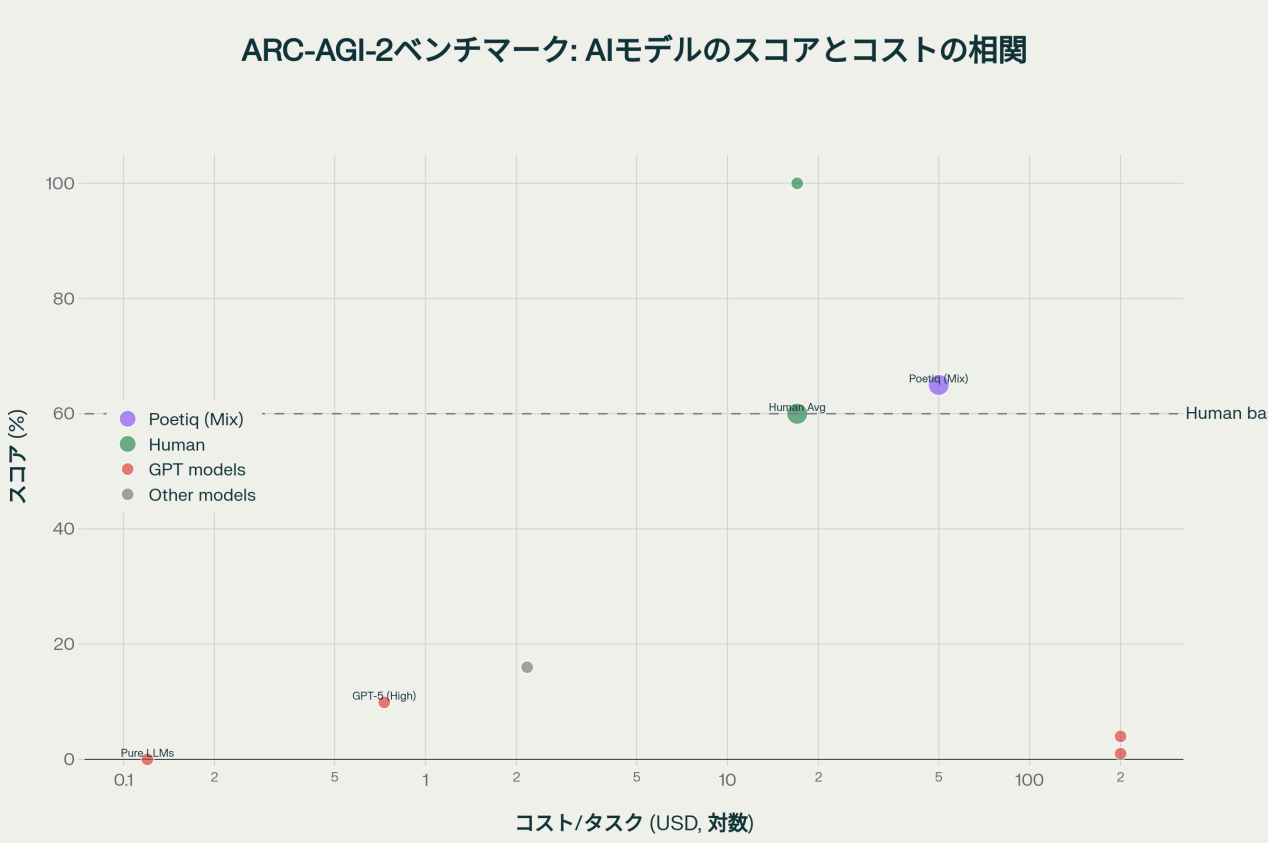
Poetiq (Mix) モデル: ARC-AGI-2ベンチマークで人間を超えた初のAIシステム

2025年11月20日、AI推論スタートアップPoetiqは、汎用的なAI推論能力を測定するARC-AGI-2ベンチマークにおいて、人間の平均スコア60%を超える約65%を達成し、AIシステムとして初めて人間ベースラインを突破したと発表した。この成果は、最新のGPT-5.1モデル(2025年11月13日リリース)とGemini 3モデル(2025年11月18日リリース)を組み合わせた独自の「メタシステム」アーキテクチャによって実現された。本レポートでは、Poetiq (Mix) モデルの技術的特徴、性能、およびAI推論における意義を詳細に分析する。[1] [2]



