コンパウンドAIシステム（Compound AI Systems）」によって切り拓かれる。単体では不完全なモデルも、相互監視とツール利用を通じて「超人的」な成果を出せる。
2. 推論スケーリングの威力: 事前学習（Pre-training）の規模拡大だけでなく、推論時（Inference-time）に計算リソースを大量投入し、数千〜数万回の試行錯誤を行うことで、AIは自身の限界を超えた難問を解決できる。これは「System 1（直感）」から「System 2（熟考）」への移行を意味する。
3. 効率性の課題: AIは「人間レベルの正答率」を達成したが、「人間レベルの効率性」には程遠い。1問に$40近いコストと電力を使用する現状は、実用化の観点からは課題が残る。今後の

競争は、この高いスコアを維持しつつ、いかに計算コスト（推論時の探索空間）を圧縮できるかという「蒸留（Distillation）」や「効率化」のフェーズに移るだろう。

Francois Chollet氏が「AIには解けない」として設計したARC-AGI-2の壁が、予想よりも早く、しかも「力技と知恵の融合」によって破られつつある事実は、AGIの到来が遠い未来の話ではないことを予感させる。Land氏のアプローチは、科学的発見や複雑なエンジニアリングなど、正解のない難問に挑むAIの未来のひな型となるだろう。

引用文献

1. Open-source just beat humans at ARC-AGI (71.6%) for $0.02 per task, 2月 4, 2026にアクセス、https://www.reddit.com/r/LocalLLaMA/comments/1p7d97m/opensource_just_beat_humans_at_arcagi_716_for_002/
2. New SOTA achieved on ARC-AGI : r/singularity - Reddit, 2月 4, 2026にアクセス、https://www.reddit.com/r/singularity/comments/1quzgg5/new_sota_achieved_on_arcagi/
3. ARC-AGI-2 human baseline surpassed (updated) - LessWrong, 2月 4, 2026にアクセス、https://www.lesswrong.com/posts/DX3EmhmwZjTYp9PBf/arc-agi-2-human-baseline-surpassed-updated
4. GPT-5.2 & ARC-AGI-2: A Benchmark Analysis of AI Reasoning, 2月 4, 2026にアクセス、https://intuitionlabs.ai/articles/gpt-5-2-arc-agi-2-benchmark
5. ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems, 2月 4, 2026にアクセス、https://www.arxiv.org/pdf/2505.11831
6. What is ARC-AGI? - ARC Prize, 2月 4, 2026にアクセス、https://arcprize.org/arc-agi
7. The ARC Prize: Efficiency, Intuition, and AGI, with Mike Knoop, co ..., 2月 4, 2026にアクセス、https://www.cognitiverevolution.ai/the-arc-prize-efficiency-intuition-and-agi-with-mike-knoop-co-founder-of-zapier/
8. ARC-AGI-2, 2月 4, 2026にアクセス、https://arcprize.org/arc-agi/2/
9. ARC-AGI-2 A New Challenge for Frontier AI Reasoning Systems, 2月 4, 2026にアクセス、https://arcprize.org/blog/arc-agi-2-technical-report
10. On January 1, 2030, there will be no AGI (and AGI will still not be ..., 2月 4, 2026にアクセス、https://forum.effectivealtruism.org/posts/xkNjpGNfnAYmkFz3s/on-january-1-2030-there-will-be-no-agi-and-agi-will-still
11. Introducing GPT-5.2 - OpenAI, 2月 4, 2026にアクセス、https://openai.com/index/introducing-gpt-5-2/
12. GPT-5.2 Benchmarks (Explained) - Vellum AI, 2月 4, 2026にアクセス、https://www.vellum.ai/blog/gpt-5-2-benchmarks
13. Company: anthropic | AINews, 2月 4, 2026にアクセス、https://news.smol.ai/tags/anthropic/
14. The real bottleneck of ARC-AGI : r/ArtificialInteligence - Reddit, 2月 4, 2026にアクセス、

https://www.reddit.com/r/ArtificialInteligence/comments/1jluzxw/the_real_bottleneck_of_arcagi/
15. Multimodal Reasoning to Solve the ARC-AGI Challenge, 2月 4, 2026にアクセス、https://omseeth.github.io/blog/2025/MLLM_for_ARC/
16. Multi-Agent Architectures - Swarms, 2月 4, 2026にアクセス、https://docs.swarms.world/en/latest/swarms/concept/swarm_architectures/
17. LLM Council - Swarms, 2月 4, 2026にアクセス、https://docs.swarms.world/en/latest/swarms/structs/llm_council/
18. Zapier's Mike Knoop launches ARC Prize to Jumpstart New Ideas for ..., 2月 4, 2026にアクセス、https://sequoiacap.com/podcast/training-data-mike-knoop/
19. Implications of the inference scaling paradigm for AI safety, 2月 4, 2026にアクセス、https://www.lesswrong.com/posts/HiTjDZyWdLEGCDzqu/implications-of-the-inference-scaling-paradigm-for-ai-safety
20. How to Beat ARC-AGI by Combining Deep Learning and Program ..., 2月 4, 2026にアクセス、https://arcprize.org/blog/beat-arc-agi-deep-learning-and-program-synthesis
21. OpenAI's Answer to Gemini 3, Runway's Interactive Worlds, Disney's ..., 2月 4, 2026にアクセス、https://www.deeplearning.ai/the-batch/issue-332/