

54%

Poetiq: **54%**

Google Gemini 3 Deep Think: 45.1%

Anthropic Opus 4.5: 37.6%

AI推論の転換点：PoetiqがARC-AGI-2で巨人を打ち破る

2025年12月5日、AIスタートアップPoetiqは、最も困難な推論ベンチマークARC-AGI-2でスコア54%を達成。これは、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、推論時に思考する「テスト時計算（Test-Time Compute）」へと決定的に移行したことを示す。

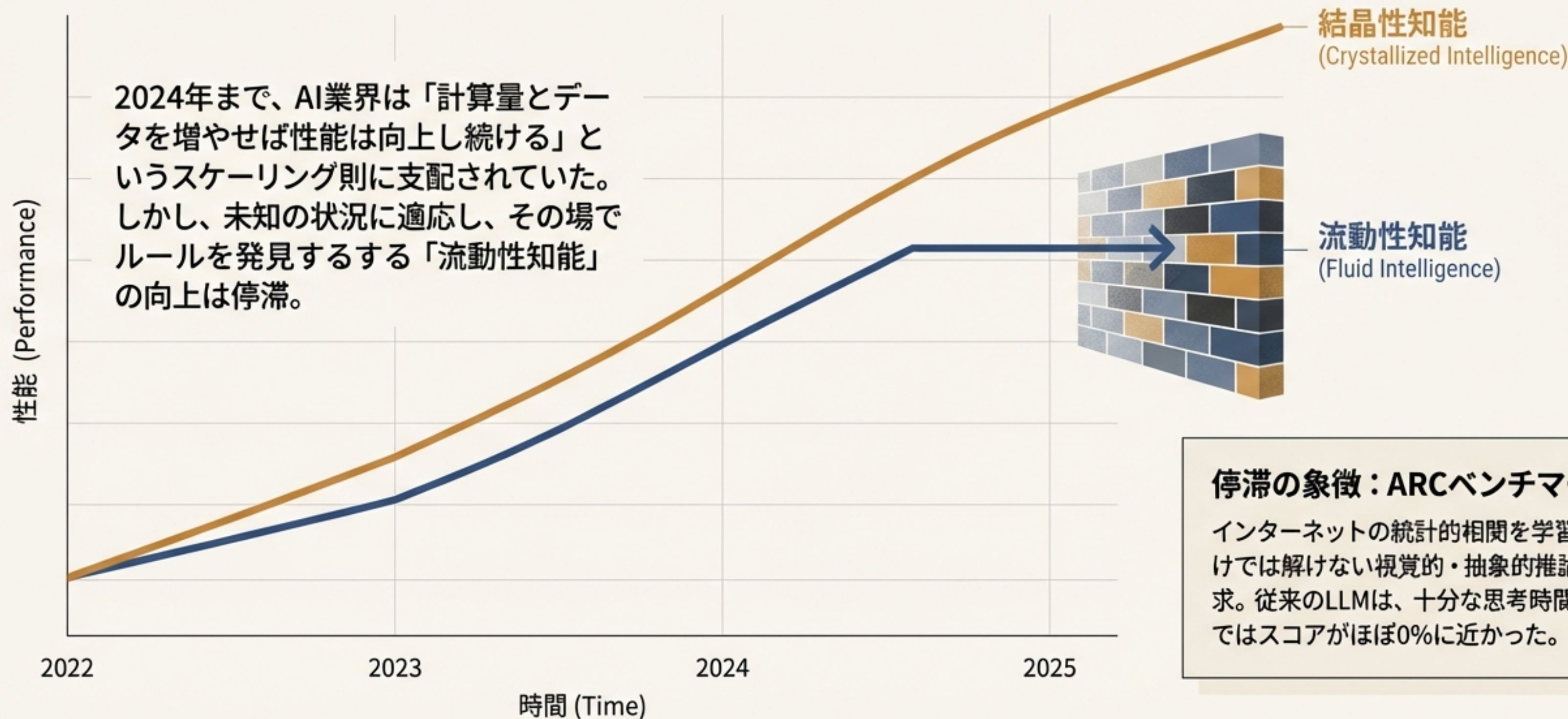


技術：既存モデルの外部で思考を進化させる「Mind Evolution」メタシステム



経済性：Googleの最高峰ソリューションの半額以下（\$30.57）でより高い性能を実現

なぜこの成果が重要なのか？ AI業界を支配した「スケーリング則」の限界



真の推論能力を測る究極の試金石、ARC-AGI-2の過酷さ

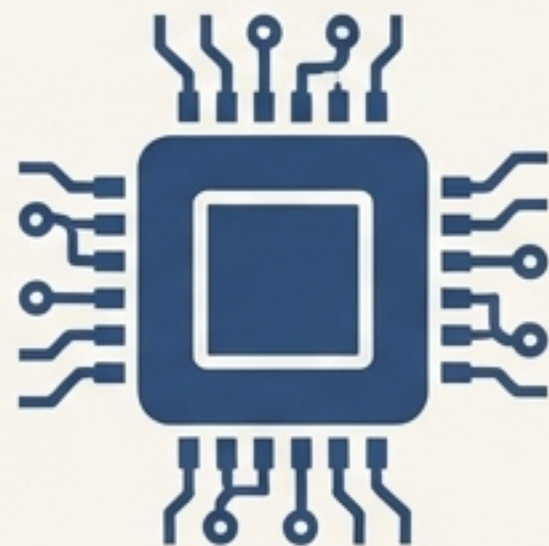
ARC-AGI-2は、単なるパターン認識ではなく、図形が持つ意味（例：「囲いの中にある」「重力に従う」）を理解する「記号的解釈」を要求するよう設計されている。



