ChatGPT

ARC-AGI-2とPoetiq AIの最新成果と今後の展望

Poetiq AI、GPT-5.2 X-Highで75%の正解率を達成

2025年12月24日、AIスタートアップ**Poetiq AI**は最新モデル**GPT-5.2 X-High**を自社のシステムに組み込み、汎用知能評価ベンチマーク**ARC-AGI-2**の公開評価データセットにおいて**75%**という驚異的な正解率を記録したと発表しました [1]。平均すると**1問あたり8ドル未満**のコストでこの精度に達しており、従来の最先端（SOTA）を約15ポイント上回る成果です [2]。ARC-AGI-2の**人間平均正解率（約60%）**も大きく凌駕しており、AIシステムが人間の平均的なテスト参加者を上回る性能を示したことになります [2]。

Poetiqの発表によれば、**GPT-5.2 X-High**を用いるにあたりPoetiq側で特別な追加学習やモデルのチューニングは一切行っておらず、既存の**Poetiqハーネス（メタシステム）**に新モデルをそのまま組み込んだだけで大幅な精度向上とコスト削減が得られたといいます [1]。これは**前世代モデルから極めて短期間での性能向上**を示すもので、**精度・費用効率**の両面で飛躍的な改善となりました [1]。Poetiqチームは「前回同様に公開評価と公式評価（非公開テスト）で同じ傾向が維持されれば、既存のどの構成よりも大幅に高いスコアが期待できる」と述べ、GPT-5.2の協力に感謝しつつ公式評価結果を見守っています [3]。

11月の発表と公式結果：65.32%から54.0%へ

今回の75%達成に先立ち、Poetiq AIは**2025年11月20日**にもARC-AGI-2関連で注目すべき結果を公表していました [4]。当時Poetiqは、新たにリリースされた**Google DeepMindのGemini 3**モデルや**OpenAIのGPT-5.1**モデルを即座に自社システムへ統合し、ARC-AGI-2公開セットで**65.32%**というスコアを達成したと報告しています [5] [6]。これは**平均的な人間テスト参加者の60%を初めて上回るモデル**となり、大きなマイルストーンでした [6]。またその際、**従来のリーダー**であったGemini 3を用いたチーム「Deep Think」の約46.1%（コスト77.16ドル/問）をも精度・効率で大きく凌駕した点も強調されています [6]。

しかしこの**65.32%**という値はあくまで公開評価セット上での社内実験結果であり、その後**ARC Prizeチーム**による公式検証（非公開テストセットでの評価）では**54.0%**というスコアに留まりました [7]。公式結果54%でも当時のSOTAを大きく塗り替える快挙でしたが、公開セットでの事前想定値より**約11ポイント低下**したことになります [7]。Poetiq自身も11月のブログで「公開から非公開評価への移行で多かれ少なかれ性能劣化が起こるのはARC-AGIで既知だが、ARC-AGI-2では両セットがより慎重に校正されているため差異は小さめだろう」と予測していました [8]。実際には二桁の低下幅となったものの、**それでも54%は初の“過半数問題解決”**であり大きなブレークスルーでした [4] [7]。ちなみにこの公式54%時点での1問あたりコストは**$30.57**で、前SOTAである**Gemini 3 Deep Thinkの45%・$77.16/問**に比べ**半分以下の費用**で達成された点も注目されました [7]。

以上を踏まえると、今回公称された**75%**という公開セット結果も、いざ公式の隠されたテストで検証するとある程度低下する可能性があります。前回は約11ポイントの差でしたから、「今回も10%以上下がるのではないか」と見る向きもあります（例えば75%→約64%程度） [2]。実際、Poetiqも「ARC-AGI-1では公開→非公開で性能低下が見られたが、ARC-AGI-2では多くのモデルで差が小さい傾向だ」と述べつつ、自社結果について「我々も同様になると予想する」と記していました [8]。いずれにせよ、公式評価で**60%台中盤**という結果になったとしても、それは依然として**人間平均を上回る新たな最高水準**であり、ARC-AGI-2史上トップの座を維持する見込みです。

ARC-AGI-2: 汎用な抽象推論を測る難関ベンチマーク

ARC-AGI-2とは、AIの**抽象的な推論力**や**適応的問題解決力**を測るために設計された評価基盤で、2024年から始まった**ARC　Prize**の一環として構築されたものです [9] 。名前が示す通り、これは**人工一般知能（AGI）**に近づくための能力を試す意図があり、単なるパターンマッチでは解けない**未知のパズル的課題**が多数含まれています [9] [10] 。各課題はこれまでAIが学習してきた分野知識の組み合わせでは太刀打ちできないよう工夫されており、**推論や論理的思考、抽象概念の応用**が要求されます [10] 。

