ChatGPT

Poetiq (Mix)モデルの性能と特徴

ARC-AGI-2ベンチマークとAIモデルの成績

ARC-AGI-2は、人間には容易だがAIには難しい推論課題によって、AIの汎用的な推論能力を測る新しいベンチマークです [1] 。ARC-AGI-2のテストで人間チームの平均正答率は約60%にも達しましたが、当初の最新AIモデルのスコアは1〜数%程度と極めて低く、人間との差が浮き彫りになりました [1] [2] 。例えばOpenAIの高性能モデル（ChatGPT系列）やAnthropicのClaudeシリーズなども軒並み一桁台〜20%未満の正答率に留まり、現行の大規模言語モデル（LLM）にとってARC-AGI-2は非常に困難な課題であることが示されています [2] 。このベンチマークではスコア（正答率）を縦軸、1タスクあたりの推論コストを横軸（対数スケール）にとった散布図で各システムの**知能効率**を評価します [3] 。コストを増やして推論時間を長くすることで性能向上が見込めるものの、多くのモデルでは計算資源の投入に対しスコアの頭打ち（**逆転現象**）が見られました [4] 。こうした背景から、大規模な計算に頼らず**新しい推論手法**を開発する必要性が指摘されています [4] 。

Poetiq (Mix)モデルとは何か

Poetiq（ポエティーク）は、LLMの上に独自の「メタシステム」を構築し、複数のモデルやステップを組み合わせて難問を解く先進的なAIシステムです [5] 。**Poetiq (Mix)**はその一構成で、**Gemini 3**（Google DeepMindの最新モデル）と**GPT-5.1**（OpenAIの最新モデル）という2つの最高峰LLMを組み合わせてタスクに挑みました [6] 。Poetiqチームは「どんな基盤モデルの上にも知能を構築できる」アーキテクチャを採用しており、新モデルが公開されてから数時間以内にGemini 3やGPT-5.1を統合することに成功したと述べています [7] 。この柔軟なメタシステムにより、それぞれのモデルの強みを活かし欠点を補い合う**混合戦略**が可能になっています。

人間平均スコアを超えた唯一のモデル

ARC-AGI-2の公開評価において、Poetiq（Mix)モデルは**約65%**という正答率を記録し、**人間平均（60%）を明確に上回る**唯一のAIシステムとなりました [8] 。リーダーボードの散布図上では、Poetiqモデル群（紫の線で連結されたポイント群）はコスト増大に伴いスコアが急激に向上しており、性能と効率の新たなパレートフロンティアを描いています [9] [10] 。特に最強構成のPoetiq (Mix)は高コスト帯で突出したスコアを示し、人間パフォーマンス（60%）を大きく凌駕しました [8] 。対照的に、GPTやClaudeといった他の主要LLMはどれも20%未満の低スコアにとどまり（中には一桁%のモデルもありました）、従来型のLLMにとってARC-AGI-2がいかに困難であるかを物語っています [2] 。この結果は、**高度な推論戦略**を導入することで初めて、人間レベルを超える汎用推論性能に達し得ることを示しています。

計算コストと性能向上の関係

ARC-AGI-2ではタスクごとの**計算コスト**を増やすことで一定の性能向上が期待できますが、その効果の度合いはモデルの**推論手法**によって大きく異なります [4] 。Poetiqモデルの場合、メタシステムによる柔軟な思考ループが功を奏し、コストを増やすほどスコアが劇的に向上しました [10] 。例えばPoetiqチームの分析では、比較的低コスト設定の構成 (Gemini-3-b) で既に高い性能を達成し、より長く「考えさせる」高コスト構成 (Gemini-3-c) ではARC-AGI-2でさらにスコアが伸び続けたと報告されています [10] 。一方、従来のLLMではコストをかけても効果が頭打ちになるケースが多く見られました [4] 。実際、ARC Prize財団の研究では「どれだけ多くの計算資源を投入しても正答率が伸びない」逆転現象が指摘されており [4] 、単純にモデルを大きく

