Poetiq (Mix) メタシステムとARC-AGI-2の突破：推論時計算における新たなパラダイムと経済性に関する包括的調査報告書

Gemini 3 pro

エグゼクティブサマリー

2025年11月下旬、人工知能研究コミュニティは「Poetiq (Mix)」システムの発表により、汎用人工知能（AGI）に向けた進歩における潜在的な転換点を迎えました。GoogleのGemini 3やOpenAIのGPT-5.1といったフロンティア・ファウンデーションモデルのリリースと時を同じくして登場したPoetiqは、単体のモデルではなく「メタシステム」として設計され、ARC-AGI-2ベンチマークにおいて人間のベースラインスコアを打破したと主張されています。本報告書は、この画期的な開発に関する技術的詳細、達成されたスコア、そして推論時計算（Inference-Time Compute）の経済性について、徹底的な調査と分析を行うものです。

本調査の核心的な発見は、Poetiq (Mix)がARC-AGI-2のパブリック評価セットにおいて**65%**を超えるスコアを達成し、人間の平均スコアである**60%**、および競合する最先端（SOTA）モデルであるGoogleのGemini 3 Deep Thinkモードの約**45.1%**を大幅に上回ったことです[1]。極めて重要な点は、この性能が単なるモデルのパラメータ数の拡大によるものではなく、複数の大規模言語モデル（LLM）—具体的にはGemini 3とGPT-5.1—を、反復的なコード生成と自己修正の推論ループを通じて統合管理する「スキャフォールディング（足場）」アプローチによるアーキテクチャ上の革新に由来していることです[4]。

さらに、本報告書ではPoetiqのアプローチがもたらす経済的影響についても詳細に分析します。再帰的推論に伴う高い計算オーバーヘッドにもかかわらず、Poetiq (Mix)は新たなパレートフロンティアを確立しました。より精度の低いGemini 3 Deep Thinkが1タスクあたり推定**77**ドル～**80**ドルのコストを要するのに対し、Poetiq (Mix)は**20**ドル～**30**ドルという大幅に低いコストで優れた精度を実現しています[1]。この「低コストでの高性能化」という逆転現象は、AI開発の焦点が、巨大なモノリシックモデルのトレーニングから、既存の知能を効果的に活用するための高度な「システム2」メタ構造の開発へと移行しつつあることを示唆しています。

本書では、Poetiqメタシステムの技術的解剖、ARC-AGI-2という課題の文脈における性能評価、そしてAGI開発における「推論のサービス化（Reasoning as a Service）」の時代への移行について詳述し

ます。

---

1. イントロダクション：AGIへの「ラストワンマイル」とARC-AGI-2の文脈

Poetiq (Mix)が達成した成果の大きさを正確に評価するためには、まずその舞台となったベンチマーク、すなわちARC-AGI-2（Abstraction and Reasoning Corpus）の特異性と、それがAI研究において持つ意味を深く理解する必要があります。

1.1 ARC-AGI-1からARC-AGI-2へ：暗記の抜け穴を塞ぐ進化

François Cholletによって2019年に提唱されたオリジナルの ARC-AGI-1は、大規模言語モデル（LLM）が得意とする「暗記」能力に対抗するために設計されました [6]。MMLUやGSM8Kといった従来のベンチマークは、インターネット上の膨大なテキストやコードを事前学習に取り込むことでスコアを向上させることが可能ですが、ARCは独自の視覚的グリッドパズルを用い、「フューショット学習（少数の例から法則を学ぶ）」と「広範な一般化能力」を測定することを目的としています。モデルは、「青いピクセルが中心に移動する」といった抽象的な変換ルールを少数の例から推論し、未知のテスト入力に適用しなければなりません。

しかし、2024年後半までに、ARC-AGI-1のスコアは「テスト時適応（Test-Time Adaptation）」手法の進化や、トレーニングデータへの類似パターンの混入により飽和し始めました [7]。これに対応するため、ARC Prize Foundationは2025年に**ARC-AGI-2**を立ち上げました。この新しいイテレーションは、以下の3つの重要な設計思想に基づいています。

