Poetiq AI による ARC-AGI-2 における GPT-5.2 X-High を用いた 75%精度達成に関する包括的分析レポート

Gemini 3 pro

1. 序論：汎用人工知能（AGI）への道程と ARC ベンチマークの重要性

1.1背景：AI 開発における新たなマイルストーン

2025 年 12 月、AI スタートアップである Poetiq AI が発表した成果は、人工知能研究コミュニティに大きな衝撃を与えた。同社は、OpenAI の最新モデルである GPT-5.2 の「X-High」推論設定と独自のメタシステムを組み合わせることにより、ARC-AGI-2（Abstraction and Reasoning Corpus for Artificial General Intelligence v2 ）ベンチマークの公開評価セット（Public Evaluation Set）において 75%という驚異的な正答率を達成したと主張している [1]。この数値が持つ意味は極めて重い。なぜなら、ARC-AGI-2 は「人間には容易だが、現在の AI には極めて困難」なタスクとして設計されており、これまでの最先端モデル（State-of-the-Art, SOTA）であっても、検証されたスコアとしては 50%台半ばが限界とされていたからである [2]。

しかし、この発表を額面通りに受け取ることは危険である。Poetiq AI は過去に、Gemini 3 を用いたシステムにおいて、公開セットで 65.32%という高いスコアを記録しながら、公式な検証環境（Semi-Private Evaluation Set）では 54.0%へと大幅にスコアを落とした経緯がある [2]。この約 11 ポイントの「スコア乖離」は、AI モデルの評価における過学習（Overfitting ）やデータ汚染（Data Contamination ）という、現代 AI が抱える根本的な問題を浮き彫りにしている。

本レポートは、Poetiq AI の 75%達成という主張の技術的背景、その実現可能性、そして「真の実力」を測るための分析を行うものである。特に、ARC-AGI-2 というベンチマークの特異性、Poetiq 独自の「再帰的自己改善」アーキテクチャ、そして GPT-5.2 という基盤モデルの性能とコスト構造を詳細に解剖し、今後の公式評価でどのような結果が予測されるかを論じる。

1.2 レポートの構成

本稿は以下の構成で論を展開する。まず第 2 章では、ARC-AGI-2 ベンチマークの設計思想とその難易度の本質について、認知科学的観点から詳述する。第 3 章では、Poetiq AI が採用し

ている技術的アプローチ、特に「メタシステム」と「再帰的自己改善」のメカニズムを解説する。第 4 章では、GPT-5.2 X-High モデルの特性と、それが Poetiq のシステムに与える影響を分析する。第 5 章では、本レポートの核心である「スコア乖離」のメカニズムを統計的および構造的観点から検証し、75%という数値の信頼性を評価する。第 6 章では、コストパフォーマンスの観点から実用性を議論し、最後に第 7 章で今後の公式評価におけるシナリオ予測と結論を述べる。

---

2. ARC-AGI-2 ベンチマーク：流動性知能の厳格なる試金石

2.1 François Chollet による知能の定義と ARC の起源

ARC-AGI ベンチマークは、Keras の開発者であり著名な AI 研究者である François Chollet によって 2019 年に提唱された。Chollet は、当時の AI 研究が「特定のタスクにおけるスキル」の向上に偏重していることを批判し、真の知能とは「スキルそのものではなく、未知の環境に適応し、新たなスキルを獲得する能力（Skill Acquisition Efficiency）」であると定義した[5]。

従来のベンチマーク（例えば MMLU や GSM8K）は、膨大なテキストデータに含まれる知識を想起したり、既存のパターンを適用したりすることで高スコアを獲得できる傾向があった。これに対し ARC は、学習データに含まれない全く新しい規則性を持つパズルを解くことを要求する。これにより、AI が「記憶」に頼ることを防ぎ、純粋な推論能力、いわゆる「流動性知能（Fluid Intelligence）」を測定することを目指している[3]。