人間 / Human

「人間には容易だがAIには極めて難しい」タスク。
十分な時間があれば

98-100%の
正答率を達成。



AI

2025年初頭の最先端システム
でもスコアは

一桁台~10%台。

ARC-AGI-2は、テストデータへのリークや総当たりの探索といった「ハッキング」手法を無効化するように設計されている。

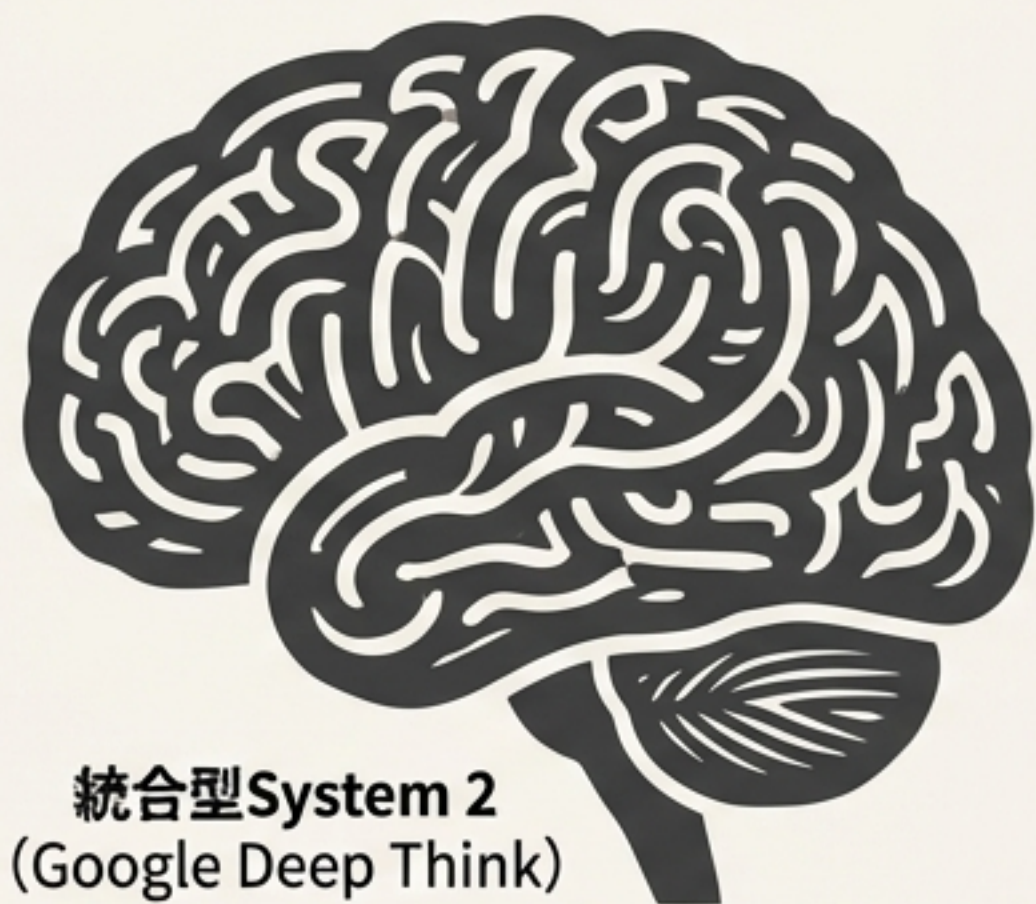
54%という数字の裏側：ARCデータセットの複雑な階層構造

データセット名	アクセス権限	主な用途	Poetiqのスコア
Public Training Set	公開 (Public)	モデルのトレーニング、DSL開発	-
Public Evaluation Set	公開 (Public)	ローカルでの検証（過学習リスク高）	-
Semi-Private Eval Set	非公開 (リーダーボード評価)	公式リーダーボードのスコア算出	54.0%
Private Evaluation Set	完全非公開 (Strictly Hidden)	賞金付きコンペの最終順位決定	不明 (コンペ優勝者は24%)

Poetiqの54%は「Semi-Private」セットでの記録。これは検証済み（Verified）スコアだが、Kaggleの賞金決定に使われる「Private」セットとは異なる。この区別が、後の過学習論争の核心となる。