1. **AIにとっての難易度向上：** ARC-AGI-2は、単なる視覚的なパターン認識を超え、記号に意味的な重要性を割り当てる必要がある「記号的解釈（Symbolic Interpretation）」を必要とする複雑なタスクを導入しました [6]。これにより、従来の対称性チェックや回転などの単純な変換では解けない問題が増加しました。
2. **人間にとっての容易性の維持：** AIにとっては飛躍的に難易度が上がった一方で、タスクは人間にとっては依然として直感的であり、「コア知識（Core Knowledge）」のみで解けるという制約が維持されています [7]。
3. **効率性への注力：** ARC-AGI-2は、単純なパズルを解くために天文学的な計算リソースを費やすブルートフォース（力任せ）なアプローチを排除し、効率性を重視しています [9]。

1.2 人間のベースライン：60%という閾値の意味

本分析において最も重要な指標の一つが「人間のベースライン」です。ARC Prize Foundationによる400人以上の人間を対象とした広範なテストの結果、ARC-AGI-2における人間の平均正答率は約**60%**であることが確立されました [6]。この数値は、単なる統計以上の深い意味を持っています。

第一に、この数値はタスクが「解可能」であり、ノイズや不適切に設定された問題ではないことを保証します。第二に、この60%というラインは、AIが「流動性知能（新しい状況に適応する能力）」において人間に並び、あるいは追い越す瞬間を定義する分水嶺となります。司法試験や医師国家試験のような「結晶性知能（知識の検索と適用）」に依存するテストとは異なり、ARCでの人間超えは、真の意味でのAGIへの到達を示唆する強力なシグナルとなります [7]。

2025年11月以前、純粋なLLMのスコアはほぼ**0%**であり、専門的な推論システムでさえ一桁台のスコアに留まっていました [8]。Gemini 3やOpenAIのoシリーズのような「システム2」思考モデルの登場により、ようやく30〜40%台への道が開かれましたが、Poetiqが主張する65%というスコアは、この漸進的な進歩のグラフにおける不連続な跳躍、すなわち「フェーズトランジション」を意味しているのです。

---

2. Poetiq (Mix)の技術的解剖学：メタシステムというアプローチ

Poetiqは、それが打ち負かしたとされるGoogleやOpenAIのファウンデーションモデルとは、根本的に異なるカテゴリに属します。それはペタバイト級のデータでゼロからトレーニングされたモデルではなく、既存のファウンデーションモデルの労働力を指揮・管理するメタシステム（**Meta-System**）、すなわち高度な認知実行アーキテクチャです。

2.1 スキャフォールディング（足場）としての知能

Poetiqのアプローチは「ソフトウェア・スキャフォールディングとしての知能」と表現するのが最も適切でしょう。その核心的な革新性は、モデルそのもののパラメータにあるのではなく、既存のモデルから推論能力を抽出・増幅する方法論にあります。

2.1.1 モデル非依存性とハイブリッド統合

「Mix」という名称が示す通り、Poetiq (Mix)は2025年11月時点で利用可能な最も強力な2つのモデルをハイブリッドに活用しています [4]。

- **Google Gemini 3**（**2025年11月18日リリース**）**:** そのマルチモーダル能力と、思考時間を長く取ることで推論を深化させる「Deep Think」モードで知られています [4]。
- **OpenAI GPT-5.1**（**2025年11月13日リリース**）**:** GPT-5シリーズの最新版であり、指示追従能力とコーディング能力の向上が特徴です [4]。

Poetiqのメタシステムは「LLM非依存（LLM-agnostic）」であり、下層の推論エンジンを自由に交換可能なモジュールとして扱います [5]。このモジュール性により、PoetiqのエンジニアはGemini 3とGPT-5.1がリリースされてからわずか数時間以内に、それらをシステムに統合し、SOTA（State-of-the-Art）の結果を出すことができました [5]。この際、システム自体が従来の勾配降下法による学習を行ったわけではなく、スキャフォールディングがこれらの新しく強力な「頭脳」にパズルを解くよう指示を与えたのです。

2.1.2 プログラム合成パラダイム：推論としてのコード

Poetiqは、モデルが自然言語で推論過程を出力する標準的な「思考の連鎖（Chain of Thought: CoT）」プロンプティングだけに依存しているわけではありません。プログラム合成（**Program Synthesis**）、具体的にはPythonコードの生成を中核に据えています [4]。