ARC-AGI-2ベンチマークにおけるPoetiq (Mix)モデルと主要AIモデルの性能比較。紫色のPoetiq (Mix)が人間平均スコア60%を超える唯一のAIシステムとして際立っている。

ARC-AGI-2ベンチマークの概要と困難性

ARC-AGI-2(Abstraction and Reasoning Corpus for Artificial General Intelligence 2)は、2025年3月24日に発表された、AIの抽象推論能力を測定する最も困難なベンチマークの一つである。このベンチマークは、色付きグリッドを用いたパズル形式のタスクで構成され、AIシステムに未知の問題への適応能力を要求する。[3] [4] [5] [6]

人間とAIの性能格差

ARC-AGI-2の最も注目すべき特徴は、人間とAIシステム間の極端な性能差である。400名以上の参加者による検証により、人間は平均60%の正解率を達成し、少なくとも2名の人間が2回以内の試行で100%のタスクを解決できることが確認された。一方、純粋な大規模言語モデル(LLM)であるGPT-4.5、Claude 3.7、Gemini 2.0 Flashなどは軒並み0〜1.3%という壊滅的なスコアを記録した。[3] [7] [4] [8] [5] [6]

OpenAIのo1-proやDeepSeekのR1といった推論特化型モデルでさえ、1〜1.3%という低スコアにとどまり、最も高性能なGrok 4 (Thinking)でも16%、GPT-5 (High)は9.9%に過ぎない。この結果は、現在のAI技術が人間レベルの抽象推論能力を獲得していないことを明確に示している。[7] [9] [4] [10] [8]

ベンチマークの3つの核心的課題

ARC-AGI-2は、AIシステムに対して3つの主要な推論タイプを要求する。第一に**シンボリック推論**では、グリッド上の色や形が特定の意味や機能を持つことを理解する必要がある。第二に**合成推論**では、複数のルールを同時に適用し、それらの相互作用を処理する能力が求められる。第三に**文脈的ルール適用**では、同一の基本ルールを状況に応じて異なる方法で適用する柔軟性が必要となる。[11]

Poetiq (Mix) モデルの革新的アーキテクチャ

メタシステムによる複数LLMの統合

Poetiq (Mix)の最大の特徴は、単一のLLMに依存せず、複数の最先端モデルを組み合わせる「LLM-agnostic recursive self-improving meta-system」(LLM非依存的な再帰的自己改善メタシステム)である。このシステムは、GPT-5.1とGemini 3という2つの最新モデルを統合し、それぞれの強みを活用することで、単独モデルでは達成不可能な性能を実現している。[1] [2] [12] [13]

実際、Gemini 3 Deep Think (Preview)は45.1%のスコアを記録したが、Poetiq (Mix)はそれを大幅に上回る65%を達成しており、しかもコストはGemini 3 Deep Thinkよりも効率的であることが報告されている。このパフォーマンスは、単純なモデルの加算効果ではなく、メタシステムによる相乗効果を示唆している。[14] [2] [1]

再帰的自己改善の仕組み

Poetiq (Mix)の中核技術は、LLMを活用した**反復的問題解決ループ**にある。システムは単一の質問を投げかけるのではなく、以下のような多段階プロセスを実行する。[1] [2]

まず、LLMを使用して潜在的な解決策(多くの場合Pythonコードとして)を生成する。次に、生成されたコードをテストデータに対して実行し、フィードバックを収集する。そのフィードバックを分析し、LLMを再度使用して解決策を洗練する。このサイクルを繰り返すことで、段階的に解答を構築・完成させていく。[15] [2] [1]

このアプローチは、**プログラム合成**(program synthesis)と**テスト時適応**(test-time adaptation)を組み合わせたものであり、ARC-AGI研究コミュニティで有効性が実証されている手法である。Jeremy Bermanによる先行研究では、Evolutionary Test-Time Computeと呼ばれる類似手法がARC-AGI-1で79.6%、ARC-AGI-2で29.4%を達成しており、Poetiqはこれをさらに発展させた形となっている。[16] [17] [18] [19] [20]