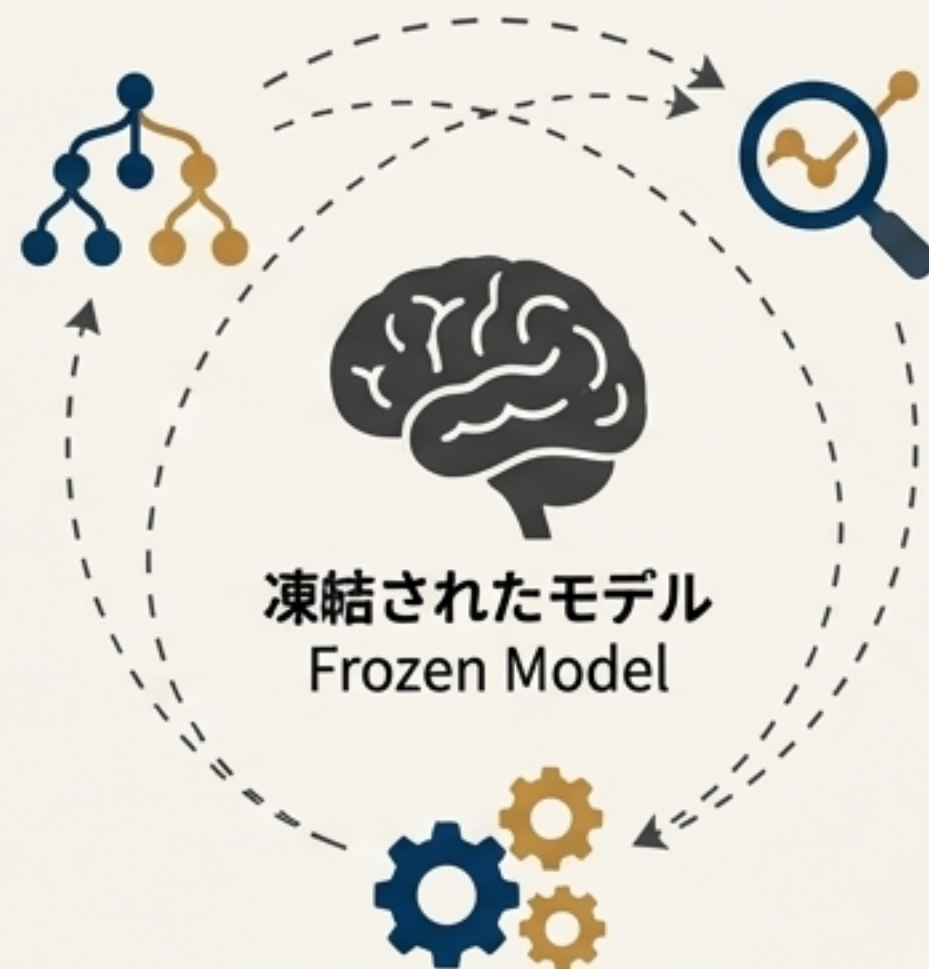
パラダイムシフト：「脳の巨大化」から「思考技法の高度化」へ

旧パラダイム / Old Paradigm



モデル内部の強化学習で知能を埋め込む。

新パラダイム / New Paradigm

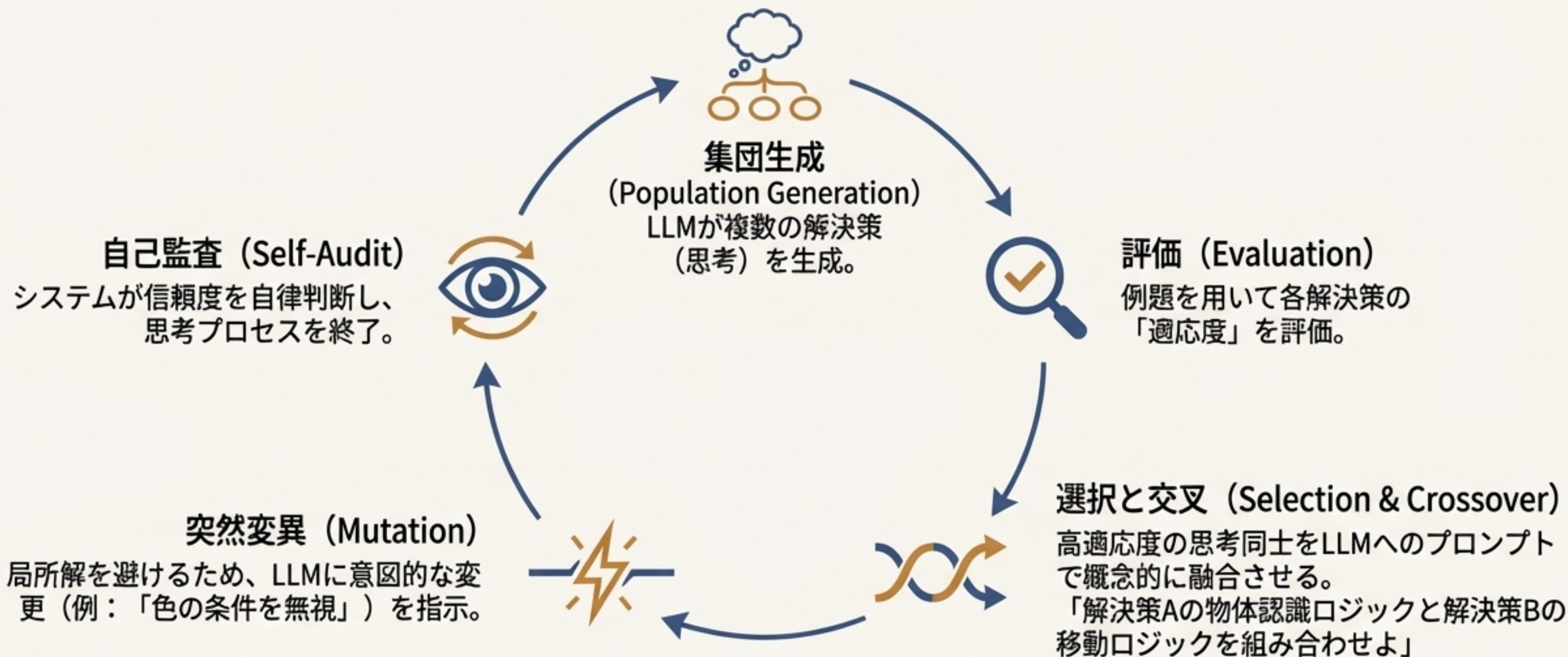