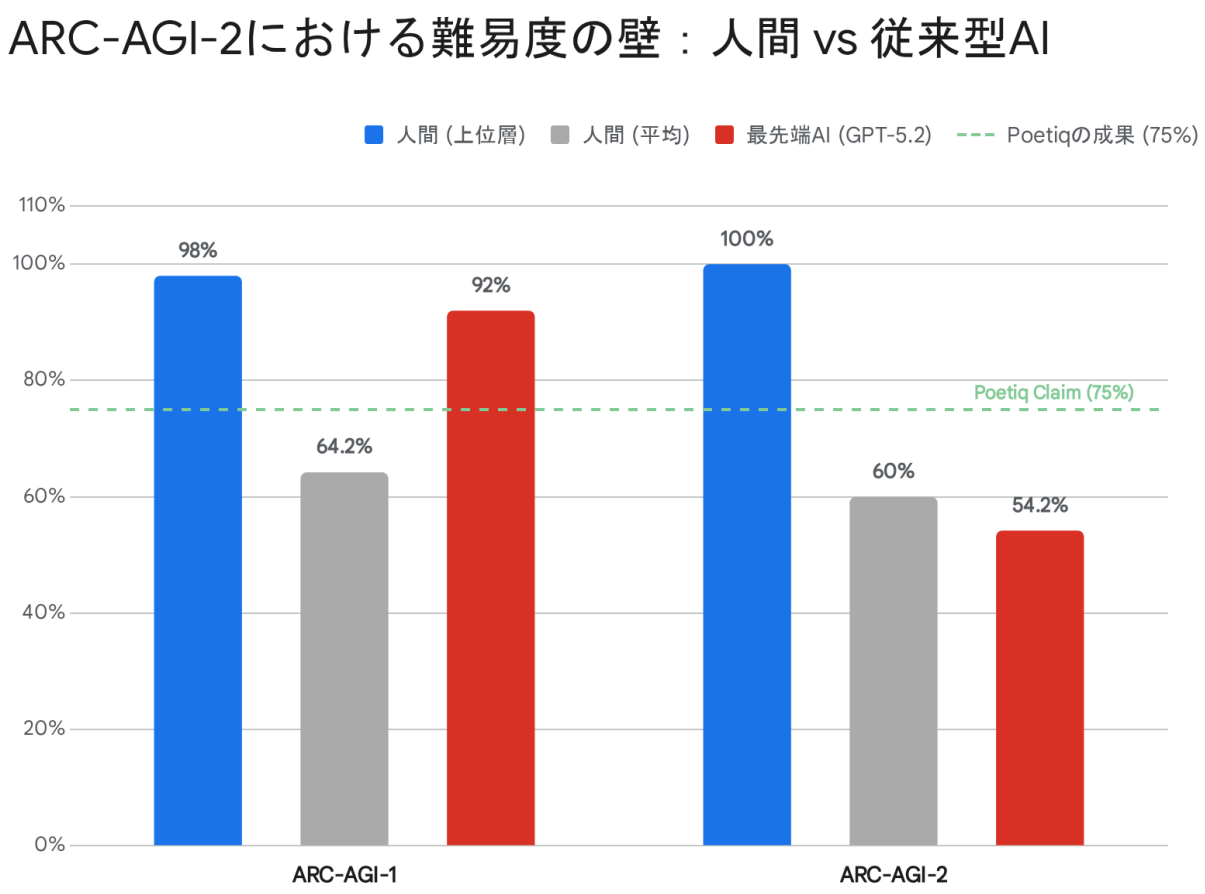
2.2 ARC-AGI-2 における難易度の質的転換

2025 年に導入された ARC-AGI-2 は、初代（v1）と比較して、その難易度と堅牢性が劇的に向上している。v1 では、一部のタスクが総当たり攻撃（Brute-force approach）や単純なピクセル操作で解決可能であったが、v2 ではそれらの抜け穴が塞がれ、より高次の抽象化能力が求められるようになった[3]。

具体的には、ARC-AGI-2 は以下の要素を強化している。

- **Core Knowledge Priors （核となる事前知識）の厳格化**: 人間が乳幼児期に獲得する物理的・空間的な基本概念（オブジェクトの永続性、対称性、包含関係など）のみを前提とし、文化的・言語的な知識を一切排除している[8]。
- **記号的解釈（Symbolic Interpretation ）**: 単なる視覚的パターンマッチングではなく、記号が持つ意味（例えば「赤い四角は『停止』を意味する」といった文脈依存のルール）を推論するタスクが増加した[9]。
- **構成的推論（Compositional Reasoning ）**: 複数のルールを組み合わせて適用する能力が

問われる。例えば、「オブジェクトを移動させる」ルールと「色を変更する」ルールを同時に、あるいは順序立てて適用する必要がある [9]。



ARC-AGI-1からv2への移行に伴い、従来のAIはスコアを落としていたが、GPT-5.2は54%まで到達し、人間（平均60%）に迫りつつある。Poetiqの75%はさらにその壁を超える成果となる。