特徴的なのは**「トレーニングによる上達」がほぼ不可能**な点です [11] 。ARC-AGI-2では大量の訓練データによる過学習を避けるため、システムが事前に解法パターンを記憶できないよう常に**新規性の高い問題**が提示されます [10] 。要するに、その場で柔軟に**初見の問題を解く汎用性**が試されるのです。この性質上、「ベンチマークそのものを学習して攻略する（ベンチマックス）」ことはできず [12] 、純粋な**推論能力の高さ**と**試行錯誤の効率**がスコアを左右します。

ARC-AGI-2の**難易度**は非常に高く設定されており、最初期は**最先端モデルですら数%〜十数%**しか正解できませんでした [13] [14] 。2025年前半に行われたKaggle競技会でも、優勝チームのプライベートテストスコアは**24.0%（$0.20/問）**に留まっています [13] 。しかしそれから1年足らずで、OpenAIやGoogleなど各社の最新LLMがARC-AGI-2に挑戦し始め状況が一変しました [15] 。例えば**Anthropic社**の**Claude Opus 4.5 (Thinkingモード，64k)**は公式評価で37.6%（$2.20/問）に達し、**商用LLM単体**ではトップクラスの成績を収めています [16] 。それでも**平均的な人間**（約53〜60%）には遠く及ばず [17] [8] 、長らく「AIは人間には到底及ばない領域」と見做されてきました。

そうした中で登場したのが**Poetiq AI**による**“リファインメントループ（洗練反復）”**手法です [18] 。2024年にARC-AGI-1で兆しが見えたこの新潮流は、2025年には各所で開花し始めました [19] 。**リファインメント（改善の反復こそ知能である）**という情報理論的観点の下、**小さなモデルでも試行錯誤を重ねて性能を引き上げる**研究が相次ぎ、ARC-AGI-2はそれを競う場にもなったのです [18] 。Poetiqのアプローチはまさにこの文脈にあり、**AIの推論過程にフィードバックループを取り入れる**ことで大規模モデルの限界を突破しようとしています [20] 。その成果が、**2025年末までに一気に人間レベルを越え、75%に迫る**という急激な進歩に繋がりました [21] 。わずか**1ヶ月前には30%にも満たなかった**スコアが今や70%以上に達しているという指摘もあり [21] [22] 、専門家たちは「何が起きているのか」と驚きをもって注目しています。

Poetiqのメタシステム（ハーネス）による問題解決

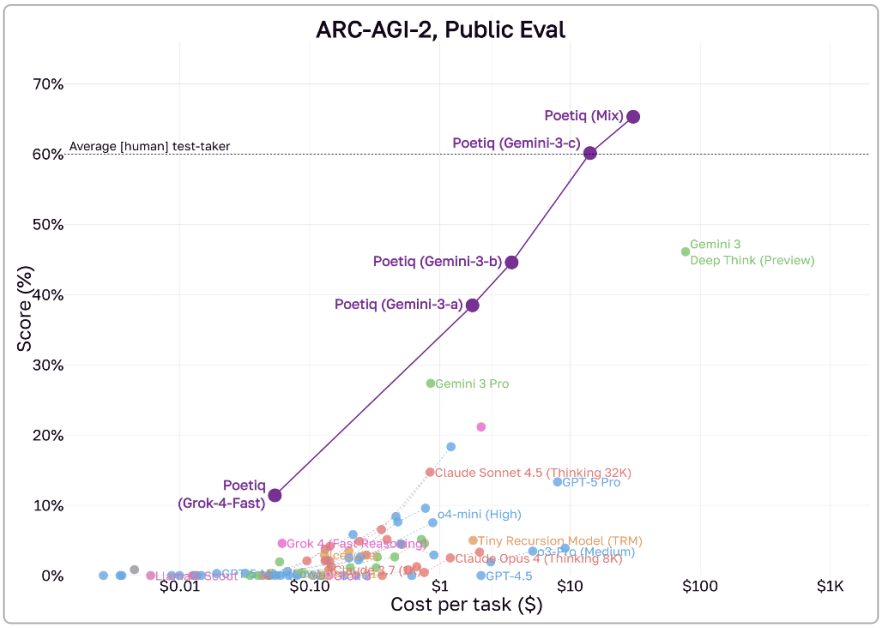
Poetiq　AIがARC-AGI-2で成果を上げている秘密は、**LLMを単独で使わず**に**複数のエージェントを組み合わせた特殊な解法フレームワーク**にあります。彼らは**「メタシステム」**あるいは**Poetiqハーネス**と呼ぶ自律的な問題解決エージェント群を開発しており、これが任意の最新モデル（GPTやGemini系列など）を土台にしてその能力を最大限に引き出す仕組みです [5] [23] 。