Poetiq メタシステム

「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エコシステムを構築する」

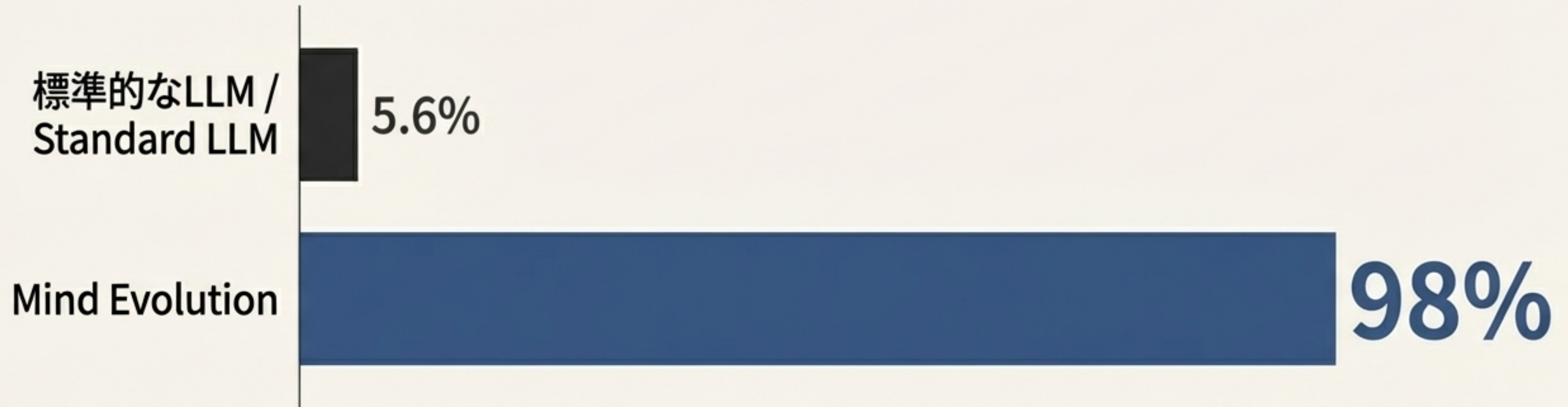
— Shumeet Baluja & Ian Fischer, Poetiq 創業者