- 推論としてのコード**:** システムはARCのタスクを単なる視覚的なマッチング問題としてではなく、プログラミングの課題として扱います。目的は、入力グリッドを受け取り、正しい出力グリッドを生成するPython関数を作成することです。
- 決定論的検証**:** コードは決定論的な実行パスを提供します。一度ルールがPythonコードとして記述されれば、それを例題に対して厳密にテストすることができます。トレーニング例（Few-shot例）を正しく解くコードが生成できれば、それがテスト入力に対しても正解を出す確率は（確実ではないものの）飛躍的に高まります。

2.1.3 再帰的自己修正ループとアンサンブル

Poetiqメタシステムは、以下のような再帰的な問題解決ループを実装しています [2]。

1. ドラフト作成**:** システムは下層のLLM（例：Gemini 3）にプロンプトを送り、グリッドを分析させ、変換ロジックをPythonコードとして提案させます。

2. 実行とフィードバック: 生成されたコードはPythonインタプリタ上で実行され、例題の入力と照らし合わされます。
3. リファインメント（改良）: コードが失敗した場合（例：例題2で誤ったグリッドを出力した、あるいはエラーが出た）、そのエラートレースと視覚的な不一致情報が再びLLMにフィードバックされます。モデルは自身のコードをデバッグし、書き直すよう求められます。
4. 投票/アンサンブル: システムは、異なるモデル（Gemini 3とGPT-5.1）や異なるプロンプト戦略を用いて複数の候補プログラムを生成します。「洗練された投票メカニズム」がこれらの結果を集約し、最も堅牢な解を選択します [4]。

2.2「システム2」のシミュレーションと自己監査

認知科学では、直感的で速い「システム1」思考と、熟考的で遅い「システム2」思考が区別されます。標準的なLLMの推論はシステム1に近いものですが、Poetiqはコード生成と実行のループを通じてシステム2を外部的にシミュレートしています。

LLMに対し、単に答えを出力させるのではなく、「立ち止まって考えさせ（コードを書き）」、「検証させ（実行し）」、「修正させる」ことで、思考プロセスを強制的に深くしています。さらに、コスト管理において極めて重要なのが**自己監査（Self-Auditing）**機能です。システムは、十分な精度の解が得られたと判断した場合に自律的にプロセスを終了させることができます。これにより、無駄な計算リソースの浪費を防ぎ、高い経済効率（パレート効率）を実現しています [2]。

2.3 開発チームと背景：Google/DeepMindのDNA

この高度なシステムを開発したPoetiqチームの背景も、その信頼性を裏付ける重要な要素です。創業チームには、Yair Alon氏やSaurabh Singh氏など、Google ResearchやGoogle DeepMindでの勤務経験を持つ研究者が含まれており、彼らの合計経験年数は53年に及びます [15]。彼らの専門知識は、単なるプロンプトエンジニアリングを超えた、LLMの能力を最大限に引き出すためのシステムアーキテクチャ設計に反映されています。

---

3. 基盤エンジンの詳細分析：Gemini 3とGPT-5.1の能力

Poetiq (Mix)の成功は、そのメタシステムとしての設計だけでなく、それを駆動するエンジンの性能向上にも大きく依存しています。2025年11月に相次いでリリースされた2つのモデルは、それぞれが前

世代からの大幅な飛躍を遂げています。

3.1 Google Gemini 3：推論の新基準

2025年11月18日にリリースされたGemini 3は、GoogleのAIモデルにおける「最も知的な」バージョンと位置づけられています [12]。

- **Deep Thinkモード:** Gemini 3の最大の特徴は「Deep Think」モードです。これは、モデルが回答を出力する前に「思考」するための時間を長く取る機能であり、いわゆる「システム2」的推論をモデル内部でネイティブに行うものです [17]。
- ベンチマーク性能:
  - **GPQA Diamond:** 博士号レベルの科学知識を問うテストで**91.9%**（Deep Think使用時さらに向上）を記録 [12]。
  - **Humanity's Last Exam:** 難問奇問を含むベンチマークで、ツールなしで**37.5%**、Deep Thinkありで**41.0%**を達成 [12]。
  - **ARC-AGI-2:** 単体（Deep Thinkモード）で**45.1%**という、Poetiq以前の最高スコアを記録しています [19]。