Data sources: LessWrong, IntuitionLabs, ARC Prize Blog

2.3 人間と AI の決定的乖離

ARC-AGI-2 の設計において最も注目すべき点は、「人間にとっては依然として容易である」という事実である。検証データによると、400 人の被験者を対象としたテストにおいて、すべてのタスクが少なくとも 2 人の人間によって解決されている（解決率 100%） [9]。平均的な人間の回答時間は 1 タスクあたり約 2.3 分であり、特別な専門知識を持たない一般人でも高い正答率を叩き出すことができる [3]。

対照的に、GPT-4 や Claude 3.5 といった 2024 年までの最先端モデルは、ARC-AGI-2 におい

てほぼ0%に近いスコアしか記録できなかった[11]。これは、現代のAIが「膨大な知識の統計的処理」には長けているものの、人間が自然に行っている「少数データからの帰納的推論」においては、幼児レベルにも達していないことを残酷なまでに示している。Poetiqが主張する75%という数値は、この絶望的なまでのギャップを一気に埋める可能性を示唆しており、それゆえに事実であれば革命的なのである。

---

3. Poetiq AIの技術的特異点：メタシステムによる推論の拡張

Poetiq AIが他社と一線を画すのは、巨大なLLMを単体で使用するのではなく、それを包含する「メタシステム（Meta-System）」あるいは「推論ハーネス（Reasoning Harness）」と呼ばれる独自のアーキテクチャを構築している点にある[12]。このシステムは、創業者のShumeet BalujaとIan FischerがGoogle等の研究機関で培った敵対的生成ネットワーク（GANs）や強化学習の知見がベースとなっていると考えられる[15]。

3.1「プロンプト」から「インターフェース」へのパラダイムシフト

Poetiqの哲学において、LLMへのプロンプトは単なる命令文ではなく、巨大な知能データベースに対する「インターフェース」として再定義されている[13]。従来のプロンプトエンジニアリングが「いかに上手くモデルに質問するか」に注力していたのに対し、Poetiqのメタシステムは「モデルから知識を引き出し、それをどのように組み合わせるか」という高次の戦略を自動生成する。

具体的には、メタシステムはタスクの性質を分析し、最適な推論戦略（例えば、「まずは色に注目する」「次は形状の変化を追う」など）を立案する。そして、その戦略に基づいて複数のモデル（Gemini 3、GPT-5.1、Claudeなど）を動的に呼び出し、それぞれの得意分野を活用する「専門家アンサンブル（Mixture of Experts similar approach）」的な挙動を示す[13]。

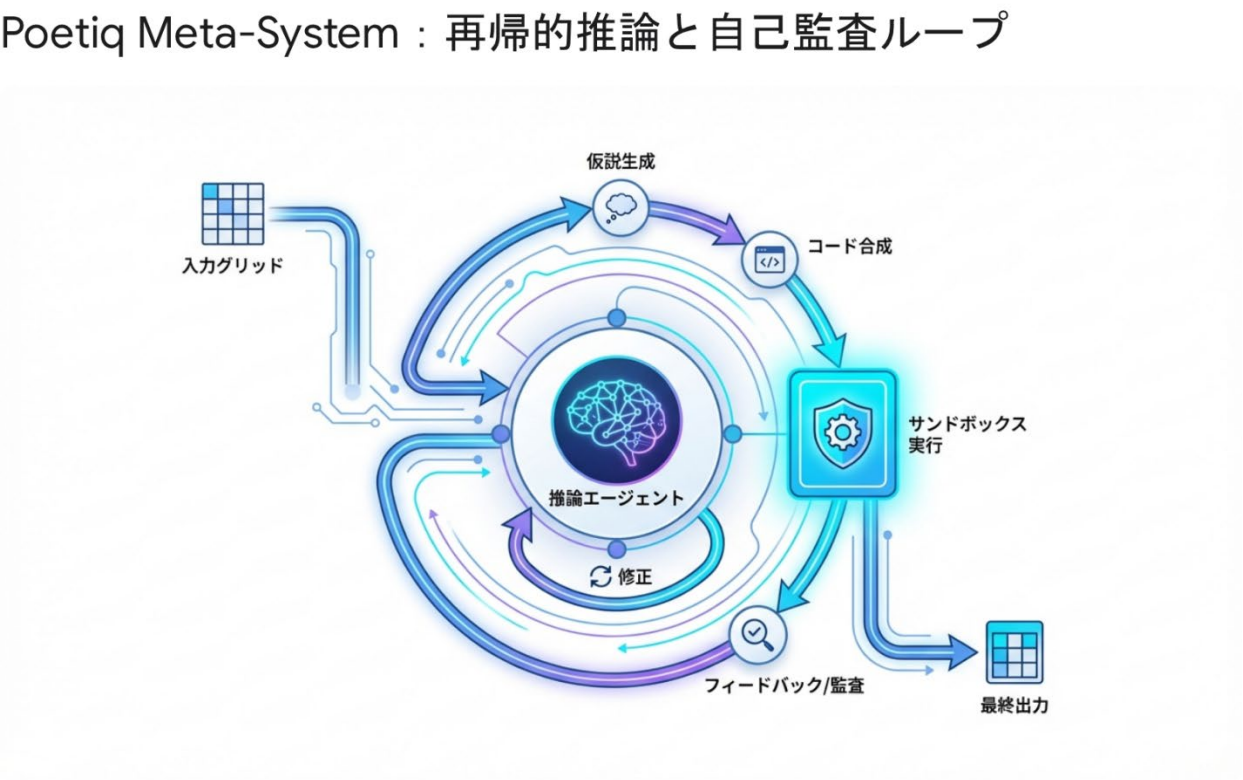
3.2 再帰的自己改善（Recursive Self-Improvement）のメカニズム