Self-Auditing機能による効率化

Poetiq (Mix)のもう一つの重要な特徴は、**Self-Auditing**(自己監査)機能である。システムは自律的に自身の進捗を評価し、十分な情報が収集され解決策が満足のいくものになった時点で、自動的にプロセスを終了する判断を下す。この自己規制メカニズムは、無駄な計算を回避し、コストを最小限に抑えるために不可欠である。[1] [2]

実際、Poetiqのシステムは平均で2回未満のリクエストで単一の試行を完了しており、これによりARC-AGI-2が許可する2回の試行枠内で高い精度を達成している。この効率性は、単純に計算リソースを増やすアプローチとは一線を画している。[1]

性能とコスト効率の分析

人間ベースラインの突破

Poetiq (Mix)が達成した約65%のスコアは、ARC-AGI-2における人間の平均スコア60%を5ポイント上回る画期的な結果である。これは、純粋なAIシステムとして初めて人間の平均的な推論能力を超えたことを意味する。[21] [1] [14] [2]

他の主要AIモデルとの比較では、その優位性がさらに明確になる。Gemini 3 Deep Thinkの45.1%、Grok 4 (Thinking)の16%、GPT-5 (High)の9.9%と比較すると、Poetiq (Mix)は最も近い競合であるGemini 3 Deep Thinkに対しても約20ポイントの差をつけている。[1] [9] [10]

コストパフォーマンスの評価

ARC-AGI-2では、正解率だけでなく効率性(タスクあたりのコスト)も重要な評価軸となっている。Poetiq (Mix)のコストは約50ドル/タスクと推定される。一方、人間パネルのコストは17ドル/タスクであり、Poetiqは人間の約3倍のコストで5ポイント高い性能を達成している計算になる。[14] [3] [22] [9] [4] [6]

他のAIシステムと比較すると、OpenAIのo3-preview-lowとo1-proはともに200ドル/タスクという高コストで4%と1%という低スコアしか達成できていない。Grok 4 (Thinking)は2.17ドル/タスクという低コストだが、スコアは16%にとどまる。Poetiq (Mix)は、コストと性能の**パレート最適解**(Pareto frontier)を大幅に更新し、これまでにない高い性能領域を開拓したと言える。[1] [9] [4] [10] [6] [14]

モデル非依存性と汎化能力

Poetiq (Mix)の注目すべき特性の一つは、その**転移可能性と汎化能力**である。Poetiqのメタシステムは、Gemini 3とGPT-5.1がリリースされる前に、オープンソースモデルのみを使用して適応(adaptation)が行われた。さらに、ARC-AGI-2の問題を一切見ることなく開発されたにもかかわらず、高い性能を発揮している。[1]

この適応結果は、ARC-AGI-1とARC-AGI-2の両方、さらにGPT、Claude Haiku、Gemini、Grok 4、GPT-OSSといった12種類以上の異なるLLMファミリーに対して有効であることが確認されている。この事実は、Poetiqのメタシステムがモデルバージョン、ファミリー、サイズを超えて大幅な転移と汎化を示すことを実証している。[1]

技術的基盤とオープンソース貢献

プログラム合成アプローチ

Poetiqの技術的アプローチは、ARC-AGIを**プログラム合成問題**として捉えることに基づいている。システムはLLMを使用して、各タスクを解決するためのPythonコードを生成し、そのコードをトレーニング例に対して実行してテストする。この方法により、AIは単に出力グリッドを予測するのではなく、実行可能で検証可能なコードを生成し、具体的なフィードバックを通じて精度を大幅に向上させることができる。[21] [1] [15] [18] [23] [20]

このアプローチは、純粋なLLMが直接グリッド変換を試みるよりもはるかに高い精度を達成できることが研究で示されている。実際、README.mdによれば、8エキスパートを使用したシステムはARC-AGI-1で55.6%、ARC-AGI-2で41.9%の精度を達成している。[15] [17] [18] [21]