3.2 OpenAI GPT-5.1：エージェント能力の進化

2025年11月13日にリリースされたGPT-5.1は、GPT-4の後継として、特にコーディングとエージェント的な自律動作において強化されています [13]。

- コーディング能力: GPT-5.1（およびその派生であるCodex-Max）は、ソフトウェアエンジニアリングタスクにおいて高い性能を示しており、これがPoetiqの「コードによる推論」アプローチにおいて重要な役割を果たしていると推測されます [13]。
- 推論性能: ARC-AGI-2におけるGPT-5.1単体のスコアは**17.6%**（Thinking Highモード）と、Gemini 3に比べて劣りますが [20]、そのコスト効率の良さと特定のタスクにおける指示追従性の高さが、Mixモデルの一部としての価値を高めています。

3.3 なぜ「Mix」なのか？

Poetiqが単一のモデルではなく「Mix」を選択した理由は、各モデルのバイアスを相殺し、探索空間を多様化するためと考えられます。Gemini 3が高度な視覚的推論を提供し、GPT-5.1が堅牢なコード生

成を行うといった分業、あるいは双方が異なるアプローチで解を生成し、投票によって正解率を高めるアンサンブル効果が、単一モデルでは到達できない65%というスコアを実現した鍵です。

---

4. パフォーマンス分析：65%の壁と人間超えの意味

Poetiq (Mix)が報告したARC-AGI-2における定量的成果は、AI研究において統計的に有意な逸脱を示しており、新たな基準を打ち立てました。

4.1 スコアの内訳と詳細比較

以下の表は、ARC-AGI-2における主要システムのスコアとステータスをまとめたものです。

システム名	**ARC-AGI-2** スコア	備考・ステータス	出典
Poetiq (Mix)	約 **65.9%**	未検証 / 自己報告 (SOTA)	1
人間平均 **(Human Average)**	**60.0%**	公式ベースライン	6
Gemini 3 Deep Think	**45.1%**	公式リーダーボード (暫定)	3
Opus 4.5 (Thinking, 64K)	**37.6%**	公式リーダーボード	3
Gemini 3 Pro	**31.1%**	公式リーダーボード	12
GPT-5.1 (Thinking)	**17.6%**	比較データ	20
従来の**LLM (GPT-4**	**< 5%**	ベースライン	8

など)			

分析:

- **SOTA**に対する優位性: Poetiq (Mix)は、次点のGemini 3 Deep Think（45.1%）に対し、約**20**ポイントもの差をつけています。通常、ARCのような高難易度ベンチマークにおける改善が一桁%刻みであることを考慮すると、これは劇的な飛躍です。
- 人間超え（**Superhuman Performance**）: 60%という人間の平均スコアを超えたことは、象徴的かつ実質的な意味を持ちます。これは、抽象的な視覚推論という、これまで生物学的知能の独壇場とされていた領域において、ハイブリッドAIシステムが平均的な人間よりも信頼できるパートナーになり得ることを示しています。
- **Deep Think**との比較: Gemini 3 Deep Thinkも45.1%と非常に優秀ですが、これは単一モデル内での思考連鎖によるものです。Poetiqの65%は、外部ツール（Python）による検証と、複数モデルの多様性を組み合わせることで、単一モデルの限界（幻覚や論理の飛躍）を補正した結果と言えます。

4.2 検証と懐疑論：パブリック評価の落とし穴

Poetiqの成果は目覚ましいものですが、科学的な慎重さを持って評価する必要があります。特にRedditなどのコミュニティでは、以下の点について議論が巻き起こっています。

- パブリック評価 **vs** プライベート評価: 報告された65%は「パブリック評価セット」に対するものです。ARC-AGIには、完全に非公開の「プライベート評価セット」が存在し、真の汎用性はそこで測定されます [14]。パブリックセットは公開されているため、意図的か否かに関わらず、システムがそのデータ分布に「過学習（Overfitting）」してしまうリスクがあります。
- ベンチマキシシング（**Benchmaxxing**）: 一部の批評家は、パブリックセットの結果だけで「AGIが解決された」と主張するのは、練習問題を暗記して満点を取った学生を天才と呼ぶようなものだと批判しています [14]。
- **Poetiq**の反論: これに対しPoetiqチームは、彼らのシステム適応（チューニング）はGemini 3やGPT-5.1のリリース前に行われ、かつオープンソースモデルのみを用いて行われたと述べています。つまり、ARC-AGI-2の問題自体はシステムに見せておらず、汎用的な推論メカニズムが機能した結果であると主張しています [2]。さらに、コードをオープンソース化することで、第三者による検証を可能にし、透明性を確保しています [24]。

