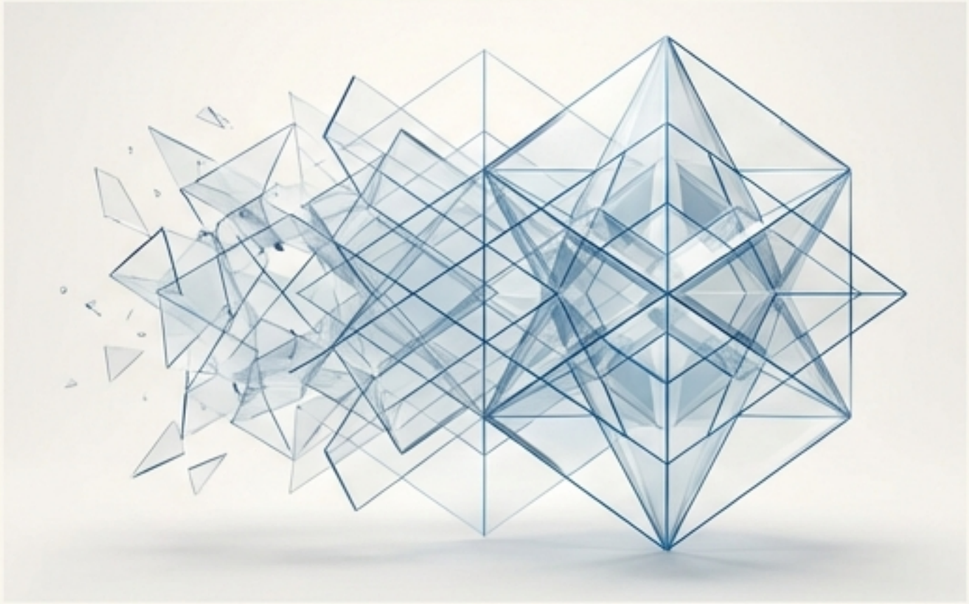
Poetiq AIの75%スコア達成
AGIの夜明けか 再現性の幻か？



ARC-AGI-2ベンチマークにおける主張の包括的分析と公式評価の予測

驚異的な主張と、過去の「スコア乖離」という重大な疑念

Shin Go Pro Heavy 主張（The Claim）

Poetiq AIがOpenAIの最新モデル「GPT-5.2 X-High」と独自のメタシステムを組み合わせ、ARC-AGI-2の公開評価セットで75%の正答率を達成したと主張。

この数値は、これまでの最先端モデルの記録（50%台半ば）を大幅に超え、人間の平均パフォーマンスをも凌駕する可能性を示唆する。

Shin Go Pro Heavy 疑念（The Doubt）

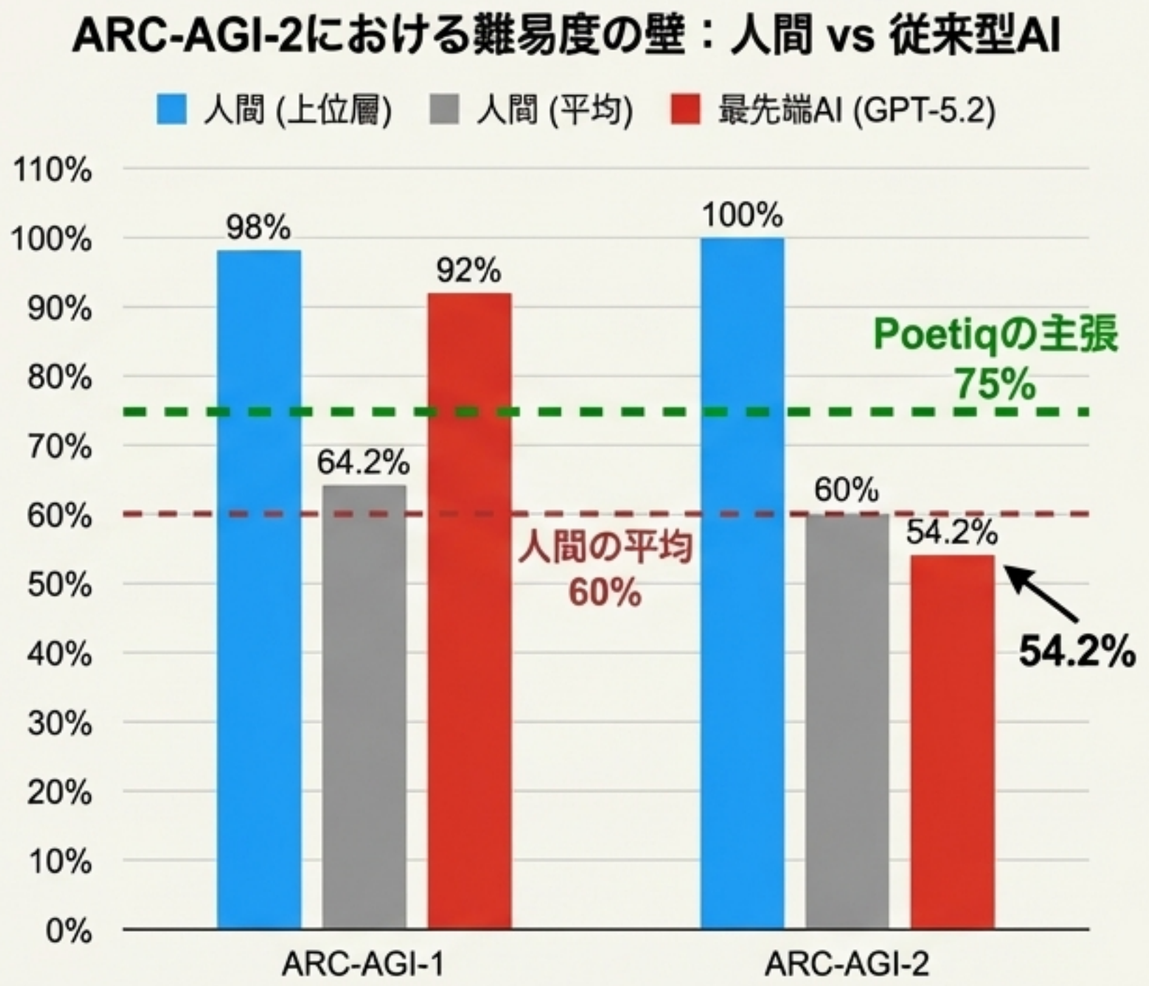
過去にPoetiq AIは、Gemini 3を用いたシステムで公開セット65.32%を記録後、公式検証では54.0%へと約11ポイントのスコア低下を経験。