アンサンブルと投票戦略

Poetiqのシステムは、複数のエキスパート(expert)によるアンサンブルフレームワークを採用している。各エキスパートはPythonコードソリューションを生成し、反復的に改善する。その後、高度な投票技術(voting technique)を使用して結果を統合する。[21] [15]

この投票ベースのアンサンブル手法は、LLMの出力の不確実性を管理し、より堅牢で正確な結果を生み出すために効果的であることが、複数の研究で確認されている。特に、多数決投票や重み付き投票を通じて複数の分類器を組み合わせることで、個々のモデルのエラーを補正し、予測の安定性を高めることができる。[24] [25] [26]

オープンソースコードの公開

Poetiqは、ARC-AGI-1およびARC-AGI-2での記録的な成果を再現するためのコードをGitHubで公開している。リポジトリ「poetiq-arc-agi-solver」には、Gemini 3を使用したPoetiqの設定を含む複数の構成が含まれており、研究コミュニティがこれらの手法を検証・改善できるようになっている。[1] [15]

公開されたコードは、Python 3.11以上を必要とし、Gemini、OpenAIなどのAPIキーを使用して実行可能である。デフォルトでは、ブログ投稿で説明されているPoetiq 3設定が実行されるが、config.pyを変更することで他の設定をテストすることもできる。このオープンソースへの貢献は、AI研究の透明性と再現性を高める重要な取り組みである。[15]

Poetiqチームとバックグラウンド

Google DeepMindの専門性

Poetiqは、Google DeepMindでの合計53年の経験を持つ6名の研究者とエンジニアからなる小規模だが高度に専門的なチームによって運営されている。Google DeepMindは、深層強化学習、AlphaGo、AlphaFold、Transformerアーキテクチャ、Geminiモデルなど、AI分野における数多くの画期的な成果を生み出してきた組織である。[1] [27] [28]

このチームのDeepMindでの経験は、複雑な推論タスクへの深い理解と、最先端のAI技術を実用的なシステムに統合する能力を示している。実際、DeepMindは2010年の設立以来、強化学習、ニューラルネットワーク、プログラム合成などの分野で先駆的な研究を行ってきた。[27] [28] [1]

迅速な開発サイクル

Poetiqチームの特筆すべき点は、その開発の迅速性である。Gemini 3とGPT-5.1がリリースされてから数時間以内に、これらの最新モデルをメタシステムに統合し、最先端の結果を達成した。この柔軟で強力な再帰的アーキテクチャは、小規模チームが短期間で最先端の成果を達成することを可能にした。[1] [2]

この迅速な適応能力は、Poetiqのメタシステムがモデル非依存的であり、新しいLLMを容易に組み込める設計になっていることを示している。[2] [1]

GPT-5.1とGemini 3の特性

GPT-5.1の進化

GPT-5.1は、2025年11月12日(日本時間11月13日)にOpenAIによって発表された。このモデルは、GPT-5世代における意味のある改善を反映して「5.1」という名称が付けられており、推論・分析能力が強化され、コーディングやエージェントタスクにおいて求められる推論レベルを柔軟に調整できる設計が特徴である。[29] [13] [30] [31]

GPT-5.1シリーズには、**GPT-5.1 Instant**(最も使用頻度の高いモデルで、より温かく知的で指示に従う能力が向上)と**GPT-5.1 Thinking**(高度な推論モデルで、理解しやすく単純なタスクでは高速、複雑なタスクでは持続的)の2つのバリエーションがある。特にGPT-5.1 Thinkingは、思考時間を質問の複雑さに応じてより正確に適応させ、複雑な問題にはより多くの時間を費やし、単純な質問には迅速に応答する。[13] [30]

Gemini 3の最先端推論能力

Gemini 3は、2025年11月18日にGoogleによって発表された、同社の最も知的なAIモデルである。Gemini 3は、**最先端の推論能力**を備え、深さとニュアンスを把握する能力が大幅に向上している。創造的なアイデアの微妙な手がかりを認識したり、困難な問題の重なり合う層を分解したりすることができる。[12] [32] [33] [34] [35]