このシステムでは、一つの質問に対して**複数のAIエージェントが協調しながら解を探します** [24] 。具体的には、あるエージェントが仮の答えや解法プランを提案し、別のエージェントがその結果を検証・批評する、といった対話型のループを回します [24] 。まるで**人間がチャットボットに指摘と再試行を繰り返して正答へ導く**ようなプロセスを、AI同士で自動的に実行しているイメージです [24] 。さらに難解な場合には**プログラムの生成と実行（プログラム合成）**も活用し、問題をコードで試行錯誤して解決することもあります [25] 。

Poetiqチームの公開リポジトリを覗いた研究者によれば、その中では**「複数のエージェントを起動して同期させ、仮説→実装→検証を連続的にループし、解が得られるか上限回数に達するまで繰り返す」**という独自のアルゴリズムが実装されているとのことです [26] 。この手法によって、単一のLLMでは解けなかったり間違えたりする問題も、**自己修正を重ねて正解に辿り着ける可能性**が飛躍的に高まります [20] 。Poetiq自身、「我々のシステムは**あらゆる既存のフロンティアモデル（最先端モデル）の上に知能を構築**するアプローチ

であり、自らコードを書くかどうかや、どのモデルにどのタスクを割り振すべきかまで自動で決定できる」と説明しています [23]。いわば**LLMを部品として組み合わせ、“考えるためのAI”を組み上げる**ことで難問を解決しているのです。

このメタシステムの強みは**モデル非依存**である点です [5]。実際、Poetiqは少人数のチームながら、Gemini 3やGPT-5シリーズ、AnthropicのClaude、さらにはオープンソースの120億パラメータモデルに至るまで**十数種類のモデルで性能向上を実証**しています [27] [28]。同じ手法を適用するだけで、各モデルの精度を軒並み引き上げつつコストを削減できることが確認されました [28]。例えば**Anthropic Claude**や**xAI Grok**などのモデルにPoetiqメタシステムを適用したところ、**2回試行が許されるARC-AGIで平均「1回未満のリクエスト」**で問題を解いてしまう（=効率が倍増する）ケースも多かったといいます [28]。これはPoetiq流の高度な最適化戦略が、従来モデルの**「思考2回分を1回で凝縮する」**ような効果を生んでいることを示唆しています。



図：**ARC-AGI-2**公開評価における**性能–コスト曲線**。紫色の点がPoetiqシステム（Gemini-3やGPT-5.1を用いた構成）で、従来より大幅に高い正解率を**低コスト**で達成し、人間平均（黒破線60%）を上回っている [6]。緑の点は従来リーダーだった**Gemini 3 Deep Think**のプレビュー版成績で、約45%/数十ドルと劣る [6]。このようにPoetiqの手法は既存モデルの組み合わせと工夫だけで**性能の次元を引き上げ、費用を引き下げ**ている。

Poetiqのシステムは**自己改良**も特徴の一つです。各タスクを解決する過程そのものが学習データとなり、次のタスクに活かされるよう設計されています [29]。多様な課題を経験するほどメタシステム自体が強化されていくため、使えば使うほど解法戦略のストックが蓄積され賢くなるという循環です [29]。さらにPoetiqはこの技術を**特定タスクに閉じず汎用のAIコンポーネント最適化**に応用できると述べており、別分野のAIシステム内に組み込んで性能向上を図ることも視野に入れています [30]。現に、「Poetiqの技術はARC風の抽象推論以外の一般的なAIエージェントにも使えるのか？」という質問に対し、Poetiq側は前向きな見解を示しています [31]。このように**ハーネス（足場）を組んでAIに考えさせる**アプローチは2026年以降さらに注目されると予想され、業界では「来年はハーネスの年になる」といった声も上がっています [32]。

業界の反応と他社の取り組み

Poetiq　　AIの台頭は、AI研究コミュニティや産業界にも大きなインパクトを与えています。**小さなスタートアップがGoogleやOpenAIの巨大モデルを凌駕する成果**を上げたことは驚きをもって受け止められ [33]、「AGIへのブレークスルーは単純なモデル巨大化ではなく**新しいアルゴリズム的工夫**から生まれる可能性」を

示す例として注目されています [34]。この結果を受け、ベンチャーへの投資や異分野からの参入も活発化し、**より抽象的な汎用能力の探求**が業界全体で加速しています [34]。