Poetiqのシステムの核心は、推論プロセスにおける「再帰的ループ」にある。従来のChain-of-Thought（思考の連鎖）プロンプティングは、一度の生成プロセス内で論理を展開するが、Poetiqのアプローチは、生成された解を実際に「テスト」し、その結果をフィードバックとしてシステムに戻すことで、回答を反復的に洗練させる[13]。

1. **仮説生成（Hypothesis Generation）**: LLMが入力グリッドから変換ルールの仮説を立て、それをPythonコードとして生成する。
2. **実行と検証（Execution & Verification）**: 生成されたコードを安全なサンドボックス環境で実行し、訓練データ（Demonstration pairs）に適用する。

3. **自己監査（Self-Auditing ）**: 実行結果が期待される出力と一致するかを厳密に検証する。一致しない場合、エラーログや出力の差異（Diff）を解析し、何が間違っていたかを言語化する。
4. **修正と再生成（Refinement ）**: フィードバック情報を元に、LLM がコードを修正・再生成する。

このプロセスは、一種の「テスト時学習（Test-Time Training）」として機能し、モデルの重みを更新することなく、そのタスク専用のスキルを一時的に獲得させることを可能にする [7]。



Poetiqのシステムは、LLMが生成した仮説（コード）を実際に実行し、その結果をフィードバックとして再びLLMに入力するループを持つ。これにより、テスト時に学習（Test-Time Training）に近い効果を得ている。

3.3 複数のエージェントによる相互監視

Poetiq のアプローチには、複数の AI エージェントが相互に監視し合う仕組みも組み込まれている可能性がある。Reddit などの議論では、3 つのエージェントを用いた構造化されたレビュープロセスにより、ハルシネーション（もっともらしい嘘）を 96%以上削減できるという研究結果が引用されており [1]、Poetiq も同様に、生成役（Generator）、批評役（Critic）、検証役（Verifier）といった役割分担を行っていると推測される。特に ARC-AGI のような正解が明確

なタスクにおいて、コード実行による検証（Ground Truth Checking）は最強のフィルタリング機能として働く。

---

4. GPT-5.2 X-High ：推論能力の極限とコストの壁

Poetiq のシステムはモデルに依存しない（Model-Agnostic）設計であるが、75%という記録的なスコアを達成するためには、基盤となる LLM の能力が不可欠である。ここでは、使用された「GPT-5.2 X-High」の実像に迫る。

4.1 "Thinking Tokens" と推論スケーリング則

GPT-5.2（特に Thinking/Pro モデル）は、OpenAI の o1 シリーズで導入された「Thinking Tokens（思考トークン）」の概念をさらに推し進めたモデルである [18]。これは、ユーザーへの回答を出力する前に、モデル内部で数千から数万トークンに及ぶ「思考の連鎖」を生成し、問題解決のプランニングや自己修正を行う機能である。

「X-High（Extra High Reasoning）」設定は、この思考プロセスに割り当てる計算資源を極大化したモードであり、通常の「High」設定よりもさらに深く、長く思考することを許可する [20]。ARC-AGI のような複雑な推論タスクにおいて、この「長考」は決定的な差を生む。モデルは、直感的な答えに飛びつくことを抑制し、複数の変換ルールをシミュレーションし、論理的な矛盾がないかを確認してから最終的な答え（またはコード）を出力する。

