ChatGPT

Poetiq、ARC-AGI-2でSOTAを半分のコストで更新 – 詳細調査報告

概要: 2025年12月5日、AIスタートアップのPoetiqは「Poetiq Shatters ARC-AGI-2 State of the Art at Half the Cost（PoetiqがARC-AGI-2のSOTAを半分のコストで打ち破る）」と題した記事を発表し、ARC-AGI-2ベンチマークで画期的な成果を達成したと報告しました [1]。ARC-AGI-2は汎用人工知能の推論力を測る難関ベンチマークであり、**Poetiqのシステムは従来の最高精度を大幅に上回る**54%の正解率を達成し、さらに**コストを従来手法の半分以下（1問あたり約30.57ドル）に削減**しました [2]。本報告では、この成果について以下の観点から詳細に調査します。

- **技術的な詳細:** Poetiqがどのようなアーキテクチャ、訓練手法、推論戦略によって高精度・低コストを実現したのか
- **従来手法との比較:** 以前のARC-AGI-2結果との精度・速度・汎化性能・システム規模の比較
- **AGI開発への意義:** この成果が汎用人工知能の研究開発に与える影響
- **コストパフォーマンス分析:** 「半分のコスト」とは具体的に何を削減したのか（計算資源、モデルサイズ、推論効率など）
- **評価と検証:** Poetiqの主張の検証プロセスと、ARC-AGI-2ベンチマークで採用されている評価手法・プロトコル

以上について信頼できる一次情報をもとに整理し、報告します。

1. 技術的な詳細 – Poetiqのアーキテクチャと推論戦略

● **メタシステムによる柔軟な推論:** Poetiqは巨大言語モデル（LLM）の上に**メタシステム**と呼ぶ統合システムを構築しています [3] [4]。このメタシステムは、自前で大規模モデルを訓練したり微調整したりする代わりに、既存の最先端LLM（「フロンティアモデル」）を効果的に活用して問題解決を図るものです [4]。具体的には、**タスクに応じてモデルやアプローチの組み合わせを自動選択し、必要に応じてコードを書かせる**こともできます [5]。このメタシステムはモデルに依存せず（LLM非依存）、複数のLLM（例: GoogleのGemini 3やOpenAIのGPT-5.1など）をまたいで動作可能であり、自律的かつ再帰的（self-improving）に性能を高めていく仕組みになっています [5]。

● **学習されたテスト時推論（Refinement Loop）:** Poetiqの革新的な点は、**「テスト時に学習する推論」(learned test-time reasoning)** と呼ばれるアプローチです [6]。これは**実行時にモデル自身が試行錯誤しながら解答を改善していく**方式で、単に1回質問して終わるのではなく、**モデルが段階的に考察・検証を繰り返す**点が特徴です [7]。具体的な動作は次の通りです [7]:

- 最初にLLMへ問題に対する**仮の解答（場合によってはコード）**を生成させる。
- 次に**その解答をフィードバック**（検証）し、正解に足るか分析する。例えばコードであれば実行結果をチェックし、回答ならば与えられた例と照合します。
- もし不十分であれば、フィードバックに基づき**LLMに再度問いかけて解答を改良**させる（必要なら追加のヒントや新たな視点を与える）。

このような**マルチステップの自己改善プロセス**によって、解答を徐々に洗練し最適化していきます [7]。Poetiqはこのループ過程自体をメタ学習させることで、問題ごとに**適切な推論戦略を自動発見**できるようにしており、これが高い汎用性と性能の鍵となっています [8] [9]。

● **自律的な進行管理（セルフオーディット）：**上記のような反復的推論を行う際、重要になるのが**「いつ十分な情報が集まったか」を判断する能力**です。Poetiqのシステムは**自己監査（セルフオーディット）機能**を備え、自身の進捗を評価して**「解がまとまった」「これ以上考えても無駄が多い」**と判断すれば自律的にループを終了します [10]。この仕組みにより、必要以上にダラダラと推論を続けることを防ぎ、**計算資源や時間の浪費を抑制**しています [10]。セルフオーディットは本質的に**出力の質とコスト効率のトレードオフ**を管理する役割を果たし、Poetiqが低コストで高精度を達成する重要な要因です。