具体的な他社動向としては、**OpenAI・Google DeepMind・Anthropic**など主要AI研究機関が相次いで**ARC-AGI-2対策の強化**を表明しました [35]。たとえばOpenAIはGPT-5.2のリリース時に**「論理推論やツール使用の強化」**を謳っていましたが、それでも単体では人間レベルの抽象推論に壁があることがARC-AGI-2で浮き彫りになっています [36]。このため、OpenAIは**モデルの巨大化だけでなく新手法（検証付き対話や自己反省機構など）の研究**に注力するとの見方があります [37]。実際、今回PoetiqがGPT-5.2を用いて結果を出した際には、OpenAIがモデル提供という形で協力しており [38]、自社モデルの潜在能力を第三者のハーネスで引き出す実験に前向きな姿勢もうかがえます。

Google DeepMindは、自社LLM群（Geminiファミリー）を活かして**合議制エージェント**の開発を進めていると報じられています。2025年11月には**Gemini 3 Pro**を活用した新しいエージェントAPI「**Deep Research**」を公開し、複雑なマルチステップのウェブ検索や推論を高精度かつ低コストで行えると発表しました [39]。これはまさに**Poetiq的なマルチステップ推論**を商用サービスに組み込む動きであり、Googleも社内ツールとしてARC-AGI-2に近い基準で性能評価を行っているようです [39]。さらにGoogleはARC-AGIとは別に、実世界の多段階リサーチタスクにAIがどこまで対応できるかを測る新ベンチマーク「**DeepSearchQA**」の提唱も始めています [39]。

Anthropicもまた、自社のClaudeシリーズに**“Thinking”**と称する長時間推論モードを搭載し、ARC-AGI-2のような難問に挑んでいます [16]。Claude Opus 4.5 (Thinking, 32Kや64K)は大規模コンテキストを活かして問題を考えさせる試みですが、先述のように公式スコアは40%未満で頭打ちでした [16]。Anthropicは現在、**合成データでの自己改善**や**外部ツール統合**による能力拡張を模索しており、ARC-AGIでPoetiqが示したような「モデルをラップする工夫」の重要性を認識しているようです [31]。

また、新興の研究機関や個人もARC-AGI-2突破に挑戦しています。**xAI（エックスAI）**のJeremy Berman氏は、自身の提案する**「進化的テスト時計算」**戦略で一時期Arc-AGI-2のパブリックリーダーボード上位に食い込んでいました [40] [41]。彼の手法はモデル自身を改変せずテスト段階で試行錯誤を繰り返す点でPoetiqと通じるものがあり、**オープンソース実装のGrok 4**などを駆使して健闘しました [40] [27]。しかし、Poetiqは**Gemini 3 Pro**や**GPT-5**など最強クラスのモデルも容易に取り込み、Berman氏の成績（例：Grok-4ベースで約30%台）を大きく引き離しています [27]。

総じて、ARC-AGI-2をめぐる競争は**「巨大モデル vs 賢い仕組み」**という構図になりつつあります。2025年末時点で、後者であるPoetiq流のアプローチが一歩リードし、各社がそれを追いかける形です [33] [34]。この動きは研究者の間でも「**汎用人工知能（AGI）実現へ向けたフロンティア**」として注目度が高く、ARC-AGI-2は単なる競技会を超えて**次世代AIの指針**とみなされています [42] [9]。

今後の展望: ARC-AGI-3と「ハーネスの時代」

ARC-AGI-2は、当初想定より遥かに早いペースで**人間水準突破**が達成されました。**初代ARC-AGI-1**は解かれるまでに約5年を要したのに対し [43]、ARC-AGI-2は開始から約1年で主要問題の**過半数をAIが解ける**ようになった計算です [43]。このため、主催のARC Prizeチームも早急に次の難関設定へ移行すべく動いています。公式には**「ARC-AGI-3は2026年第1四半期にリリース予定」**とされており [44]、現状の延長では歯が立たない新たな課題を準備中とのことです [45]。

ARC-AGI-3では、例えば**対話型の長期課題**や**マルチモーダルな推論**など、より実世界的でインタラクティブな問題が出題される可能性があります [45]。これは現在のPoetiqシステムのような静的ループに対し、新たなチャレンジを課す狙いがあるかもしれません。しかしPoetiqは「長い地平線のタスク（long horizon tasks）

であっても、既存モデルに内在する知識を引き出す工夫次第で解けるのではないか」と意気込みを見せており [46]、モデルを追加で訓練せずとも**新たな知識抽出メカニズム**で乗り切れる可能性を探っています [46]。