4.2 コーディング能力の向上と Poetiq との親和性

GPT-5.2 は、コーディングベンチマークである SWE-bench においても SOTA を記録しており [18]、そのプログラミング能力は極めて高い。これは、Python コードを生成して問題を解く Poetiq のアプローチにとって最大の武器となる。Poetiq のメタシステムが提示する修正指示を正確に理解し、複雑なアルゴリズムを実装できるだけのコーディング能力が、GPT-5.2 には備わっているのである。

4.3 コストという重い課題

しかし、GPT-5.2 X-High の使用には莫大なコストが伴う。推論トークンはユーザーには見えないが、API の課金対象となる出力トークンとしてカウントされるため、1 回の回答生成に数ドルかかることも珍しくない 19。
Poetiq は以前のモデルで「1 タスクあたり$1.90」というコスト効率をアピールしていたが 3、GPT-5.2 X-High を使用した場合、そのコストは跳ね上がる可能性がある。Reddit 上の議論では、ARC-AGI-2 の 120 問を解くために数百ドルから数千ドル規模のクレジットを消費する可能性が指摘されており 24、これは Poetiq が掲げる「効率的な推論」という目標と矛盾するリスクを孕んでいる。

5. スコア乖離の深層分析：75%は真実か？

Poetiq AI の発表で最も懸念される点は、過去の実績に見られる「公開スコアと検証スコアの乖離」である。前回の Gemini 3 ベースのシステムでは、公開セットで 65.32%を記録したものの、検証セットでは 54.0%に留まった。この約 11.3ポイントの下落はなぜ起きたのか、そして今回の 75%にも同様の現象が起きるのかを分析する。

5.1 データセット構造と「一般化ギャップ」

ARC-AGI-2 のデータセットは、以下の 4 つに分類される [11]。

データセット	タスク数	役割	公開状況
Training Set	1000	AI の学習用	公開（Public）
Public Evaluation Set	120	開発者の自己評価用	公開（Public）
Semi-Private Evaluation Set	120	リーダーボード用検証	非公開（Private）
Private Evaluation Set	100	最終コンペティション用	完全非公開（Hidden）

ここで重要なのは、**Public Evaluation Set** でのスコア（65.32%）と、**Semi-Private Evaluation Set** でのスコア（54%）の差である。これを機械学習の用語で「一般化ギャップ（Generalization Gap）」と呼ぶ。

5.2 乖離の主因 1：データ汚染（Data Contamination）

最大の要因は、学習データへの汚染である。GPT-5.2 のようなフロンティアモデルは、インターネット上のあらゆるテキストを学習しており、その中には GitHub や Kaggle で公開されている ARC の過去問や解答コードも含まれている可能性が極めて高い 25。
Public Evaluation Set は長期間公開されているため、GPT-5.2 はその「答え」を知っている（記憶している）可能性がある。一方、Semi-Private Set は非公開であるため、モデルは純粋

な推論能力で解かなければならない。記憶に頼って解いた問題と、推論で解いた問題の差が、そのままスコアの差として表れる。

5.3 乖離の主因 2：メタシステムの過学習（Overfitting ）

モデルだけでなく、Poetiq のメタシステム自体も過学習している可能性がある。開発チームは Public Evaluation Set を使ってシステムのプロンプトやロジックを調整しているはずである。例えば、「Public セットには回転を含むタスクが多い」と気づけば、回転を優先的に試すようシステムを調整するだろう。しかし、Semi-Private セットの分布がわずかに異なれば、その調整は逆効果になることもある。ARC Prize 主催者は、両者のスコア差が 10％を超えた場合、それは「過学習」であると見なす基準を設けている [8]。前回の 11.3%差は、まさにこの過学習の領域にあった。

5.4 75%の「補正後実力値」の試算

前回の乖離率（約 17%減：65.32% → 54.0％）を、今回の 75%に適用してシミュレーションを行うと、以下のようになる。

数理的推計:

$$\text{Estimated Verification Score} = \text{Public Score} \times (1 - \text{Discount Rate})$$

$$75.0\% \times (1 - 0.173) \approx 62.0\%$$

また、過去の事例や他のモデル（o3 など）の傾向 [17] を踏まえると、より保守的に 20 ポイント程度の下落を見積もる声もある。それでもなお、60％前後のスコアに着地する可能性が高い。