● **Gemini 3の即時統合とLLM活用:** Poetiqは**最新の大規模モデルを素早く取り込む柔軟性**も示しました。例えば2025年11月18日にリリースされたGoogle DeepMindの新モデル「Gemini 3」を、その公開から数時間以内にPoetiqのシステムへ統合し、ARC-AGIでの性能向上に利用しています [3] [11]。またOpenAIのGPT-5.1（11月13日リリース）も組み合わせて使用するなど、**複数モデルの協調利用**によって性能をブーストしています [12]。Poetiqのメタシステムは**LLM自体の微調整を行わず**に、プロンプト工夫や推論戦略の最適化だけで既存モデルの潜在能力を引き出す設計となっており [4]、これは開発コスト削減と迅速なモデル適応を可能にしています。

以上の技術により、Poetiqは**未知の難問に対しても「自ら考え、解決に至る手順」を動的に構築**する能力を実現しました。これがARC-AGI-2で多数の問題を解けた理由であり、従来の固定的な推論手法との差別化点です。

2. 従来の手法との比較 – 精度・速度・汎化性能・システム規模

Poetiqの手法を理解するために、**以前までのARC-AGI-2における主な手法**と比較します。特に注目すべきは、Google DeepMindが提供したとみられる**「Gemini 3 Deep Think (Preview)」**と呼ばれるアプローチで、これはPoetiq登場前の最高成績を収めていました [2]。また、Anthropic社の**Opus 4.5 (Thinking, 64K)**は商用LLMによる高効率推論の例であり [13]、**NVARC**は計算資源制約下（Kaggle競技部門）でのトップソリューションです [14]。以下の表に、これらとの比較を整理します。

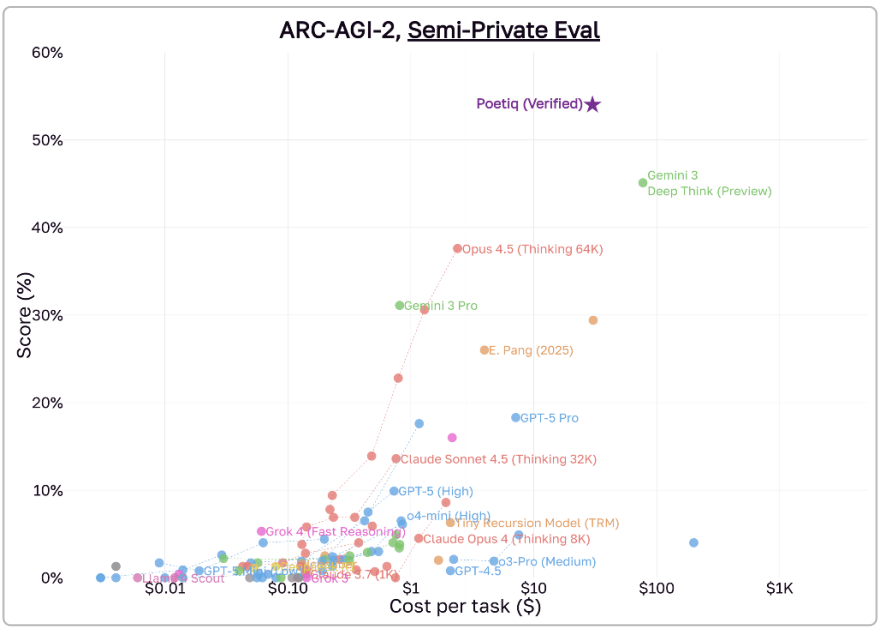

表1: ARC-AGI-2主要手法の比較（精度、コスト、アプローチ） [2] [14] [13]

