2025年におけるAI推論の転換点：PoetiqによるARC-AGI-2の突破と進化的推論メタシステムの台頭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12月5日、AIスタートアップであるPoetiq AIは、汎用人工知能（AGI）の進捗を測る最も厳格なベンチマークの一つであるARC-AGI-2において、既存の最高記録を大幅に更新する**54%**というスコアを達成したと発表しました [1]。この成果は、GoogleのGemini 3 Deep Think（45.1%）やAnthropicのOpus 4.5（37.6%）といった巨大テック企業の最先端モデルを凌駕するものであり [2]、AI開発のパラダイムが「モデルの巨大化（スケーリング）」から「推論時の計算資源活用（Test-Time Compute）」へと移行していることを決定的に印象付ける出来事となりました。

本報告書は、Poetiqの発表内容、その背後にある技術的アーキテクチャである「Mind Evolution（精神の進化）」アルゴリズム、そしてこの成果が示唆するAI産業への経済的・技術的影響について、利用可能なリサーチ資料に基づき徹底的に分析したものです。Poetiqのアプローチは、既存の基盤モデル（LLM）を「再学習」させるのではなく、その外部に「思考の進化」を促すメタシステムを構築するという点で特異であり、1タスクあたり**30.57**ドルというコストで、GoogleのDeep Thinkソリューション（77.16ドル）の半額以下でより高い推論能力を実現しています [1]。

本稿では、ARC-AGI-2ベンチマークの特異性、Poetiqが採用した「学習されたテスト時推論（Learned Test Time Reasoning）」のメカニズム、そして学術界やコミュニティから寄せられる「過学習（Overfitting）」への懸念についても公平かつ詳細に検証を行います。

第1章：2025年後半におけるAI推論の現在地と「壁」

1.1 スケーリング則の限界と流動性知能の停滞

2024年まで、AI業界は「スケーリング則（Scaling Laws）」というドグマに支配されていました。データ量と計算量を増やせば、AIの性能は対数線形的に向上するという経験則です。しかし、2024年後半から2025年にかけて、この法則に陰りが見え始めました。知識の検索や定型的なタスク処理能力（結晶性知能）は向上し続けたものの、未知の状況に適応し、新しいルールをその場で発見する能力、すなわち「流動性知能（Fluid Intelligence）」の向上が停滞したのです [4]。

この停滞を最も如実に示したのが、François Cholletによって提唱されたARC-AGIベンチマークでした。インターネット上のテキストデータの統計的相関を学習するだけでは解けない、視覚的かつ抽象的な推論を要するこのテストにおいて、GPT-4クラスのモデルでさえ、推論時間を十分に与えられな

い場合（System 1思考）、そのスコアはほぼ0%に近い状態が続いていました [4]。

1.2「System 2」思考へのパラダイムシフト

この閉塞感を打破するために浮上したのが、「System 2」思考の実装です。これは、人間の認知プロセスにおける「直感的で速い思考（System 1）」に対し、「論理的で遅い思考（System 2）」をAIに模倣させるアプローチです。OpenAIのoシリーズやGoogleのGemini 3 Deep Thinkは、モデルが回答を出力する前に内部で思考の連鎖（Chain of Thought）や探索を行うことで、この能力を獲得しようとしました [5]。

しかし、これらのアプローチは依然として「モデル内部」のトレーニングや強化学習に依存していました。これに対し、Poetiqのアプローチは、凍結されたモデル（Frozen Model）を外部から操作し、推論プロセスそのものを進化させるという点で一線を画しています。Poetiqの共同創業者であるShumeet Baluja氏とIan Fischer氏は、モデルの中に知能を組み込むのではなく、「モデルの周囲に完全な知的なエコシステムを構築する」という哲学を掲げています [6]。これは、AIの進化が単なる「脳の巨大化」から「思考技法の高度化」へと移行したことを意味します。