シナリオ	想定される減少幅	推定検証スコア	評価
楽観的（Optimistic ）	-5% 〜 -8%	**69% 〜 71%**	歴史的快挙。人間の平均を完全に凌駕。
現実的（Realistic ）	-10% 〜 -15%	**60% 〜 64%**	SOTA 更新。人間の平均と同等レベル。
悲観的（Pessimistic ）	-20% 〜 -25%	**50% 〜 55%**	現行 SOTA と同等。コスト増に見合わず。

特筆すべきは、たとえ「現実的シナリオ（60〜64%）」であったとしても、これは現在の公式記録（54%）を明確に上回り、ARC-AGI-2 における人間の平均スコア（60%）に到達することを意味する点である。

---

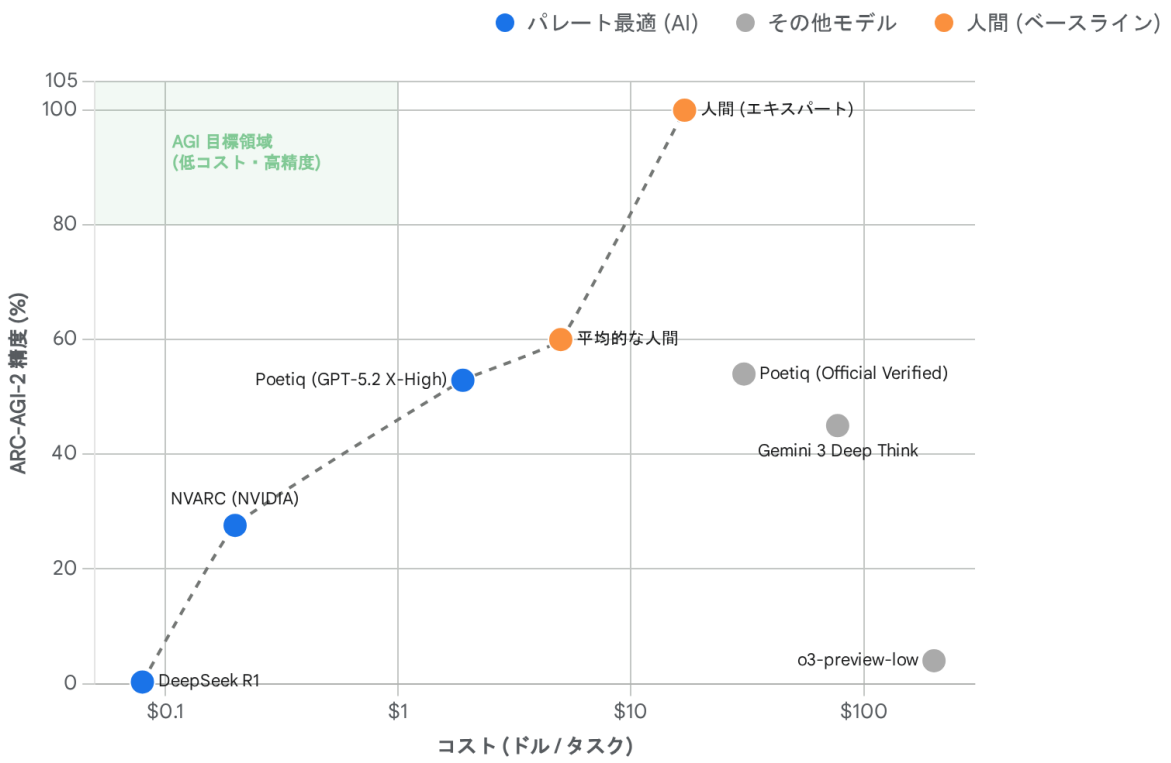
6. コストパフォーマンスと実用性の評価

技術的な達成度と同時に評価すべきなのが、経済的な持続可能性である。

6.1 パレート境界（Pareto Frontier ）の移動

Poetiq は、自社のシステムが「パレート境界（コストと精度のトレードオフにおける最適ライン）」を押し上げたと主張している 2。
確かに、GPT-5.2 X-High を用いた Poetiq のシステムは最高精度を叩き出すが、コストもまた最高レベルにある。一方で、NVIDIA の Kaggle 優勝チーム（NVARC）は、比較的小規模なモデル（8B クラス）をファインチューニングし、1 タスクあたり$0.20 という低コストで 27.6% のスコアを出している 7。