Poetiqの心臓部：「Mind Evolution」 アルゴリズムの解剖



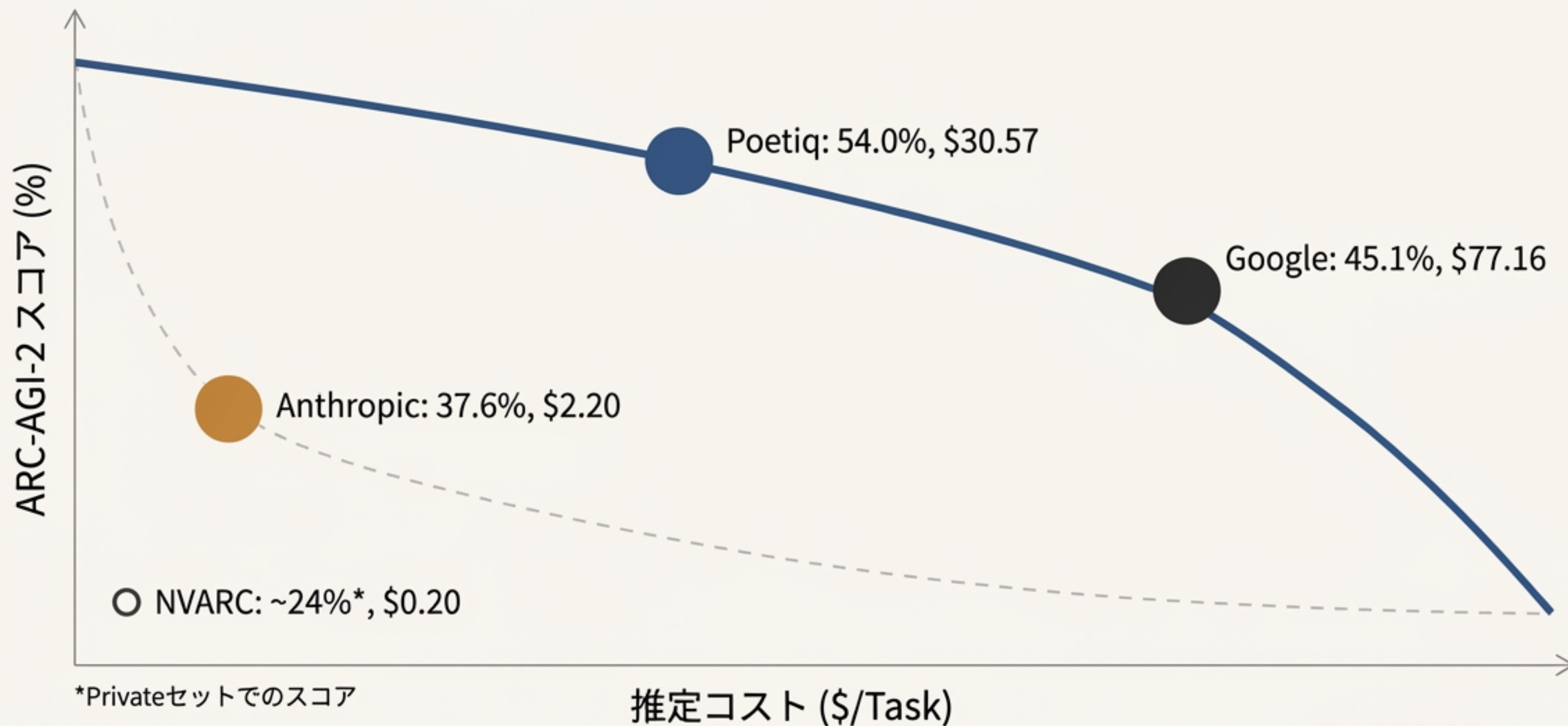
このプロセスは「学習されたテスト時推論 (Learned Test Time Reasoning)」として機能。
タスクごとに専用の解決プログラム (DSL) をその場で進化させている。