第2章：ベンチマークの深層分析（ARC-AGI-2）

Poetiqの「54%」という数字の重みを理解するためには、ARC-AGI-2というテストの過酷さと、その評価体系の複雑さを理解する必要があります。

2.1 ARC-AGI-1からARC-AGI-2への進化

初代ARC-AGI（2019年）は、AIにとって「容易だが人間には難しい」タスクではなく、「人間には容易だがAIには極めて難しい」タスクを集めたものでした。しかし、2024年までに、AIコミュニティは膨大な計算資源を用いた総当たり的な探索や、テストデータへの微妙な漏洩（リーク）を利用してスコアをハッキングする手法を見出し始めました。

これに対抗して2025年に導入されたARC-AGI-2は、より高度な「記号的解釈（Symbolic Interpretation）」を要求するように再設計されました [4]。例えば、単に色を合わせるだけでなく、図形が持つ意味（「囲いの中にある」「重力に従う」など）を理解しなければ解けない問題が増加しました。

- 人間のパフォーマンス: 人間はこのテストにおいて、十分な時間があれば98〜100%の正答率を叩き出します [7]。これは、ARCが測定しているのが「人間が生まれつき持っている汎用的な推論能力」であることを裏付けています。
- AIの現状: 2025年初頭の時点では、純粋なLLMはほぼ0%、最先端の推論システムでも一桁台後半から10%台に留まっていました [4]。

2.2 評価データセットの階層構造と「半公開」の罠

Poetiqの成果を評価する上で最も重要なのが、ARC Prizeにおけるデータセットの区分です。ここに

は、しばしば誤解や論争の種となる複雑な構造が存在します [8]。

以下の表は、ARC-AGIにおけるデータセットの分類とその役割を整理したものです。

表1: **ARC-AGI** データセットの分類とアクセス権限

データセット名	タスク数	アクセス権限	主な用途	**Poetiq**のスコア
Public Training Set	1,000+	公開 (Open Source)	モデルのトレーニング、DSL（ドメイン固有言語）の開発	-
Public Evaluation Set	120	公開 (Open Source)	ローカルでの検証用。モデルが内容を記憶（過学習）してしまうリスクが高い。	-
Semi-Private Eval	120	非公開 (Kaggle/Leaderboard上で評価)	公式リーダーボードのスコア算出用。タスク自体は見えないが、提出を通じてフィードバックを得られる。	**54.0%** [1]
Private Evaluation	120	完全非公開 (Strictly Hidden)	賞金付きコンペティション（Kaggle）の最終順位決定用。提出者は内容を一切知ることができない。	不明 (コンペ優勝者は24%)

Poetiqが54%を記録したのは**Semi-Private Evaluation Set**です。これは「検証済み（Verified）」とされていますが、Kaggleの賞金決定に使われる**Private Evaluation Set**とは異なります [1]。この区別は、後述する「過学習論争」において中心的な論点となります。

第3章：Poetiqの技術的解剖 ——「Mind Evolution」メタシステム

Poetiqの勝因は、より強力なモデルをトレーニングしたことではなく、既存のモデル（Gemini 3やGPT-5.1）をより賢く「使う」ためのメタシステムを構築した点にあります。その核となる技術が「Mind Evolution（精神の進化）」です。

3.1 学術的基盤：arXiv:2501.09891 "Evolving Deeper LLM Thinking"

Poetiqの創業者らとGoogle DeepMindの研究チームによる共著論文「Evolving Deeper LLM Thinking」[10] は、このシステムの理論的支柱となっています。この論文では、LLMの推論能力を進化的アルゴリズムによって飛躍的に向上させる手法が提案されています。

従来の「Chain of Thought（思考の連鎖）」や「Best-of-N（多数決）」が、モデルから独立したサンプルを生成するのに対し、Mind Evolutionは生成された思考（Thought）を遺伝子のように扱います。