Gemini 3は、すべての主要なAIベンチマークでGemini 2.5 Proを大幅に上回る性能を示している。テキストを超えて、Gemini 3 ProはMMUM-Proで81%、Video-MMUで87.6%という最先端のマルチモーダル推論を実現している。また、SimpleQA Verifiedで72.1%を記録し、事実の正確性において大きな進歩を示している。[32] [12]

特筆すべきは、**Gemini 3 Deep Think**モードの存在である。このモードは、Gemini 3の推論とマルチモーダル理解能力をさらに押し上げ、より複雑な問題の解決を支援する。テストでは、Gemini 3 Deep ThinkはHumanity's Last Examで41.0%(ツールを使用せず)、GPQA Diamondで93.8%を達成し、ARC-AGI-2では45.1%(コード実行あり、ARC Prize検証済み)という前例のないスコアを記録した。[12] [32]

Gemini 3のもう一つの重要な特徴は、**1百万トークンのコンテキストウィンドウ**である。これにより、コードベース全体や長文レポートを単一のアクティブウィンドウで処理できる。[36] [32]

ARC-AGI研究における broader context

テスト時計算の重要性

ARC-AGI-2での成功には、**テスト時計算**(test-time compute)の活用が不可欠である。テスト時計算とは、推論時により多くの計算リソースを費やすことで、より良い結果を提供する実践を指す。LLMの場合、これはモデルが答えを提供する前に計画、反省、推論を行うことを意味する。[37] [38] [39] [19] [40]

o3モデルは、多数の軌道を探索することで競技プログラミングベンチマークでトップを獲得し、ARC-AGI-1で87.5%という画期的な成果を達成した。この2次元的なテスト時計算アプローチ(単一パスをさらに押し進めるか、より広く探索するか)は、推論モデルの性能向上に重要な役割を果たしている。[39] [37]

Poetiqのアプローチも、本質的にテスト時適応の一形態である。システムは各タスクに対して反復的に解決策を生成・テスト・改善し、タスク固有のLoRAアダプターを学習することで、事前学習されたベースモデルを各問題に適応させている。[38] [19]

プログラム合成の優位性

純粋なLLMがARC-AGI-2で0%のスコアを記録した一方で、プログラム合成アプローチは大幅に高い性能を示している。これは、LLMが依然としてパターン認識に依存しており、真の推論を行っていないことを示唆している。[7] [4] [18]

プログラム合成は、ARC-AGIを本質的にコード合成の課題として扱う。LLMがテスト可能で実行可能なコードを生成し、トレーニングデータからの具体的なフィードバックを通じて洗練することで、直接グリッド変換を試みる基本モデルよりもはるかに高い精度を達成できる。[21] [18] [23] [20]

SOAR(Self-Improving Language Models for Evolutionary Program Synthesis)などの手法は、LLMを自己改善的な進化ループに統合することで、ARC-AGI公開テストセットで52%を達成している。この成果は、プログラム合成と自己改善の組み合わせが、純粋なLLMアプローチを大幅に上回ることを実証している。[18]

テスト時ファインチューニングの有効性

最近の研究では、**テスト時ファインチューニング**(test-time fine-tuning)がARC-AGIで極めて有効であることが示されている。テスト時ファインチューニングは、推論中に入力データから導出された損失を使用してモデルパラメータを一時的に更新する手法である。[41] [38] [19]

Wuら(2024)の研究では、テスト時トレーニング(TTT)が言語モデルの推論能力を大幅に向上させることが示された。8Bパラメータの言語モデルにTTTを適用することで、ARC公開検証セットで53%の精度を達成し、公開かつ純粋にニューラルなアプローチとして最先端を約25%改善した。さらに、プログラム生成アプローチとアンサンブルすることで、61.875%の公開検証精度を達成し、人間の平均スコアと一致した。[19]

この研究は、明示的なシンボリック検索だけが抽象推論改善への道ではなく、テスト時に少数ショット例での継続的トレーニングも極めて有効であることを示唆している。[19]