したり推論時間を延ばすだけでは汎用的な知能には直結しないことが示唆されています。**Poetiq (Mix)**の成功は、計算コストの増加を最大限に生かすには洗練された推論アルゴリズムが不可欠であることを強調しています。

Poetiqメタシステムの推論手法

Poetiqの革新的な点は、**LLMを道具として使う再帰的な問題解決ループ**にあります [11]。単一のプロンプトで答えを引き出す従来手法と異なり、PoetiqシステムはまずLLMに仮の解答やコードを書かせ、それを自動的に実行・検証します [12]。得られたフィードバックを分析し、再びLLMに改良案を提案させる――この過程を繰り返すことで、解答を段階的に洗練させていきます [12]。システムは自ら**進捗を監査**し、十分な解が得られたと判断すればプロセスを打ち切るため、無駄な計算を避けコストを抑える工夫も備えています [13]。さらにこのメタシステムは**モデル非依存**で柔軟性が高く、複数のLLMを組み合わせる最適な戦略を自動で選択します [14]。例えば「どの部分問題を解くのにどのモデルを使うか」「必要ならコードを書いて計算させるか」といった判断もシステム自身が行います [14]。Poetiq (Mix)ではその能力を最大限に活用し、Gemini 3とGPT-5.1という異なるモデルを適材適所で活かすことで、人間超えのスコア達成に至りました [6]。Poetiqチームはこの手法を**LLM上で知能を構築する**アプローチと表現しており、実際に小規模なチームで最先端モデルを即座に統合しARC-AGIでの記録的性能を達成しています [15]。なお、このPoetiqメタシステムのコードや具体的な構成（Poetiq (Mix)を含む複数の設定）はオープンソースで公開されており、再現や研究が可能となっています [16]。

まとめ：Poetiq (Mix)の意義

ARC-AGI-2ベンチマークにおけるPoetiq (Mix)の成果は、**計算コストを増やすだけでは不十分で、適切な推論戦略の採用が決定的**であることを示しました。Poetiq (Mix)は高度なメタ推論により限界を打ち破り、約65%というスコアで人間平均（60%）を上回ったのに対し [8]、他の最先端LLMは大規模モデルであっても20%に満たないスコアにとどまりました [2]。この**圧倒的な性能差**は、AIの知能を引き上げるにはモデル自体の巨大化や高額な計算資源以上に、**問題解決のプロセスを工夫すること**が重要であることを示唆しています [4]。Poetiqのアプローチによって、限られた試行回数でも効率良く知識を引き出し組み合わせることが可能となり、コスト対効果に優れた汎用AI推論の道が拓けることが証明されたと言えるでしょう [17] [18]。この成果はAGI（汎用人工知能）研究に新たな指針を与え、今後はPoetiqのような**自己改良型の推論システム**が、真に人間並みまたはそれ以上の柔軟な知能を実現する鍵になると期待されています。

参考文献: Poetiq公式ブログ [6] [8] [14]、Reddit発表 [7] [5]、Nazology解説記事 [2] [4] など。

---

[1] 汎用AI向け新テスト「ARC-AGI-2」で人間がAIに圧勝　AIの新たな課題が浮き彫りに | Plus Web3 media
https://plus-web3.com/media/_1607-250325arc/

[2] [4] 主要なAIモデルがAGIテストで全滅：汎用人工知能の高い壁 - ナゾロジー
https://nazology.kusuguru.co.jp/archives/173968

[3] ARC Prize - Leaderboard
https://arcprize.org/leaderboard

[5] [7] [15] [16] Poetiq Did It!!! Poetiq Has Beaten the Human Baseline on Arc-AGI 2 (<60%) | "Poetiq's approach of building intelligence on top of any model allowed us to integrate the newly released Gemini 3 and GPT-5.1 models within hours of their release to achieve the SOTA-results presented here." : r/accelerate
https://www.reddit.com/r/accelerate/comments/1p2grr3/poetiq_did_it_poetiq_has_beaten_the_human/

6 8 9 10 11 12 13 14 17 18 Poetiq | Traversing the Frontier of Superintelligence
https://poetiq.ai/posts/arcagi_announcement/