メカニズムの詳細：

1. 集団生成 **(Population Generation)**: まず、LLMに対してタスクの解決策（推論プロセスを含む）を複数生成させます。これが第1世代の「個体群」となります。
2. 評価 **(Evaluation)**: ARCのようなタスクでは、入力と出力のペア（例題）が与えられています。生成された解決策（プログラムやルール）をこれらの例題に適用し、正解と一致するかどうかで「適応度（Fitness）」を評価します。正解のないタスク（TravelPlannerなど）においては、内部的な論理整合性や制約条件の充足度が評価基準となります [10]。
3. 選択と交叉 **(Selection & Crossover)**: 適応度の高い解決策を選抜し、それらを「交叉」させます。ここがLLM特有の革新点です。従来の遺伝的アルゴリズムではビット列を交換していましたが、Mind EvolutionではLLMに対し、「解決策Aの物体認識ロジックと、解決策Bの移動ロジックを組み合わせて、新しい解決策Cを作れ」といったプロンプトを与え、言語レベルでの概念的な融合を行います。
4. 突然変異 **(Mutation)**: 局所解（Local Optima）に陥るのを防ぐため、解決策に意図的な変更を加えます。例えば、「色の条件を無視してみる」「グリッドを逆回転させてみる」といった変異をLLMに指示し、探索空間を広げます [12]。
5. 自己監査と終了条件 **(Self-Audit)**: システムは解決策が十分に信頼できるレベルに達したかを自律的に判断し、思考プロセスを終了させます。これにより、簡単な問題には少ないコストを、難問には高いコストを配分することが可能になります [13]。

このプロセスにより、単独の推論では到達できない複雑な解法を、部分的な成功の組み合わせによって構築することが可能になります。実際に、論文では「TravelPlanner」ベンチマークにおいて、標準的なLLMの成功率が5.6%であったのに対し、Mind Evolutionは98%以上の成功率を達成したと報告されています [12]。

3.2「学習されたテスト時推論」の実装

Poetiqはこの進化的アプローチを「Learned Test Time Reasoning（学習されたテスト時推論）」と呼

んでいます[1]。これは、推論（Test Time）において、モデルが動的に学習（適応）しているかのように振る舞うことを意味します。

ARCタスクにおいて、これは「タスクごとのDSL（ドメイン固有言語）の即席生成」として機能していると考えられます。Poetiqのシステムは、与えられた数個の例題から、そのタスク専用の解決プログラムを進化させ、それをテスト入力に適用します。これにより、トレーニングデータに含まれていない未知のパターンに対しても、その場で適応することが可能になります。

第4章：競合比較と経済性分析

2025年12月現在のリーダーボードは、異なる哲学を持つシステム同士の戦場となっています。

4.1 競合システムの分析

1. Google: Gemini 3 Deep Think

- アプローチ: 統合型System 2。モデル内部に強化学習（RL）とモンテカルロ木探索（MCTS）のような探索機能を統合[5]。
- パフォーマンス: 45.1% (ARC-AGI-2)[2]。
- コスト: 約77.16ドル/タスク[1]。
- 特徴: Googleの巨大な計算資源を背景にした力技のアプローチ。「Deep Think」モードは、複数の仮説を並列に検証するため、極めて高コストです。

2. Kaggle優勝チーム (NVARC / The ARChitects)

- アプローチ: 高度に最適化されたDSL探索と、軽量モデルによるガイド。
- パフォーマンス: ~24% (Private Set) / ~53% (Public/Semi-Private)[15]。
- コスト: 約0.20ドル/タスク (コンペの制約)[15]。
- 特徴: 圧倒的なコスト効率。しかし、計算資源の制約（1タスクあたり数十秒〜数分）があるため、Poetiqのような深遠な探索は不可能です。

3. Poetiq Meta-System

- アプローチ: 進化的メタシステム（ラッパー）。
- パフォーマンス: 54.0% (Semi-Private)[1]。
- コスト: 約30.57ドル/タスク。
- 特徴: コストと性能のバランスにおいて、新しいパレート最適を達成。Googleの半額以下のコストで、より高いスコアを実現しています。

以下の表は、主要システムのコスト対効果を比較したものです。

表2: ARC-AGI-2における主要システムのコスト・パフォーマンス比較 (2025年12月)