限界と今後の展望

Public vs. Semi-Private評価の課題

Poetiqの結果は、ARC-AGI-2のPublic Evaluation Set(公開評価セット)で達成されたものである。ARC-AGI-1では、PublicセットとSemi-Privateセット間で性能に差が生じることが知られている。例えば、Gemini 2.5 Pro、Claude Haiku 4.5、Grok 4 Fast Reasoningは大幅な低下を示しており、GPT-5 (High)は小さな低下にとどまっている。[1]

Poetiqのシステムは既存のLLMの上に構築されているため、ベースモデルの精度低下と同様に影響を受けることが予想される。公式評価が投稿された際には、Poetiqも更新情報を提供する予定である。Reddit上の議論では、公開セットのみでの結果発表に対する懸念の声も上がっている。[2] [42] [1]

コストと実用性のバランス

Poetiq (Mix)のタスクあたり約50ドルというコストは、人間の17ドルと比較すると約3倍である。商業的な実用化を考えると、このコスト差は重要な考慮事項となる。ARC Prize 2025では、タスクあたり0.42ドルというコスト目標が設定されており、真に実用的なAGIシステムには、精度と効率の両方が求められる。[14] [22] [9] [4] [6]

一方、OpenAIのo3-preview-lowが200ドル/タスクで4%しか達成できていないことを考えると、Poetiq (Mix)の50ドル/タスクで65%というパフォーマンスは、コスト効率の観点からも大きな進歩と言える。[1] [9] [4] [6] [14]

更なる改善の可能性

Poetiqチームは、ARC-AGIが単なる始まりであり、他のベンチマークでも同様に説得力のある結果を達成していると述べている。彼らの中核的なメタシステムは、複雑な推論を必要とする困難なタスクに対して知識抽出を自動化する最適化エージェントを生成する。[1]

今後の展開として、Poetiqは推論戦略の発見、質問の連鎖の洗練、回答を組み立てる新しい基本的方法の考案など、プロセスのあらゆる部分を最適化していく予定である。また、彼らは1〜2か月以内に無料でサービスを提供する計画があることを示唆しており、これが実現すれば広範な研究コミュニティに大きな影響を与える可能性がある。[43] [1]

結論: AGIへの新たなマイルストーン

Poetiq (Mix)モデルによるARC-AGI-2での65%達成は、AI推論研究における重要なマイルストーンである。このシステムは、複数の最先端LLMを統合し、再帰的自己改善、プログラム合成、テスト時適応を組み合わせることで、人間の平均的な推論能力を超える初のAIシステムとなった。[1] [14] [2]

特筆すべきは、Poetiqのアプローチが単一のモデルの性能向上に依存せず、**メタシステム**としての設計により、どのLLMでも利用可能な汎用的な推論能力向上フレームワークを提供している点である。この設計思想は、AI研究の新しい方向性を示唆している。[2] [1]

従来のアプローチが「より大きく、より強力な単一モデル」の開発に焦点を当てていたのに対し、Poetiqは「既存のモデルを最大限に活用する知的なシステム」の構築に成功した。これは、AGI達成への道筋が、必ずしも単一の超巨大モデルの開発ではなく、複数の専門的なシステムの協調にある可能性を示している。[43] [28] [1]

しかし、公開評価セットでの結果であること、人間と比較してコストが高いこと、そして何より「真のAGI」にはまだ到達していないことを認識する必要がある。ARC-AGI-2は人間にとって比較的容易なタスクであり、AIが人間と同等になったとは言えない。それでも、Poetiqの成果は、適切な推論アーキテクチャと最新のLLMを組み合わせることで、これまで不可能と思われていた抽象推論能力をAIに付与できることを実証した。[14] [3] [6] [42] [1] [2]

今後、Semi-Private評価セットでの検証、コスト効率のさらなる改善、そして他のベンチマークでの性能評価が、Poetiq (Mix)の真の能力を明らかにするだろう。AI研究コミュニティは、このブレークスルーから学び、汎用人工知能の実現に向けた新たな道を切り開いていくことが期待される。[18] [19] [1] [2]