ARC特化ではない汎用性：TravelPlannerベンチマークでの圧倒的成果



「Mind Evolution」は特定の視覚パズルに限定されない。全く異なるドメインである旅行計画タスク「TravelPlanner」においても、標準的なLLMの性能を劇的に向上させ、98%以上の成功率を達成。これは、手法そのものがドメイン非依存の汎用的な推論能力を持つことを強く示唆する。

コストと性能の競争地図を塗り替える：新パレートフロンティアの出現



Poetiqは、Googleの半額以下のコストで、より高いスコアを達成。
これは単なる改善ではなく、コスト対効果の根本的な再定義である。

知能の「上限突破」を、現実的なコストで

\$30.57

Poetiq

vs.

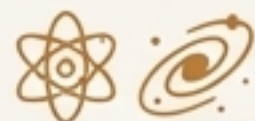
\$77.16

Google

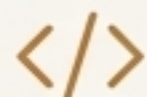
30ドルというコストは、チャットボットとしては高額だが、専門家レベルの推論としては破格である。

人間の専門家 (Human Expert)

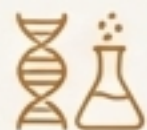
数時間で数百ドル～数千ドル



新しい物理法則の発見



複雑なコードベースのデバッグ



創薬・材料科学

Poetiq

30ドル、数分で解決する可能性

Poetiqが証明したのは、推論時の計算量（時間）を増やすことで、安価なベースモデルでも最高級モデルを凌駕できるという事実。

技術的信頼性の源泉：創業者の血統と研究の系譜

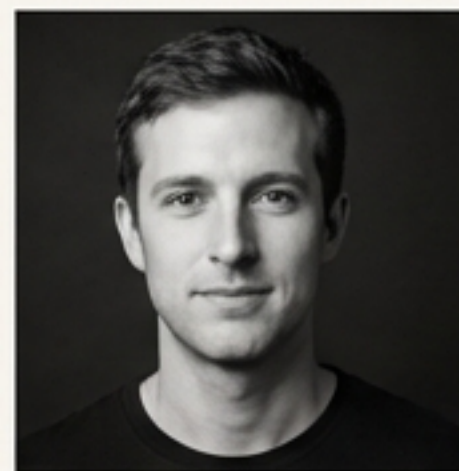
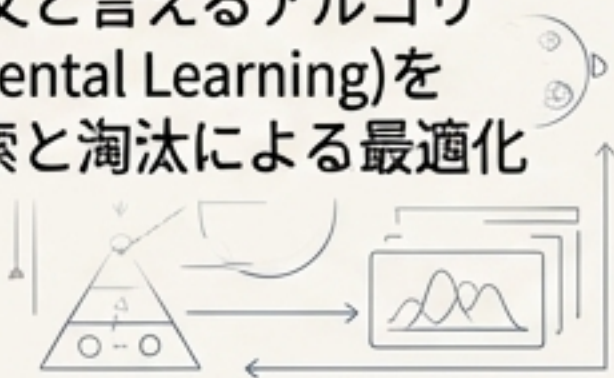