---

5. 経済分析：推論コストと新たなパレートフロンティア

Poetiq (Mix)の真の革新性は、単なるスコアの高さ以上に、その経済効率性にあります。巨大モデ

ルの時代において、推論コストは実用化の最大の障壁です。Poetiqは「より高い知能は指数関数的に高いコストを要する」という従来の常識に挑戦しています。

5.1 タスクあたりのコスト：効率性のギャップ

利用可能なデータに基づき、ARC-AGI-2の1タスクを解決するためにかかるコストを比較分析します。

システム	手法	推定コスト/タスク	精度 (スコア)	効率比 (スコア/$)
Poetiq (Mix)	メタシステム + API	約 **$20.00 - $30.00**	**~65%**	**2.1 - 3.2**
Gemini 3 Deep Think	ネイティブ System 2	約 **$77.16 - $80.00**	**45.1%**	**0.5 - 0.6**
人間パネル	生物学的知能	約 **$17.00**	**60.0%**	**3.5**
Opus 4.5 (Thinking)	ネイティブ CoT	約 **$2.40**	**37.6%**	**15.6**
Gemini 3 Pro	ネイティブ推論	約 **$0.81**	**31.1%**	**38.4**

データ出典: [1]

洞察:

1. 「深い思考」の高コスト: PoetiqとGemini 3 Deep Thinkは、標準的な推論（$0.81）と比較して桁違いに高価です（$20以上）。これは、1つのパズルを解くために、何百行ものコードを生成し、シミュレーションを実行し、数百万トークンを処理するという「推論時計算（Inference-Time Compute）」の重さを反映しています。
2. **Poetiq**の圧倒的なコスト効率: PoetiqはGemini 3のAPIを使用しているにもかかわらず、Gemini 3 Deep Thinkモードの約**1/3〜1/4**のコスト（$20対$80）で、より高いスコアを出しています。
    - メカニズム: これは「自己監査」機能によるものです。Deep Thinkモードはスコア最大化のために固定された長時間の思考を行う可能性がありますが、PoetiqはPython実行によって正解（と思われる解）が見つかった瞬間に計算を打ち切るため、APIコール数を劇的に削減できます [2]。

3. 人間とのコストパリティ: 興味深いことに、Poetiqのコスト（$20-$30）は、人間パネルのコスト（$17）に接近しています。AIが特定のタスクにおいて人間より賢いだけでなく、コスト的にも同等になりつつあるという「クロスオーバーポイント」に我々は立っています。

5.2 パレートフロンティアの再描画

経済学におけるパレートフロンティアは、コストと精度の最適なトレードオフ曲線を示します。通常、最後の数パーセントの精度を上げるためにコストは指数関数的に増加します（収穫逓減の法則）。

Poetiqは、まったく新しいパレートフロンティアを確立したと主張しています [2]。

- 曲線のシフト: 以前のパラダイムでは、45%のスコアを出すのに80ドルかかりました。Poetiqは65%を20ドルで達成しています。これは曲線上の移動ではなく、曲線全体が「より良く、より安く」という方向へシフトしたことを意味します。
- スキャフォールディングの乗数効果: これは、生のモデル（Gemini 3）をそのまま使うよりも、賢いメタシステムで包んだ方が効率的であることを証明しています。メタシステムの制御ロジックが付加する価値は、APIコールの追加コストを上回るのです。

5.3 低コスト領域の開拓：GrokとOSSモデル

Poetiqはハイエンドな「Mix」だけでなく、低コスト領域でも成果を上げています [11]。

- **Poetiq (Grok-4-Fast):** xAIのGrok-4 Fast Reasoningモデル上に構築され、ベースモデルよりも安価かつ高精度であると報告されています。これは、メタシステムがモデルの「癖」を理解し、無駄な推論を省いていることを示唆します。
- **Poetiq (GPT-OSS-b):** GPT-OSS-120B（オープンウェイトモデル）を使用し、1問あたり1セント未満で驚くべき精度を示しています。これは、予算に応じた「知能の調整」が可能であることを意味し、実用化の観点から極めて重要です。