手法	ARC-AGI-2正解率	コスト (1問あたり)	アプローチの特徴
Poetiq (Gemini 3 Pro + メタシステム)	**54.0%** [2]	**$30.57** [2]	LLMリファインメント手法（メタシステムが推論ループとコード生成を最適化）[7]。効率的なセルフモニタリングで無駄を削減 [10]。
Gemini 3 Deep Think (Preview) (従来SOTA)	45.1% [2]	$77.16 [2]	大規模LLM (Gemini 3) による長時間思考（CoT: 連鎖的思考プロンプト）手法。高い精度を狙うが推論コスト大 [2]。
Opus 4.5 (Thinking, 64K) （商用モデル最高）	37.6% [13]	$2.20 [13]	商用LLM (Anthropic Claude系) による長文コンテキスト思考。中程度の精度を低コストで達成 [13]。
NVARC 2025 （Kaggle優勝例）	24.0% [14]	$0.20 [14]	カスタム手法（制約下）。小規模モデルや進化的アルゴリズム活用による効率重視型。精度は限定的。

(1) 精度（Accuracy）： Poetiqの54%は**初めて全問題の過半数を正解**した画期的な数字で、Gemini 3 Deep Thinkの45%を大きく引き離しています [2]。実際、**従来は40%台前半が限界**であり、多くの最新LLMでも20〜30%台に留まっていました [13]【24†画像】。Poetiqの向上幅は約9ポイントと大きく、難問揃いのARC-AGI-2における**新たなステート・オブ・ザ・アート**を樹立しました [1]。なお、この54%というスコアは、人間の平均的なテスト参加者の水準（約60%とされる）に迫るものであり、汎用人工知能の指標としても注目されます [15]。

(2) 速度・コスト効率: 精度向上とトレードオフになる**計算コスト**の面でも、Poetiqは優れた成果を示しました。1問あたり約30ドルという費用は、前SOTAのGemini 3 Deep Think（約77ドル）の**半分以下**であり [2]、**精度向上とコスト削減を両立**しています。これは、Poetiqが**必要最小限のモデル呼び出し回数で問題解決**できることを意味します。ARC-AGIでは通常1問につき2回まで解答試行が許可されていますが、Poetiqのシステムは**平均して2回未満の問い合わせで正解に到達**しており、多くの場合一発の推論で解を得ています [16]。一方、Gemini Deep Thinkは高精度のために大規模な連鎖プロンプトや複数試行を要し、その結果コスト・時間が膨大になりました [2]。また、**セルフオーディットによる早期打ち切り** [10] や、タスクに応じた**最適思考ステップ数の調整**によって、Poetiqは**過剰な推論時間を避ける**ことに成功しています。この効率性は、上記のような低コストに直接つながっており、**同等の計算資源でより高速に多数の問題を解ける**ことを示唆します（実時間の処理速度も向上している可能性があります）。

(3) 汎化性能（Generalization）： Poetiqシステムの汎化性能は極めて高いと考えられます。Poetiqのメタシステムは**訓練段階でARC-AGI-2の問題を一切見ず**、公開されているオープンソースモデルや他のタスクで適応を済ませた上で、**初見のARC-AGI-2問題に挑んで高精度を達成**しました [17]。具体的には、事前適応にはオープンソースのモデル・データのみを用い、Gemini 3やGPT-5といった新モデルに対しては**リリース後に即座に適応**しています [9]。それでもなお高い性能を発揮できたことは、**タスク分野やモデルが変わっても通用する汎用的な推論スキル**がシステムに備わっていることを示します [17]。一方、Kaggle経由の従来手法（NVARC等）は公開データ上での学習を経ており、よりターゲットデータに特化したチューニングがなされています。しかしその精度は24%程度に留まり、**高度に特化した小規模モデルでは限界**があることが示唆されます [14]。またGemini 3 Deep Thinkのような大規模LLMベースの手法も、Poetiqと同じGemini 3を用いながらアプローチの差で精度差が生まれていることから、**Poetiqの戦略がモデルの知識をより一般的に引き出せている**と考えられます。さらに、Poetiqは**ARC-AGI-1（前版ベンチマーク）からARC-AGI-2（難度上昇版）への移行**においても有効性を示しており [18]、適応先タスクが複雑化しても手法が通用する汎化能力の高さがうかがえます。