Shumeet Baluja

共同CEO、元Google研究者

Expertise: **進化計算** (Evolutionary Computation) のパイオニア

1990年代に「Mind Evolution」の祖父と言えるアルゴリズム**PBIL** (Population-Based Incremental Learning)を共同開発。彼の研究は一貫して、探索と淘汰による最適化に焦点を当てている。



Ian Fischer

共同CEO、元Google DeepMind

Expertise: 「学習する方法を学習する」メタ学習の系譜

Expertise: 「学習する方法を学習する」メタ学習 (Meta-Learning) の専門家



Poetiqは単なる「LLMラッパー」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。



論争と懐疑①：Semi-Private Setへの過学習（Overfitting）疑惑



懸念 (The Concern)

リーダーボードのスコアをフィードバックとして使い、「Mind Evolution」のパラメータ（突然変異率、プロンプト形式など）を調整し続ければ、特定の120問にのみ特化したシステムが完成してしまう可能性がある。

コミュニティでの揶揄：「Graduate Student Descent」



反論 (The Rebuttal)

Poetiq's Claim: 使用したベースモデル（Gemini 3, GPT-5.1）はリリース直後のもので、ARCデータで再学習する時間はなかった。

ARC Prize Org's Stance: Semi-PrivateとPrivateセットは「統計的に同分布（IDD）」になるよう設計されており、一方での高スコアは他方での高スコアを示唆するはず。

論争と懐疑②：コストの壁と「ベンチマーク・マーケティング」説

Point 1: コストの壁 (The Cost Barrier)



批判 (Criticism)

「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」

反論 (Rebuttal)

コストは初期段階のものであり、最適化で急速に低下する。ビジネスモデルは、創薬や金融分析など、30ドルが誤差となる高付加価値な推論がターゲット。

Point 2: 「ベンチマーク・マーケティング」説 (The "Benchmark Marketing" Theory)



批判 (Criticism)

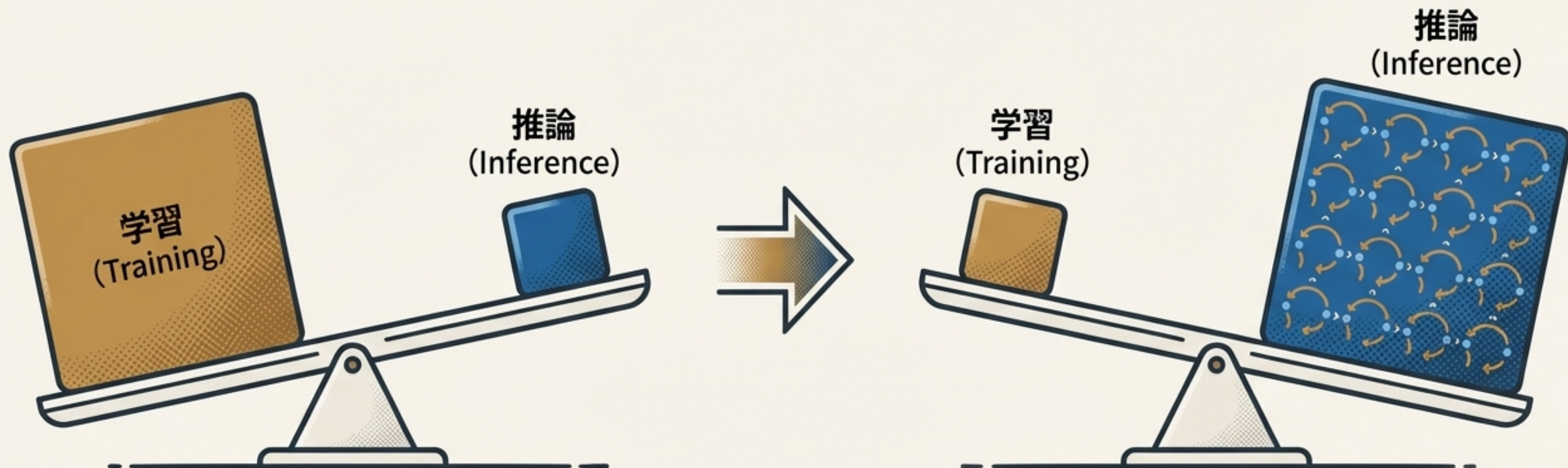
「ARCでの高得点がAGIの実現を意味するわけではない」

反論 (Rebuttal)