---

6. 技術的深掘り：コードと再帰的探索の役割

なぜPoetiqはこれほど効率的なのでしょうか？その秘密は、推論媒体としての「Pythonコード」と、探索アルゴリズムとしての「再帰」にあります。

6.1 言語としてのPythonの優位性

自然言語は曖昧ですが、Pythonコードは厳密です。ARCのような論理パズルにおいて、自然言語で「右上の赤いピクセルを左に動かす」と考えるよりも、grid[y, x] = grid[y, x-1] とコードを書くほうが、検証可能です。

- **検証可能性:** 生成されたコードは即座にテスト可能です。エラーが出れば、それは明確な「失敗」のシグナルとなり、モデルは次の試行で修正できます。自然言語の思考では、論理が破綻していても気づかずに突き進んでしまう（幻覚）リスクがありますが、コード実行はそのリスクを「コンパイルエラー」や「アサーションエラー」として顕在化させます。

6.2 探索戦略：Best-of-Nと木の探索

Poetiqは、単一の解を生成して終わりではありません。複数のエキスパート（モデル）を用いて複数の解候補を生成し、それらを評価する「投票メカニズム」を採用しています [4]。これは「Best-of-N」探索や「Tree of Thoughts（思考の木）」アプローチの一形態と考えられます。

- **多様性の確保:** Gemini 3とGPT-5.1という異なるアーキテクチャを持つモデルを組み合わせることで、解の探索空間が広がります。Geminiが苦手とするパターンをGPTが得意とする場合があり（逆もまた然り）、メタシステムはこの相補性を利用しています。

6.3 ゼロショット適応の主張

Poetiqは、ARC-AGI-2の問題に対してゼロショットで適応したと主張しています [14]。これは、メタシステムのロジック（プロンプトの構造、エラー処理のフローなど）が汎用的であり、特定の問題セットに過学習していないことを示唆します。もしこれが真実であれば、このメタシステムはARC以外の推論タスク（例えば数学やコーディングコンテスト）にも転用可能である可能性が高いです。

---

7. 広範な影響と将来の展望

Poetiq (Mix)のARC-AGI-2における成功は、AI開発の未来に深い示唆を与えています。

7.1 モデルスケーリングからシステムスケーリングへの移行

2020年代前半の「スケーリング則」は、トレーニング時の計算量（より多くのデータ、より大きなGPU）に焦点を当てていました。しかし、2025年は**推論時計算（Inference-Time Compute）**の時代として定義されつつあります。Poetiqは、GPT-6を待たなくとも、既存のGPT-5.1に十分な「考える時間」と「正しい道具（Python）」を与えれば、AGIレベルの問題を解決できることを証明しました。これは、OpenAIのo1やGoogleのDeep Thinkが切り開いたトレンドを、外部的なスキャフォールディングによってさらに加速させるものです。

7.2 ファウンデーションモデルのコモディティ化

もしメタシステムがGeminiとGPTを自由に組み合わせて使えるなら、下層のモデルは交換可能な「部品（コモディティ）」になります。Poetiqのようなシステムインテグレーターにとって、Gemini 3は単なる「推論ユニット」に過ぎません。これにより、AI業界の価値の源泉が、モデル作成者（Google/OpenAI）から、それらを統合して最高の結果を出す認知アーキテクチャ設計者（Poetiqのような企業）へとシフトする可能性があります。

7.3 AGIは解決されたのか？

Poetiqのブログタイトル「ARC-AGI-2 is Solved（ARC-AGI-2は解決された）」[11] は、マーケティング的な誇張を含んでいる可能性がありますが、データは強力です。

- 「解決」の定義: もし「人間のベースラインを超えること」を解決と定義するなら、答えはイエスです。
- 汎用性の課題: しかし、ARC-AGI-2（視覚的グリッド論理）を解くことが、小説を書くことやロボットを操作することに直結するわけではありません。真のAGIは、これらの異なるドメイン間をシームレスに移動できる必要があります。
- 残されたギャップ: 65%というスコアは人間平均を超えていますが、依然として35%の誤答率があります。人間の「エキスパート」はおそらく90-100%のスコアを出すでしょう。真のAGIは平均的な人間ではなく、最も優れた人間に匹敵すべきです。