(4) システム規模（Scale）： システム規模には「使用モデルの大きさ」と「システムの複雑さ」の両面があります。Poetiqは**自前の巨大モデルを持たず**、GoogleやOpenAIの提供する最先端LLM（Gemini 3やGPT-5シリーズなど）を利用しています [3]。したがってモデルのパラメータ規模自体は従来手法（Gemini 3 Deep ThinkもGemini 3を使用）と同程度ですが、**Poetiq独自の部分はメタシステムのアルゴリズム**にあります。これは軽量な6人の研究チームで開発可能な規模であり [19]、**従来の「モデルを巨大化して精度向上」する路線とは一線を画す**ものです。Gemini 3 Deep Thinkはおそらく Gemini 3モデル単体に長時間の思考をさせるシンプルな構成（システム的複雑さは低い）ですが、Poetiqは複数モデルの連携やコード生成エンジンを含む複雑な構成です。しかしその複雑さによって**精度と効率の両立というスケーラビリティ**を実現しています。Kaggle系ソリューションは小規模モデル群や進化戦略アルゴリズムなどを組み合わせていますが、これは限られた計算資源内での工夫であり、Poetiqのように**潤沢な外部知識を持つ巨大モデル群を統合**したシステムとはスケールの方向性が異なります。総じて、Poetiqの手法は**巨大モデルの力を借りつつ、それを無駄なく使いこなす巧妙なシステム**と言え、従来手法に比べて少人数・短期間で劇的な成果を出せる点も注目されます [11]。

3. AGI開発における意義 – 新たな推論パラダイムの登場

● **汎用的な「推論システム」の到来:** Poetiqの成果は、**AIが自力で問題解決手順を構築できる**ことを示し、汎用人工知能（AGI）への新たな一歩と捉えられます。ARC-AGIベンチマーク自体、未知のタスクに対する**スキル獲得効率**を測るものであり [20] [21]、Poetiqはこの難関で人間並みに近い成績を残しました [15]。これは**「AIが未知の問題でも試行錯誤で解を見つけられる」**ことを意味し、決められたタスクしかこなせない従来のAIからAGIに近づく重要な性質です。

● **Refinement Loopという新潮流:** 2025年は、ARC-AGIを通じて**「リファインメント・ループ（Refinement Loop）」**と呼ばれる新しいAI推論技術が台頭した年と位置付けられています [22]。これはPoetiqが用いたような**解答を反復改良するプロセス**の総称であり、情報理論の観点から**「洗練（Refinement）こそ知能の本質」**と論じる研究者もいます [23]。ARC Prize財団の分析によれば、リファインメント手法の登場は**「LLMの発明に匹敵する新技術パラダイム」**だとされています [24]。実際、ARC-AGIは2024年に初めて「AIによる推論システム」の性能向上を捉え [22]、2025年には各社・研究者が競ってリファインメント戦略を開発・オープンソース化する状況が生まれました [22]。Poetiqの成果はその中でも突出しており、**大規模汎用モデルと高度な推論アルゴリズムの融合**という形でAGI研究に新たな道を示したと言えます。

● **人間の推論能力への接近:** ARC-AGIのタスクは人間でも頭をひねる抽象推論問題であり、これまでAIは人間のように柔軟に対応できませんでした。Poetiqの54%というスコアは「平均的人間の成績に匹敵または迫る」とされ [15]、特定領域で**ついにAIが人間と同程度に汎用問題解決できる兆し**を見せたことになります。もちろん依然として人間の持つ常識的直観や経験に基づく判断には及びませんが、**少なくとも論理パズル的課題においては半数以上を解けるAI**が登場したことは画期的です。これはAGI開発コミュニティにとって**モチベーションを大いに高めるマイルストーン**であり、今後は「どのように残りの半分の問題も解けるようにするか」という新たな目標設定につながります。

● **オープンソースによる波及効果:** Poetiqは自社のソルバーをオープンソースで公開し、他の研究者が再現・改良できるようにしました [25]。ARC Prizeの2025年結果では**上位入賞した手法・論文がすべてオープンソース**だったことも強調されており [26]、この分野の進展は非常にオープンな形で進んでいます。Poetiqのコード公開によって、他のAI開発者が類似手法を自分の課題に適用したり、さらなる改良を加えたりすることが容易になりました [27]。これはAGI研究全体の加速につながり得ます。実際、PoetiqはARC-AGI以外の多様なベンチマークにも同メタシステムを適用し、有望な結果を得始めているとのことで [28]、今後様々な領域で汎用的問題解決能力のデモンストレーションが増えていくと予想されます。