システム名	開発組織	**ARC-AGI-2 スコア (Semi-Private)**	推定コスト **($/Task)**	技術的アプローチ
Poetiq Meta-System	Poetiq AI	**54.0%**	**$30.57**	進化的探索ラッパー (Mind Evolution)
Gemini 3 Deep Think	Google	45.1%	$77.16	統合型RL + 並列推論
Opus 4.5 (Thinking)	Anthropic	37.6%	$2.20	Chain of Thought (長文脈)
Gemini 3 Pro	Google	31.1%	$0.81	標準的な推論
NVARC Solution	Kaggle Team	24.0% (Private スコア*)	$0.20	DSL合成 + テスト時学習

注: *NVARC*のスコアは*Private Set*のものであり、他システムの*Semi-Private*スコアと直接比較する際には注意が必要です。

4.2 パレートフロンティアの移動と「1思考あたり」のコスト

Poetiqの成果における最大の意義は、知能の「価格破壊」ではなく、知能の「上限突破」を現実的なコスト曲線上で実現した点にあります。これまでのAIは、性能を上げるためにはモデルサイズを指数関数的に大きくする必要がありました。しかしPoetiqは、**推論時の計算量（時間）**を増やすことで、より安価なベースモデル（Gemini 3 Proなど）を用いても、最高級のモデル（Gemini 3 Deep Think）を凌駕できることを証明しました [1]。

30ドルというコストは、チャットボットとしては高額ですが、専門家レベルの推論としては破格です。例えば、新しい物理法則の発見や、複雑なコードベースのデバッグといったタスクにおいて、人間の専門家を雇えば数時間で数百ドルかかります。Poetiqはそれを30ドル、数分で解決する可能性を示しています。

第5章：創業者の血統と理論的背景

Poetiqの技術的信頼性は、その創業メンバーの経歴に強く裏打ちされています。

5.1 Shumeet Balujaと進化計算の歴史

共同CEOであるShumeet Baluja氏は、Googleにおける長年の研究者であり、進化計算（Evolutionary Computation）の分野におけるパイオニアの一人です [6]。

- **PBIL (Population-Based Incremental Learning)**: Baluja氏が1990年代に共同開発したこのアルゴリズムは、遺伝的アルゴリズムと競合学習を組み合わせたものであり、現在の「Mind Evolution」の祖父とも言える技術です。
- 研究の系譜: 彼の研究リストには、「Evolved GANs（進化的GAN）」「Adversarial Transformation Networks（敵対的変換ネットワーク）」といったタイトルが並びます [17]。これらはすべて、「勾配降下法（Backpropagation）」だけに頼らず、探索と淘汰によってシステムを最適化するアプローチです。

5.2 Ian Fischerとメタ学習

もう一人の共同CEO、Ian Fischer氏もGoogle DeepMind出身であり、「学習する方法を学習する（Meta-Learning）」分野の専門家です [6]。彼らのバックグラウンドは、Poetiqが単なる「LLMラッパー」のスタートアップではなく、進化的アルゴリズムと深層学習の融合という、AI研究の未踏領域に挑むディープテック企業であることを示しています。

第6章：論争と懐疑的視点 —— 過学習と「ベンチマーク・マーケティング」

54%という圧倒的なスコアに対し、AIコミュニティからは称賛と共に強い懐疑の声も上がっています。RedditやHacker Newsでの議論 [13] を中心に、主な批判点を分析します。

6.1 過学習（Overfitting）の懸念

最大の懸念は、Poetiqのシステムが「Semi-Private Set」に過学習しているのではないかという点です。

- メカニズム: Semi-Private Setはタスク自体は非公開ですが、システムを提出してスコアを得ることは可能です。開発者がこのスコアをフィードバックとして、「Mind Evolution」のパラメータ（突然変異率やプロンプトの形式）を調整し続ければ、その特定の120問に対してのみ高い性能を発揮するシステムが出来上がってしまう可能性があります（これを「Graduate Student Descent」と揶揄することもあります）。
- 反論: Poetiq側は、使用したベースモデル（Gemini 3, GPT-5.1）はリリースから数時間〜数日しか経過しておらず、モデル自体をARCデータでトレーニングする時間はなかったと主張しています [13]。また、ARC Prize運営チームも、Semi-PrivateとPrivateは「統計的に同分布（IDD）」になるよう調整されており、一方での高スコアは他方での高スコアを示唆するはずだとしています [7]。