論文で示されたTravelPlannerベンチマークでの**98%**というスコアは、この手法が**ドメイン非依存の汎用的な推論能力**を持つことを示している。

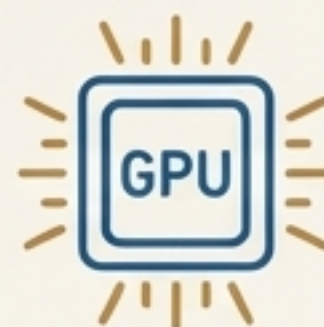
未来への示唆①：「学習」から「推論」へ、AIインフラの地殻変動

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た。



「推論クラウド」の勃興

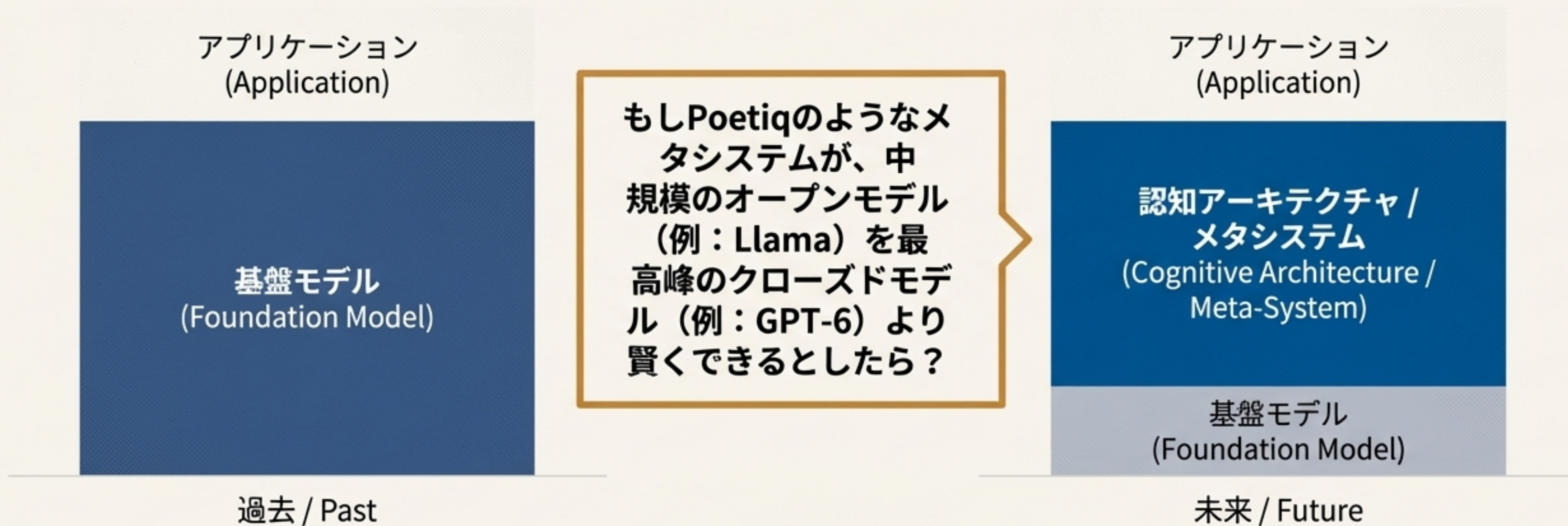
推論時に数分間「考え続ける」AIが標準化すれば、推論計算市場は現在の数千倍に膨れ上がる可能性がある。



NVIDIAへの影響

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。

未来への示唆②：基盤モデルのコモディティ化と認知アーキテクチャの台頭



- 基盤モデル開発企業の競争優位性が揺らぐ。
- 価値の源泉は「パラメータ数」から、Poetiqのようなミドルウェアが提供する「認知アーキテクチャ（ソフトウェア）」へと移行する。

結論：思考する機械の誕生と、次なる壁 ARC-AGI-3

Poetiqの54%達成は、AIが「パターンの記憶再生装置」から「能動的な思考装置」へと進化したマイルストーンである。進化的アルゴリズムによるメタシステムは、巨大資本による統合型アプローチにコストと性能の両面で勝利した。



ARC-AGI-3

次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ト（主体性）が問われる。

*Shumeet Baluja*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」を、シリコン上で数分間に圧縮しようとしている。その試みが成功したとき、我々は真の汎用人工知能を目撃することになる。