---

8. 詳細比較：Poetiq (Mix) vs. Gemini 3 Deep Think

最後に、主要な競合である2つのシステムの違いを明確にします。

特徴	**Poetiq (Mix)**	**Gemini 3 Deep Think**
タイプ	メタシステム (ソフトウェアラッパー)	統合モデルモード (End-to-End)
構成要素	Gemini 3 + GPT-5.1 + Pythonインタプリタ	Gemini 3 (推論用に微調整)
推論媒体	Pythonコード (プログラム合成)	自然言語 + 内部トークン
エラー修正	明示的: Pythonのエラーを検知して再プロンプト	暗黙的: 内部のアテンション機構 (不透明)
停止条件	動的 (自己監査により早期終了)	固定/ヒューリスティック (計算予算に基づく)
適応性	高 (下層モデルを交換可能)	低 (モノリシックなアーキテクチャ)
コストプロファイル	~$20 (早期終了により効率的)	~$80 (計算集約的)
ARC-AGI-2 スコア	**~65.9%**	**45.1%**

結論としての優位性:
Poetiqのアーキテクチャは、**検証問題（Verification Problem）**を効果的に解決しています。Gemini 3 Deep Thinkが「思考」するとき、その思考連鎖は誤っている可能性がありますが、最終的なグリッドを出力するまで外部からのチェックを受けません。一方、Poetiqはモデルにコードを書かせることで、Pythonインタプリタを「敵対的（Adversarial）」な検証者として利用します。コードが例題で失敗すれば、システムは提出前にそれが間違っていることを知ることができます。この「テスト駆動型開発（Test-Driven Development）」のアプローチをAI推論に適用したことが、幻覚を排除し、スコアを飛躍させた最大の要因です。

結論

Poetiq (Mix)の出現とARC-AGI-2におけるそのパフォーマンスは、人工知能の歴史における重要なマイルストーンです。これは、確率的なファウンデーションモデルを決定論的な制御ロジックで包み込む再帰的メタシステムの有効性を実証する概念実証（PoC）となりました。

65%というスコアを達成することで、Poetiqは、厳密な検証の足場と反復的な改良のプロセスによって導かれれば、現在のフロンティアモデル（Gemini 3, GPT-5.1）が抽象推論タスクにおいて人間の能力を凌駕する潜在的な知能を持っていることを経験的に示しました。

さらに、費用対効果分析は驚くべき効率性を明らかにしました。推論リソースをインテリジェントに管理することで、Poetiqは競合他社の4分の1のコストでSOTAのパフォーマンスを提供しています。これは、AGIへの道が単にデータセンターを巨大化することにあるのではなく、私たちがすでに持っているモデルの労働力を効果的に管理・監査できる、よりスマートなソフトウェアアーキテクチャの設計にあることを示唆しています。

2026年に向けて、業界の焦点はこのようなエージェント的メタシステムへと急速にシフトしていくでしょう。「Mix」アプローチ—モデルのハイブリッド化とコード実行のようなツールの統合—は、もはやニッチなテクニックではなく、超人的な推論を実現するための確立された青写真となったのです。