6.2 コストの壁と実用性

「1問解くのに30ドルもかかるシステムは実用的ではない」という批判もあります [20]。

- 反論: Poetiqの創業者は、このコストは初期段階のものであり、最適化によって急速に低下すると予測しています。また、彼らのビジネスモデルは、大量の安価なリクエストを処理することではなく、高付加価値な推論（創薬、材料科学、金融分析など）を提供することにあります [6]。これらの分野では、30ドルは誤差の範囲です。

6.3「ベンチマーク・マーケティング」説

一部の批評家は、PoetiqがARCという特定のベンチマークに特化しすぎていると指摘します。「ARCで高得点を取ったからといって、AGIが実現したわけではない」という主張です [18]。

- 分析: 確かにARCは特定の種類の視覚的パズルに限定されています。しかし、「Mind Evolution」論文において、全く異なるドメインである「TravelPlanner（旅行計画）」でも98%のスコアを達成している事実は [10]、この手法がドメイン非依存の汎用的な推論能力を持っていることを強く示唆しています。

第7章：将来展望とAI産業へのインプリケーション

7.1「推論クラウド」の勃興

Poetiqの成功は、AIの価値が「学習（Training）」から「推論（Inference）」へとシフトしたことを決定づけました。今後、NVIDIAのGPU需要は、巨大モデルの学習用から、推論時の膨大な探索計算用へと比重を移していくでしょう。推論時に数分間「考え続ける」AIが標準化すれば、推論計算市場は現在の数千倍に膨れ上がる可能性があります [21]。

7.2 基盤モデルのコモディティ化

もしPoetiqのようなメタシステムを使えば、中規模のオープンモデル（例：LlamaやGemini Pro）でも、最高峰のクローズドモデル（GPT-6やGemini Ultra）を凌駕できるとしたら、基盤モデル開発企業の競争優位性は揺らぎます。価値の源泉は「パラメータ数」から「認知アーキテクチャ（ソフトウェア）」へと移行し、Poetiqのようなミドルウェア企業が覇権を握るシナリオが現実味を帯びてきます。

7.3 ARC-AGI-3とAGIへの道のり

ARC-AGI-2で50%を超えた今、次なるフロンティアはARC-AGI-3です [4]。ここでは、静的なパズル解決だけでなく、環境との相互作用やエージェンシー（主体性）が問われることになります。Poetiqの「Mind Evolution」が、静的な推論だけでなく、動的な環境適応においても機能するかどうかが、真のAGI（汎用人工知能）実現への試金石となるでしょう。

結論

PoetiqによるARC-AGI-2での54%達成は、単なるリーダーボードの更新ではありません。それは、AIが「パターンの記憶再生装置」から「能動的な思考装置」へと進化したことを示すマイルストーンです。

GoogleのDeep Thinkが示した「巨大資本による統合型アプローチ」に対し、Poetiqは「進化的アルゴリズムによるメタシステムアプローチ」で対抗し、コストと性能の両面で勝利を収めました。30ドルというコストや過学習への懸念は残るものの、「推論時の探索（**Inference-time Search**）」こそが次なるスケーリング則であるという事実は、もはや疑いようがありません。

Shumeet Baluja氏らが率いるPoetiqは、生物進化が数十億年かけて行った「試行錯誤と淘汰」のプロセスを、シリコンチップ上で数分間に圧縮して再現しようとしています。その試みが成功したとき、私たちは真の意味での「人工知能」を目撃することになるのかもしれません。

参考文献・データソース

- [15] ARC Prize 2025 Results Analysis
- [3] ARC-AGI Leaderboard Data
- [1] Poetiq.ai: "Poetiq Shatters ARC-AGI-2 State of the Art"
- [4] ARC-AGI Benchmark Definitions
- [6] Poetiq Team Background
- [5] Google Gemini 3 Deep Think Announcement
- [2] Google Gemini 3 Product Page
- [10] arXiv:2501.09891 "Evolving Deeper LLM Thinking"
- [11] DeepMind Research: Mind Evolution
- [18] Reddit Discussion on Overfitting
- [7] ARC Prize Blog on Set Difficulty
- [7] ARC-AGI-2 Human Baselines