コスト vs 精度：AI推論のパレート境界線



Poetiq (GPT-5.2 X-High) は最高の精度（Y軸）を達成するが、コスト（X軸）も高い。対照的にNVARCなどは低コストで健闘している。真のAGIは、図の左上（低コスト・高精度）を目指す必要がある。

Data sources: GreaterWrong, NVIDIA Blog, Poetiq.ai, Intuition Labs, ARC Prize

6.2 「富豪的アプローチ」の限界と可能性

GPT-5.2 X-High を用いた Poetiq のアプローチは、現時点では「富豪的アプローチ（Brute-force compute）」の側面が強い。1 問解くのに数千円かかるシステムは、産業応用としては限定的である。しかし、AI の歴史が示す通り、今日のスーパーコンピュータレベルの計算量は、数年後にはスマートフォンレベルにまでダウンサイズされる可能性がある（例えば、蒸留モデルやハードウェアの進化による）。

したがって、現時点で「コスト度外視で最高性能を証明した」ことには、将来のロードマップを示すという意味で大きな価値がある。Poetiq もまた、より安価なモデル（Grok-4-Fast や GPT-OSS など）を用いた構成でコストダウンを図っており、スケーラビリティを意識した開発を行っている点は評価できる 13。

7. 今後の公式評価での見通しと結論

7.1 公式評価のタイムラインと予測

Poetiq AI は、この 75%の結果を ARC Prize の公式リーダーボードに登録し、検証を受けるプロセスに入ると考えられる。検証は非公開の Semi-Private セットで行われるため、前述の「一般化ギャップ」の影響が確実に現れる。

筆者の予測としては、検証スコアは 62%〜64%の範囲に着地すると見る。
根拠は以下の通りである。

1. **ポジティブ要因:** GPT-5.2 自体の汎用能力の高さと、Poetiq の再帰的修正システムの堅牢性は、単なる暗記では説明できないレベルにある。特にコード生成と実行による検証ループは、未知の問題に対しても強い汎化性能を持つ。
2. **ネガティブ要因:** データ汚染の影響は避けられず、10 ポイント程度の下落は統計的な必然である。また、GPT-5.2 X-High のコスト制約により、検証時の試行回数や探索深さが制限される可能性もある。

7.2 結論：AI は「知能の壁」を超えたか？

もし検証スコアが 60% を超えれば、それは AI 史における重要な転換点となる。ARC-AGI-2 における 60% という数字は、平均的な人間が持つ流動性知能と同等のパフォーマンスを意味するからだ。Poetiq AI の成果は、以下の 3 つの事実を我々に突きつけている。

1. **推論は「計算」できる:** 膨大な計算資源（Thinking Tokens）と適切なアルゴリズム（再帰的修正）を組み合わせれば、直感やひらめきに近い推論プロセスを AI で再現・代替できる。
2. **システムアプローチの勝利:** 単一のニューラルネットワークのパラメータを巨大化させるだけでは限界があり、コード実行や外部ツール、自己監査を含む「システム」として AI を構築する必要がある。
3. **AGI への道筋:** ARC-AGI-2 のような「暗記不能」なベンチマークの攻略は、AI が特定の狭いタスク（Narrow AI）から、汎用的な問題解決能力（General Intelligence）へと足を踏み入れたことを示唆している。

Poetiq AI の 75%という自己申告スコアは、割り引いて見るべきではあるが、決して無視してよい数字ではない。それは、AI が「人間並みの推論」を手に入れる日が、もはや SF の世界の話ではなく、エンジニアリングの射程圏内に入ったことを告げるファンファーレなのである。今後の公式検証結果が待たれる。