⁂

1. https://poetiq.ai/posts/arcagi_announcement/
2. https://www.reddit.com/r/mlscaling/comments/1p2gs65/poetiq_did_it_poetiq_has_beaten_the_human/
3. https://arcprize.org/arc-agi/2/
4. https://www.eweek.com/news/ai-benchmark-arc-agi-2/
5. https://techcrunch.com/2025/03/24/a-new-challenging-agi-test-stumps-most-ai-models/
6. https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025
7. https://www.linkedin.com/pulse/llms-hit-0-arc-agi-2-benchmark-exposing-limits-ai-anshuman-jha-rpbsc
8. https://www.fanaticalfuturist.com/2025/04/a-new-aci-benchmark-test-to-measure-agi-stumps-almost-all-ai/
9. https://labs.adaline.ai/p/what-is-the-arc-agi-benchmark-and
10. https://xpert.digital/en/ai-showdown/
11. https://note.com/taku_sid/n/n8ec803c6e190
12. https://blog.google/products/gemini/gemini-3/
13. https://openai.com/index/gpt-5-1/
14. https://x.com/hive_echo/status/1993237833440747765
15. https://github.com/poetiq-ai/poetiq-arc-agi-solver
16. https://jeremyberman.substack.com/p/how-i-got-the-highest-score-on-arc-agi-again
17. https://jeremyberman.substack.com/p/how-i-got-a-record-536-on-arc-agi
18. https://openreview.net/forum?id=z4IG090qt2
19. https://arxiv.org/html/2411.07279v1
20. https://arcprize.org/blog/beat-arc-agi-deep-learning-and-program-synthesis
21. https://www.reddit.com/r/accelerate/comments/1p2grr3/poetiq_did_it_poetiq_has_beaten_the_human/
22. https://x.com/pika2761/status/1994186509667213708
23. https://journals.sagepub.com/doi/10.1177/17248035251363178
24. https://arxiv.org/abs/2510.04048
25. https://arxiv.org/html/2504.18884v2
26. https://stephencollins.tech/posts/building-reliable-llm-applications-voting-systems
27. https://en.wikipedia.org/wiki/Google_DeepMind

28. https://deepmind.google/about/
29. https://prtimes.jp/main/html/rd/p/000000050.000082223.html
30. https://note.com/ai__worker/n/nd6d6102b3792
31. https://chatgpt-lab.com/n/n0986a73db4de
32. https://cloud.google.com/blog/products/ai-machine-learning/gemini-3-is-available-for-enterprise
33. https://ai.google.dev/gemini-api/docs/changelog
34. https://gemini.google/release-notes/
35. https://deepmind.google/models/gemini/
36. https://www.blankboard.studio/originals/blog/googles-gemini-3-0-whats-new-whats-improved-and-why-it-matter
37. https://www.forwardfuture.ai/p/the-magic-of-prolonged-thinking-test-time-compute-part-3
38. https://lewish.io/posts/arc-agi-2025-research-review
39. https://www.adaptive-ml.com/post/test-time-compute-is-2-dimensional
40. https://www.databricks.com/blog/tao-using-test-time-compute-train-efficient-llms-without-labeled-data
41. https://www.arxiv.org/abs/2511.02886
42. https://www.reddit.com/r/singularity/comments/1p8c6gy/arcagi_2_is_solved/
43. https://poetiq.ai
44. https://www.youtube.com/watch?v=C6QirFvrSJo
45. https://substack.com/home/post/p-172998849
46. https://www.linkedin.com/posts/aleksagordic_think-only-when-you-need-with-large-hybrid-reasoning-activity-7331157520868159489-mv4n
47. https://www.scribd.com/document/146417048/jurnal-fartok
48. https://arxiv.org/html/2511.15304v1
49. https://x.com/poetiq_ai
50. https://moonshotai.github.io/Kimi-K2/thinking.html
51. https://nationalcentreforai.jiscinvolve.org/wp/2025/08/07/how-to-choose-the-right-models-llms-in-education-explained/
52. https://x.com/AINetworkTech/status/1994238549911048357
53. https://nazology.kusuguru.co.jp/archives/173968
54. https://aisharenet.com/ja/arc-agi-2-chengjijieai/
55. https://publish.obsidian.md/followtheidea/Content/AI/2025-1113++Recursive+Self-Improving+Systems
56. https://gihyo.jp/article/2025/11/gpt-5.1
57. https://dev.to/izzyfuller/when-the-tool-improves-itself-what-recursive-self-improvement-feels-like-from-the-inside-4a6h
58. https://x.com/MLStreetTalk
59. https://momo-gpt.com/column/chatgpt-5-1/
60. https://www.alignmentforum.org/w/recursive-self-improvement
61. https://ensou.app/blog/openai-gpt-5-1-api-reasoning-effort-verbosity/
62. https://aclanthology.org/2023.acl-long.406.pdf