一方、**公式リーダーボードの行方**も注目です。ARC-AGI-2公式でPoetiqの新システム（GPT-5.2 X-High版）がどの程度のスコアを叩き出すかは、年明けにも明らかになるでしょう。仮に**60%台後半**に届けば、人間テスト参加者（平均53〜60%程度 [17]）を明確に上回ることになり、**「AIが人間越えを果たした汎用知能テスト」**として広く報じられる可能性があります。また、ARC Prizeには**グランドプライズ**（おそらく人類水準を大幅に超えるか100%近い達成への賞金）が設定されていますが、依然未受賞のままです [47]。Poetiqや他チームが80〜90%台に迫れば受賞条件に近づくかもしれません。もっとも運営側も難易度を上げてくるため、**真のAGI（あらゆる初見の問題を難なく解く存在）には未だ距離がある**とも言えます [47]。

今後数年は、**「モデルの知能を引き出す仕組み作り」**がAGI研究のキーワードとなりそうです。Poetiqの成功以降、業界ではこの種の**ハーネス型AI**への関心が急速に高まっています [32]。2026年は各社から類似の**推論オーケストレーション**技術が発表され、**AIシステム同士が対話しながら問題解決する**のが当たり前になるかもしれません [32]。たとえば、ChatGPTのような対話AIに裏で別のチェックAIが常駐し回答を検証・改善してくれるようなサービスが登場する可能性もあります。実際、AnthropicやOpenAIも既に**自己反省(Self-Reflection)**や**ツール使用**をモデルに組み込む試みを始めています [36]。

さらに先を見れば、**本当に汎用的な問題解決AI**が姿を現す可能性があります。ARC-AGIシリーズで培われた技術を現実世界の課題に適用すれば、今までAIでは解決困難だった複雑な計画立案・学術的発見・創造的設計などにも突破口が開けるでしょう。Poetiqは既に「**我々のメタシステムは複雑な実世界の問題にも適用可能**」と述べ、パートナー企業と応用研究を進めています [48]。わずか6名の精鋭チーム [49] ながら、大企業を出し抜く成果を上げた彼らの動向は、AGI開発競争において引き続き注目の的となるはずです。

参考文献・出典: Poetiq AI公式ブログやARC Prize公式発表、業界ニュース等 [2] [1] [6] [7] [33] など（本文中に引用箇所を示しています）。ARC-AGI-2のリーダーボードや技術詳細については公式サイト [50] および技術報告書 [17] も参照してください。今後の動きについてはPoetiq AIや主要AI企業の続報に注目しましょう。

---

[1] [3] [31] [32] [38] We finally had a moment to run our system with GPT-5.2 X-High on ARC-AGI-2! Using the same Poetiq harness as before, we saw results as high as 75% at under $8 / problem using GPT-5.2 X-High on the… | Shumeet Baluja
https://www.linkedin.com/posts/shumeetbaluja_we-finally-had-a-moment-to-run-our-system-activity-7409324615379943424-nU-O

[2] [12] [21] [22] [24] [25] [26] [43] [44] Poetiq Achieves SOTA on ARC-AGI 2 Public Eval : r/singularity
https://www.reddit.com/r/singularity/comments/1pu5mhk/poetiq_achieves_sota_on_arcagi_2_public_eval/

[4] [7] [29] [30] [46] [48] [49] Poetiq | ARC-AGI-2 SOTA at Half the Cost
https://poetiq.ai/posts/arcagi_verified/

[5] [8] [23] [27] [28] [40] [41] Poetiq | Traversing the Frontier of Superintelligence
https://poetiq.ai/posts/arcagi_announcement/

[6] [20] [39] Microsoft denies cutting AI sales targets - Sync #549
https://www.humanityredefined.com/p/sync-549?utm_source=substack&utm_medium=email&utm_content=share&action=share

[9] [10] [11] [14] [33] [34] [35] [36] [37] [42] ARC-AGI-2 results redefine AI benchmarks: GPT-5.2 advancements, startup surprises, and industry implications
https://www.datastudios.org/post/arc-agi-2-results-redefine-ai-benchmarks-gpt-5-2-advancements-startup-surprises-and-industry-impl

13 15 16 18 19 45 47 50 ARC Prize 2025 Results and Analysis
https://arcprize.org/blog/arc-prize-2025-results-analysis

17 ARC-AGI-2 human baseline surpassed (updated) — LessWrong
https://www.lesswrong.com/posts/DX3EmhmwZjTYp9PBf/ai-performance-has-surpassed-a-human-baseline-on-arc-agi-2