引用文献

1. Poetiq Achieves SOTA on ARC-AGI 2 Public Eval : r/singularity, 12月 24, 2025 にアクセス、

https://www.reddit.com/r/singularity/comments/1pu5mhk/poetiq_achieves_sota_on_arcagi_2_public_eval/
2. ARC-AGI-2 SOTA at Half the Cost - Poetiq, 12 月 24, 2025 にアクセス、https://poetiq.ai/posts/arcagi_verified/
3. GPT-5.2 & ARC-AGI-2: A Benchmark Analysis of AI Reasoning, 12 月 24, 2025 にアクセス、https://intuitionlabs.ai/articles/gpt-5-2-arc-agi-2-benchmark
4. Startup Poetiq just achieved an "Attention is All You Need ... - Reddit, 12 月 24, 2025 にアクセス、https://www.reddit.com/r/DeepSeek/comments/1p9da7t/startup_poetiq_just_achieved_an_attention_is_all/
5. ARC-AGI-1, 12 月 24, 2025 にアクセス、https://arcprize.org/arc-agi/1/
6. ARC-AGI v2 Leaderboard - LLM Stats, 12 月 24, 2025 にアクセス、https://llm-stats.com/benchmarks/arc-agi-v2
7. NVIDIA Kaggle Grandmasters Win Artificial General Intelligence ..., 12 月 24, 2025 にアクセス、https://developer.nvidia.com/blog/nvidia-kaggle-grandmasters-win-artificial-general-intelligence-competition/
8. arXiv:2412.04604v2 [cs.AI] 8 Jan 2025, 12 月 24, 2025 にアクセス、https://arxiv.org/pdf/2412.04604
9. ARC-AGI-2 A New Challenge for Frontier AI Reasoning Systems, 12 月 24, 2025 にアクセス、https://arcprize.org/blog/arc-agi-2-technical-report
10. ARC-AGI-2 human baseline surpassed (updated) - LessWrong, 12 月 24, 2025 にアクセス、https://www.lesswrong.com/posts/DX3EmhmwZjTYp9PBf/ai-performance-has-surpassed-a-human-baseline-on-arc-agi-2
11. Announcing ARC-AGI-2 and ARC Prize 2025, 12 月 24, 2025 にアクセス、https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025
12. Poetiq's AI Reasoning Layer Hits 54% on ARC-AGI-2 at Half the Cost, 12 月 24, 2025 にアクセス、https://www.vktr.com/ai-news/poetiqs-ai-reasoning-layer-hits-54-on-arc-agi-2-at-half-the-cost/
13. Traversing the Frontier of Superintelligence - Poetiq, 12 月 24, 2025 にアクセス、https://poetiq.ai/posts/arcagi_announcement/
14. Andrei Savu @andreisavu - Twitter Profile | Instalker, 12 月 24, 2025 にアクセス、https://www.instalker.org/andreisavu
15. Poetiq, 12 月 24, 2025 にアクセス、https://poetiq.ai/
16. Using a GAN to generate adversarial examples to facial image ..., 12 月 24, 2025 にアクセス、https://library.imaging.org/admin/apis/public/api/ist/website/downloadArticle/ei/34/4/MWSF-210
17. Open-source just beat humans at ARC-AGI (71.6%) for $0.02 per task, 12 月 24, 2025 にアクセス、https://www.reddit.com/r/LocalLLaMA/comments/1p7d97m/opensource_just_beat_humans_at_arcagi_716_for_002/
18. GPT-5.2: Pricing, Context Window, Benchmarks, and More - LLM Stats, 12 月 24,

2025 にアクセス、 https://llm-stats.com/models/gpt-5.2-2025-12-11
19. Reasoning models | OpenAI API, 12 月 24, 2025 にアクセス、 https://platform.openai.com/docs/guides/reasoning
20. GPT-5.2 X-High Review: OpenAI's Best Coding Model - YouTube, 12 月 24, 2025 にアクセス、 https://www.youtube.com/watch?v=duNxIdEQr_I
21. GPT 5.2 (X-High) and GPT 5.2 Pro (X-High) - TypingMind, 12 月 24, 2025 にアクセス、 https://feedback.typingmind.com/p/gpt-52-x-high-and-gpt-52-pro-x-high
22. OpenAI GPT-o1 API Pricing: A Comprehensive Guide, 12 月 24, 2025 にアクセス、 https://francpetracci.medium.com/openai-gpt-o1-api-pricing-a-comprehensive-guide-b93fdaed217c
23. ARC-AGI-2 human baseline surpassed (updated) - LessWrong 2.0 ..., 12 月 24, 2025 にアクセス、 https://www.greaterwrong.com/posts/DX3EmhmwZjTYp9PBf/arc-agi-2-human-baseline-surpassed
24. r/windsurf - Reddit, 12 月 24, 2025 にアクセス、 https://www.reddit.com/r/windsurf/best/
25. Introducing the ARC-AGI Public Leaderboard, 12 月 24, 2025 にアクセス、 https://arcprize.org/blog/introducing-arc-agi-public-leaderboard
26. 54% on ARC-AGI 2 is now Officially Verified : r/singularity - Reddit, 12 月 24, 2025 にアクセス、 https://www.reddit.com/r/singularity/comments/1pfk4t0/54_on_arcagi_2_is_now_officially_verified/