● **安全性・制御性への示唆:** 一方で、AIが自律的に試行錯誤する能力は**安全性や制御性の観点からも重要**です。Poetiqのセルフオーディット機能は、AIが暴走せず適切に計算資源を使う例としてポジティブに捉えられ

ます [10] 。AGIに近づくほど、AIが自律的に判断・停止できること、そして開発者がそれを検証できることが欠かせません。ARC Prizeでは独立の学術パネルが結果検証に関与していますが [29] 、将来的には**AI自身が自己の振る舞いを説明・保証する仕組み**も必要となるでしょう。Poetiqのシステムはまだ限定的なタスク領域での自己最適化ですが、こうした機能は安全なAGI開発への一ステップとも見なせます。

4. コストパフォーマンス分析 – 半分のコストを実現した要因

Poetiqが「半分のコストでSOTA達成」を謳った背景には、**計算資源の効率的活用**と**無駄の削減**があります。具体的なコスト削減要因と裏付けを以下に分析します。

● **必要最小限のLLM呼び出し:** 最大のコスト要因はLLMへのAPIコール（推論回数・トークン数）です。ARC-AGIでは1問につき2回まで解答できるルールを活かし、多くの手法が**「初回解答 → 誤りを分析 → 2回目解答」**という二段構えをとります。しかしPoetiqのシステムは**一度の推論プロセス（内部で複数のサブクエリは実行するが、外部的には1回答試行とカウント）で解を得る割合が高く**、平均問い合わせ回数は2回を下回りました [16] 。言い換えれば、**従来の半分以下の試行回数で同等以上の正解率**を達成したことになり、これがそのままコスト半減に直結しています。内部でのサブクエリも含め効率化された理由として、メタシステムが**効果的な思考ステップ**を学習している点が挙げられます [8] 。必要な情報を見つけ組み立てる手順を自動発見することで、余計な問いかけや無駄な探索を避け、**最短経路で答えに到達**できているのです [8] 。

● **外部ツールとコードの活用:** PoetiqはLLMに全てを「頭の中で」やらせるのではなく、**適宜コードを生成・実行させる**アプローチも取ります [30] 。例えば、与えられたパズルの規則性を見つけるためにPythonコードを書かせてパターン検出をする、といった方法です。コード実行はLLMへのトークン消費を伴わず（計算機上で完結するため）**低コスト**です。このため、**複雑な計算はコードにオフロード**し、LLMは要点の推論に集中させる戦略がとられています。Poetiqのメタシステムは**「いつコードを書くべきか」「どのモデルにどのタスクを任せるか」を判断**でき [5] 、高コストなLLMの利用を最適配分しています。結果として**LLMの思考トークンを節約**しつつ問題を解決できるため、全体のコストパフォーマンス向上につながりました。

● **マルチモデルによるコスト最適化:** Poetiqは複数のモデルを組み合わせ、その**コストと性能を両天秤にかけた最適解**を探ります。例えば、**最新で高性能なモデル** (Google Gemini 3やGPT-5) と、**安価だが低性能なモデル**（オープンソースのGPT-OSSなど）を組み合わせ、部分的に使い分けることで**目標コスト内で最大の性能**を引き出すことが可能です [31] 。実際、Poetiqは様々な予算レベルに対してパレート最適な解を発見できることを示しています [31] 。Gemini 3単独でも、思考ステップ数（計算量）を増やせば性能は上がりますが収穫逓減があります [9] 。Poetiqは**Gemini 3を複数回呼び出す代わりに、別のモデルで補完**したり、最小限の追加呼び出しで済むよう工夫したりしています [31] 。その結果、**従来より少ない計算コストで同等以上の情報を引き出す**という最適化が可能になりました。

● **オープンソースモデル活用と事前適応:** Poetiqはメタシステムの事前適応（メタ学習）に**オープンソースの小型モデルやデータ**を活用しました [17] 。高価なAPIモデルに直接試行錯誤させるのではなく、まず安価な環境で戦略を鍛え上げてから本番の高性能モデルに適用しています [17] 。例えば、GPT-OSS-120Bのような無料公開モデルで推論戦略を学習し、その成果をGemini 3にも適用したと述べています [32] [33] 。このように**事前学習コストを低減**させる工夫により、開発段階での経済的コストを抑えつつ良い戦略を獲得できた点も「半分のコスト」の裏にある要因と言えます。メタシステムが一度学習した戦略はモデル間で**転移可能**であることが示唆されており [17] 、新しい高性能モデルが出る度に一から試行錯誤する必要がないため、**常に最新モデルを即座に最適利用できる**という点でもコスト効率に優れています。