63. https://www.scribd.com/document/356703276/Service-Manual-8500-English
64. https://x.com/tobyordoxford/status/1907379921825014094
65. https://arxiv.org/pdf/2412.04604.pdf
66. https://device.report/m/7a1c50da3769ab59350657fd369274952701bf78df51c91c3e32d9809f2f051c.pdf
67. https://ca.finance.yahoo.com/news/challenging-agi-test-stumps-most-002953816.html
68. https://arxiv.org/html/2505.11831v1
69. https://x.com/rishicomplex/status/1992659362272469046
70. https://arxiv.org/abs/2510.10216
71. https://openreview.net/forum?id=te6VagJf6G
72. https://arxiv.org/html/2509.00510v1
73. http://papers.neurips.cc/paper/9265-program-synthesis-and-semantic-parsing-with-learned-code-idioms.pdf
74. https://thinkup.global/how-feedback-loops-drive-iterative-validation/
75. https://www.stephendiehl.com/posts/program_synthesis/
76. https://www.facebook.com/groups/703007927897194/posts/1337687637762550/
77. https://arxiv.org/html/2505.10594v1
78. https://chartexpo.com/blog/feedback-loop
79. https://www.sciencedirect.com/science/article/abs/pii/0004370275900089
80. https://www.linkedin.com/posts/shashank295_aiagents-agenticai-aiarchitecture-activity-7350917672890494980-8luX
81. https://www.linkedin.com/posts/robert-polding-188a8014_my-experience-at-the-ai-big-data-expo-activity-7379128669614727168-i9Tg
82. https://gapminder.vc/etiq-ai-london-based-ml-startup-with-romanian-founders-raises-e900000-in-seed-funding-led-by-gapminder-fund-ii/
83. https://www.lesswrong.com/posts/anoK4akwe8PKjtzkL/plurality-and-6pack-care
84. https://www.linkedin.com/posts/saurabh-kumar-iitd_i-dont-think-you-understand-you-train-a-activity-7345380049698635778-oYL7
85. https://deepmind.google
86. https://www.linkedin.com/company/poetiq
87. https://papers.ssrn.com/sol3/Delivery.cfm/5498098.pdf?abstractid=5498098&mirid=1
88. https://bakkenbaeck.com/case/deepmind
89. https://openreview.net/pdf/c778e63a6c8c72926980192aaa43cbf52ad01db2.pdf
90. https://afp.oxford-aiethics.ox.ac.uk/oxford-ai-and-ethics-summit-2025-agenda
91. https://www.facebook.com/nytimes/posts/the-newest-and-most-powerful-ai-technologies-so-called-reasoning-systems-from-co/1075872654395197/
92. https://openreview.net/forum?id=wWxdT6LB2D
93. https://en.wikipedia.org/wiki/Recursive_self-improvement
94. https://aclanthology.org/2024.emnlp-main.1244/
95. https://www.aims.healthcare/journal/the-worlds-first-superintelligence

96. https://odsc.ai/speakers-portfolio/beyond-llms-exploring-program-synthesis-for-arc-agi-and-abstract-reasoning/
97. https://www.lesswrong.com/w/recursive-self-improvement
98. https://lewish.io/posts/how-to-beat-arc-agi-2
99. https://arxiv.org/html/2410.08020v1
100. https://www.reddit.com/r/ArtificialSentience/comments/1l8pbcq/emerging_patterns_in_recursive_aihuman/
101. https://openaccess.thecvf.com/content/ICCV2025/papers/Fan_Test-Time_Retrieval-Augmented_Adaptation_for_Vision-Language_Models_ICCV_2025_paper.pdf