報告書作成: シニアAIリサーチアナリスト
日付: 2025年11月28日

引用文献

1. 速報：これが本当なら、もう人類超えてる！【AGIが近い】, 11月 28, 2025にアクセス、 https://www.youtube.com/watch?v=kJTYx_3W3qk
2. Traversing the Frontier of Superintelligence - Poetiq, 11月 28, 2025にアクセス、 https://poetiq.ai/posts/arcagi_announcement/
3. Leaderboard - ARC Prize, 11月 28, 2025にアクセス、 https://arcprize.org/leaderboard
4. 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/ - Reddit, 11月 28, 2025にアクセス、 https://www.reddit.com/r/accelerate/comments/1p2grr3/poetiq_did_it_poetiq_has_beaten_the_human/
5. 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/ - Reddit, 11月 28, 2025にアクセス、https://www.reddit.com/r/mlscaling/comments/1p2gs65/poetiq_did_it_poetiq_has_beaten_the_human/
6. ARC-AGI-2, 11月 28, 2025にアクセス、https://arcprize.org/arc-agi/2/
7. Announcing ARC-AGI-2 and ARC Prize 2025, 11月 28, 2025にアクセス、https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025
8. ARC-AGI-2 A New Challenge for Frontier AI Reasoning Systems, 11月 28, 2025にアクセス、https://arcprize.org/blog/arc-agi-2-technical-report
9. AI Models Struggle with New ARC-AGI-2 Benchmark, Raising Doubts About AGI Progress | by Ebimaro Jessica | Artificial Synapse Media | Medium, 11月 28, 2025にアクセス、https://medium.com/artificial-synapse-media/ai-models-struggle-with-new-arc-agi-2-benchmark-raising-doubts-about-agi-progress-1c5b2fcb9cf6
10. Artificial General Intelligence (AGI): How far are we from achieving it? | by Lucky Malefu, 11月 28, 2025にアクセス、https://medium.com/@i.am.lvcky/artificial-general-intelligence-agi-how-far-are-we-from-achieving-it-30a5deb399e7
11. ARC-AGI 2 is Solved : r/accelerate - Reddit, 11月 28, 2025にアクセス、https://www.reddit.com/r/accelerate/comments/1p8cvxv/arcagi_2_is_solved/
12. Gemini 3: Google's Most Powerful LLM - DataCamp, 11月 28, 2025にアクセス、https://www.datacamp.com/blog/gemini-3
13. GPT-5.1-Codex-Max vs Gemini 3 Pro: Next-Generation AI Coding Titans - Medium, 11月 28, 2025にアクセス、https://medium.com/@leucopsis/gpt-5-1-codex-max-vs-gemini-3-pro-next-generation-ai-coding-titans-877cc9054345
14. ARC-AGI 2 is Solved : r/singularity - Reddit, 11月 28, 2025にアクセス、https://www.reddit.com/r/singularity/comments/1p8c6gy/arcagi_2_is_solved/
15. Poetiq, 11月 28, 2025にアクセス、https://poetiq.ai/
16. Yair Alon - Google Scholar, 11月 28, 2025にアクセス、https://scholar.google.co.uk/citations?user=Kd1_AzgAAAAJ&hl=ja
17. Google's new Deep Think feature in Gemini is terrifyingly good. Google is trading speed for intelligence. Forget Chain-of-Thought. Parallel Thinking is here. A deep dive into Gemini Deep Think and the new era of AI reasoning. : r/ThinkingDeeplyAI - Reddit, 11月 28, 2025にアクセス、https://www.reddit.com/r/ThinkingDeeplyAI/comments/1mfe0zn/googles_new_deep_think_feature_in_gemini_is/
18. A new era of intelligence with Gemini 3 - Google Blog, 11月 28, 2025にアクセス、https://blog.google/products/gemini/gemini-3/
19. Gemini 3 Deep Think outperforms Gemini 3 Pro's already impressive performance on Humanity's Last Exam (41.0% without the use of tools) and GPQA Diamond (93.8%). It also achieves an unprecedented 45.1% on ARC-AGI-2 (with code execution, ARC Prize Verified) : r/accelerate - Reddit, 11月 28, 2025にアクセス、https://www.reddit.com/r/accelerate/comments/1p0g0a2/gemini_3_deep_think_o

utperforms_gemini_3_pros/
20. GPT-5.1 AR -AGI scores.Achieving SOTA in ARC-AGI-1. : r/singularity - Reddit, 11月 28, 2025にアクセス、https://www.reddit.com/r/singularity/comments/1ozsogm/gpt51_ar_agi_scoresachieving_sota_in_arcagi1/
21. Gemini 3.0 vs GPT-5.1 vs Claude 4.5 vs Grok 4.1: AI Model Comparison - Clarifai, 11月 28, 2025にアクセス、https://www.clarifai.com/blog/gemini-3.0-vs-other-models
22. New AI Benchmark ARC-AGI-2 'Significantly Raises the Bar for AI' - eWeek, 11月 28, 2025にアクセス、https://www.eweek.com/news/ai-benchmark-arc-agi-2/
23. Poetiq's approach of building intelligence on top of Gemini 3 & GPT-5.1 models to achieve SOTA-results. : r/singularity - Reddit, 11月 28, 2025にアクセス、https://www.reddit.com/r/singularity/comments/1p2tzqw/poetiqs_approach_of_building_intelligence_on_top/
24. 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. - GitHub, 11月 28, 2025にアクセス、https://github.com/poetiq-ai/poetiq-arc-agi-solver
25. Comparison of Gemini 3 to other models on ARC-AGI 1 & 2 : r/singularity - Reddit, 11月 28, 2025にアクセス、https://www.reddit.com/r/singularity/comments/1p0g63r/comparison_of_gemini_3_to_other_models_on_arcagi/