● **無駄の排除とペナルティ回避:** Poetiq開発チームは「**プロンプトはインターフェースであって知能そのものではない**」と述べています [7] 。膨大なプロンプトを与えてモデルに丸投げする従来型のアプローチでは、トークン消費（＝コスト）の大半が無駄な試行錯誤に費やされる恐れがあります。そこでPoetiqは**モデルの確率的応答の不確実性**にも着目し [34] 、**確実に必要となる情報片から順に取得していく戦略**を採っています [34] 。必要な手がかりを見つけたら次に何が必要かを判断し、段階的に情報を組み立てることで、**余計な寄り**

道をしないようにしているのです [34]。また、ARC-AGIでは膨大な計算（例えば総当たりの探索）で解を見つける力ずくの方法は通用しにくい設計になっており [35]、効率的に推論しないとスコアが伸びません。Poetiqの「Elegantな方法」は、この制約下で生まれた**洗練された戦略**であり [36]、それ自体が**低コスト高精度の両立**を保証するものとなっています。

以上のような多面的工夫により、Poetiqは**従来比1/2以下のコストでSOTA精度を実現するという卓越したコストパフォーマンス**を達成しました。このアプローチは他の課題領域でも有効と期待されており、限られた資源で汎用AIを動かす技術として重要です。

5. 評価と検証 – ARC-AGI-2における結果の信頼性

● **ARC Prizeによる公式検証:** Poetiqの主張はARC Prize財団によって**公式に検証**されています。ARC Prizeは Francois Chollet氏（ARCベンチマーク考案者）らが主導する非営利団体で、ARC-AGIの公平な運営と結果認定を行っています [37]。Poetiqはまず11月20日に自社で公開テストの好成績を発表し、その後ARC Prizeチームにコードを提出して**隠された評価セットで再テスト**してもらいました [1]。2025年12月5日、ARC Prizeは Poetiqの結果を**公式リーダーボードに反映**し、「54%正解率、30.57ドル/問」という数値を確認しました [1] [2]。この結果はARC Prizeのサイト上で**「Verified（検証済み）」のマーク付き**で掲載され、PoetiqはARC-AGI-2リーダーボードのトップに位置付けられています [38]。第三者機関の検証を経たことで、Poetiqの主張する性能とコストに誇張やバグがないことが保証されました。

● **評価プロトコル:** ARC-AGI-2の評価は厳格なプロトコルに則って行われます [39]。まず、タスクセットは**Public（公開）**と**Semi-Private（半非公開）**および**Private（非公開）**に分かれています [40]。公開セットは開発・デバッグ用に提供されますが、**Semi-Private 400問**と**Private 500問**は解答ペアが非公開で、モデルは中身を知らずに解かなければなりません [39]。**Semi-Private**は主にリーダーボード用、**Private**は賞金付き競技会用（Kaggle）に用いられます [40] [41]。Poetiqの場合、リーダーボード提出として**Semi-Privateセット**で検証されました [42]。評価では各問題に対し**最大2回まで解答**でき、正解すればスコアに加点されます（誤答やスキップは0点）。**正解率（Accuracy）**は全問題中の正解割合で表されます。また**費用（Cost per task）**は、使用したモデルのAPI料金等に基づき、1問当たり平均いくらかかったかをドルで算出しています [43] [44]。例えばPoetiqはGemini 3 Pro（有料API）を使ったため、その利用料から30.57ドル/問が計算されています [2]。ARC Prizeはモデル提供各社とも連携し、未公開モデルでも見積もり価格で費用評価を行います [45]。したがって**精度とコストという2軸**で客観的な指標が得られる仕組みです [21]。