引用文献

1. Poetiq Shatters ARC-AGI-2 State of the Art at Half the Cost, 12月 6, 2025にアクセス、https://poetiq.ai/posts/arcagi_verified/
2. A new era of intelligence with Gemini 3 - Google Blog, 12月 6, 2025にアクセス、https://blog.google/products/gemini/gemini-3/
3. Leaderboard - ARC Prize, 12月 6, 2025にアクセス、https://arcprize.org/leaderboard
4. ARC-AGI-2, 12月 6, 2025にアクセス、https://arcprize.org/arc-agi/2/
5. Google Rolls Out Gemini 3 Deep Think to AI Ultra Users, 12月 6, 2025にアクセス、https://www.eweek.com/news/google-launches-gemini-3-deep-think-ai-ultra-users/
6. Poetiq, 12月 6, 2025にアクセス、https://poetiq.ai/

7. Announcing ARC-AGI-2 and ARC Prize 2025, 12月 6, 2025にアクセス、https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025
8. Guide - ARC Prize, 12月 6, 2025にアクセス、https://arcprize.org/guide
9. ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems - arXiv, 12月 6, 2025にアクセス、https://arxiv.org/html/2505.11831v1
10. Evolving Deeper LLM Thinking - arXiv, 12月 6, 2025にアクセス、https://arxiv.org/html/2501.09891v1
11. Evolving Deeper LLM Thinking - Google DeepMind, 12月 6, 2025にアクセス、https://deepmind.google/research/publications/evolving-deeper-llm-thinking/
12. Mind Evolution: DeepMind is Teaching AI to Think Deeper Through Natural Selection, 12月 6, 2025にアクセス、https://gregrobison.medium.com/mind-evolution-deepmind-is-teaching-ai-to-think-deeper-through-natural-selection-abaabc1e52ed
13. 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, 12月 6, 2025にアクセス、https://www.reddit.com/r/accelerate/comments/1p2grr3/poetiq_did_it_poetiq_has_beaten_the_human/
14. Roll over, Darwin: How Google DeepMind's 'mind evolution' could enhance AI thinking, 12月 6, 2025にアクセス、https://www.zdnet.com/article/roll-over-darwin-how-google-deepminds-mind-evolution-could-enhance-ai-thinking/
15. ARC Prize 2025 Results and Analysis, 12月 6, 2025にアクセス、https://arcprize.org/blog/arc-prize-2025-results-analysis
16. Shumeet Baluja's research works | Google Inc. and other places - ResearchGate, 12月 6, 2025にアクセス、https://www.researchgate.net/scientific-contributions/Shumeet-Baluja-3225574/publications/2
17. Shumeet Baluja - Google Scholar, 12月 6, 2025にアクセス、https://scholar.google.com/citations?user=PggaADkAAAAJ&hl=en
18. ARC-AGI 2 is Solved : r/singularity - Reddit, 12月 6, 2025にアクセス、https://www.reddit.com/r/singularity/comments/1p8c6gy/arcagi_2_is_solved/
19. Traversing the Frontier of Superintelligence - Poetiq, 12月 6, 2025にアクセス、https://poetiq.ai/posts/arcagi_announcement/
20. Gemini 3 Deep Think benchmarks : r/singularity - Reddit, 12月 6, 2025にアクセス、https://www.reddit.com/r/singularity/comments/1p0fspc/gemini_3_deep_think_benchmarks/
21. Understanding Test-Time Compute: A New Mechanism Allowing AI to “Think Harder” | by Rendy Dalimunthe | Medium, 12月 6, 2025にアクセス、https://medium.com/@rendysatriadalimunthe/understanding-test-time-compute-a-new-mechanism-allowing-ai-to-think-harder-19e017abc540
22. Evolutionary Test-Time Compute: trade time & token for creativity - Alex Dong, 12月 6, 2025にアクセス、

https://alexdong.com/llm-evolutionary-test-time-compute.html