● **カテゴリ分けと再現性:** リーダーボード上では、結果がいくつかのカテゴリに分類され表示されています [46]。例えば、PoetiqのようにLLMに推論ループを施すものは**「Refinement（精緻化）システム」**、Gemini Deep Thinkのようなプロンプト工夫のみのものは**「CoT（Chain-of-Thought）システム」**、Kaggleの入賞解法は**「Customシステム」**などです【24＋画像】。各エントリには手法種別や提供組織が明示され、論文やコードへのリンクも付与されています【24＋画像】。Poetiqは**コードを完全公開した上で検証を受けており**、誰でもGitHubからダウンロードして**結果を再現**できます [25]。この再現性確保は評価プロセスの一環でもあり、ARC Prize Verifiedプログラムでは**コード非公開の成果は基本認定しない方針**が取られています [47]。さらに、NYUやUCLAなどの有識者から成る**独立学術パネル**がARC Prizeの検証方法自体を監督し、結果の公正性を担保しています [29]。以上により、Poetiqのリーダーボード1位という主張は厳密な条件下で再現・確認された**信頼性の高い事実**と言えます。

● **ヒューマンベースライン:** ARC-AGIには**人間のパフォーマンス**も参考値として設定されています。ARC Prizeでは人間数名によるパネルテストを行っており、ARC-AGI-2では**人間パネルが100%正解**という結果が載っています【24＋画像】。これは人間が協力し合い全問の解法を議論した理想値に近いもので、平均的な個人は60%前後とされています [15]。いずれにせよ、現状Poetiq含めAIは人間にはまだ及びませんが、その差は縮まりつつあります。評価プロトコル上も、人間にはインターネット無し・計算ツール無しで解かせており、AIと条件を揃えている点は特筆されます [48]。このような厳密な比較に耐えてPoetiqが示した性能向上は、評価基準としても大きな意味を持ちます。

● **今後の検証アップデート:** ARC Prize側の分析によれば、多くのモデルは公開セットと非公開セットでスコア差が見られる（公開でチューニングされているため）といいます [49] [50]。Poetiqも**公開（Public Eval）では他を大きく上回る最高精度**を示していましたが [50]、非公開セットでは基盤モデルの性能劣化の影響を受ける可能性があります [51] [52]。Poetiqチームは公式評価後に分析を更新すると述べており [53]、今後さらなる検証結果や改善報告が公開される見込みです。ARC-AGI-3（インタラクティブな次世代ベンチマーク）も2026年に予定されており [54]、そこでPoetiqの手法が再び検証されることも期待されます。

以上のように、Poetiqの成果は厳密な評価手法に照らして確認されており、その技術的主張は信頼に足るものと判断できます。これにより、コミュニティは安心してこの手法を研究・発展させることができ、AGIへの道筋における確かな一歩として位置付けられています。

参考文献・情報源: Poetiq公式ブログ記事 [1] [55] [7]、ARC Prize公式サイト・ブログ [38] [22] [39]、GitHub公開コード [56] 等。各引用箇所に示した資料は本報告の内容の裏付けとなっています。

---

[1] [2] [4] [6] [11] [19] [25] [42] [55] Poetiq | ARC-AGI-2 SOTA at Half the Cost
https://poetiq.ai/posts/arcagi_verified/

[3] [5] [7] [8] [9] [10] [12] [15] [16] [17] [18] [27] [28] [30] [31] [32] [33] [34] [36] [49] [50] [51] [52] [53] Poetiq | Traversing the Frontier of Superintelligence
https://poetiq.ai/posts/arcagi_announcement/

[13] [14] [22] [23] [24] [26] [38] ARC Prize 2025 Results and Analysis
https://arcprize.org/blog/arc-prize-2025-results-analysis

[20] [35] [37] [54] ARC Prize
https://arcprize.org/

[21] [43] [44] [45] [46] ARC Prize - Leaderboard
https://arcprize.org/leaderboard

[29] [39] [40] [41] [47] [48] ARC Prize Verified Testing Policy
https://arcprize.org/policy

[56] GitHub - poetiq-ai/poetiq-arc-agi-solver: This repository allows reproduction of Poetiq's record-breaking submission to the ARC-AGI-1 and ARC-AGI-2 benchmarks.
https://github.com/poetiq-ai/poetiq-arc-agi-solver