この「スコア乖離」は、過学習やデータ汚染といった、AI評価における根深い問題を示唆する。

本レポートが解明する問い：この75%という数値の信頼性は？
そして公式評価後の「真の実力」は何点になるのか？

ARC-AGI-2：AIの「記憶」を封じ、「真の知能」を問う試金石

- **提唱者**：François Chollet（Keras開発者）
- **知能の定義**：「スキルそのものではなく、未知の環境に適応し、新たなスキルを獲得する能力（Skill Acquisiton Efficiency）」
  Efficiency）
- **目的**：従来の知識想起型ベンチマーク（MMLU等）と異なり、学習データにない新規パズルを解かせることで、純粋な推論能力、すなわち 「流動性知能（Fluid Intelligence）」を測定する。
- **人間との比較**：人間には容易だが、AIには極めて困難。人間の解決率は100%（全タスクが最低2人の人間により解決）、平均スコアは60%。

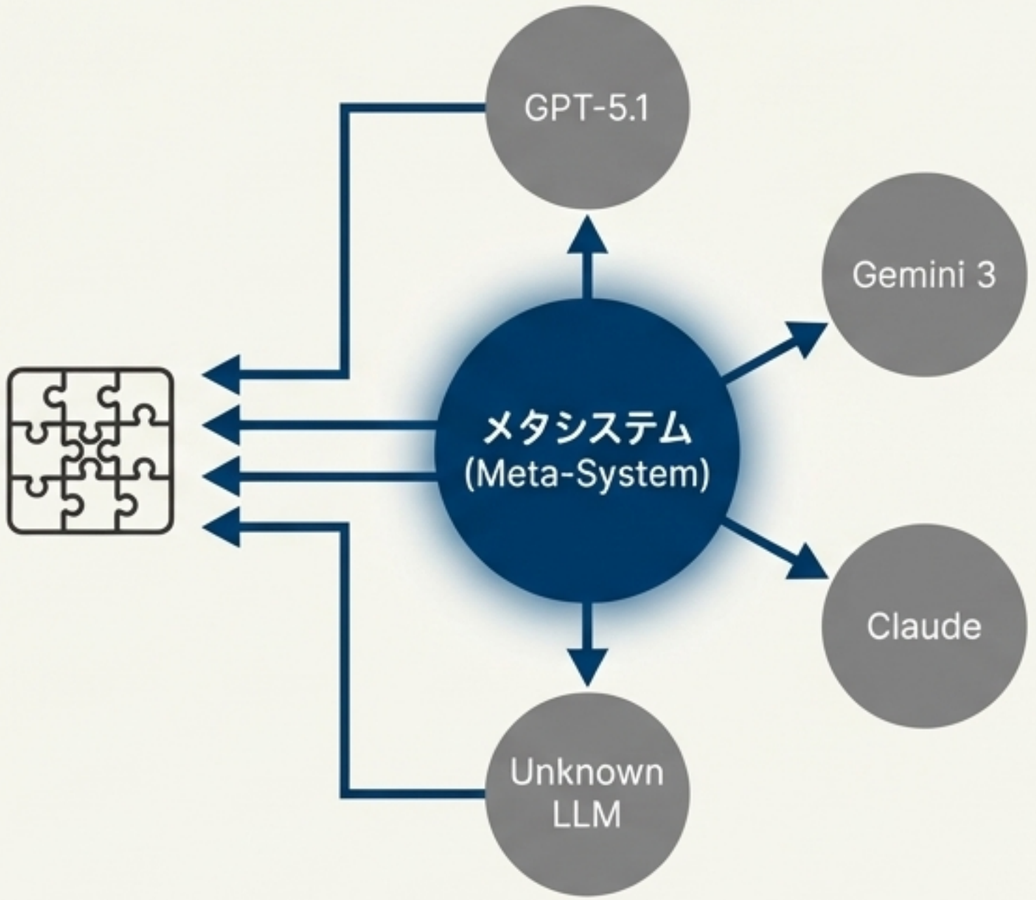


Poetiqの特異点：LLMを「使う」のではなく、推論プロセスを「指揮」するメタシステム

Poetiqのアーキテクチャは、巨大なLLMを単体で使うのではなく、それを包含する「メタシステム（Meta-System）」または「推論ハーネス（Reasoning Harness）」を構築している点に本質がある。

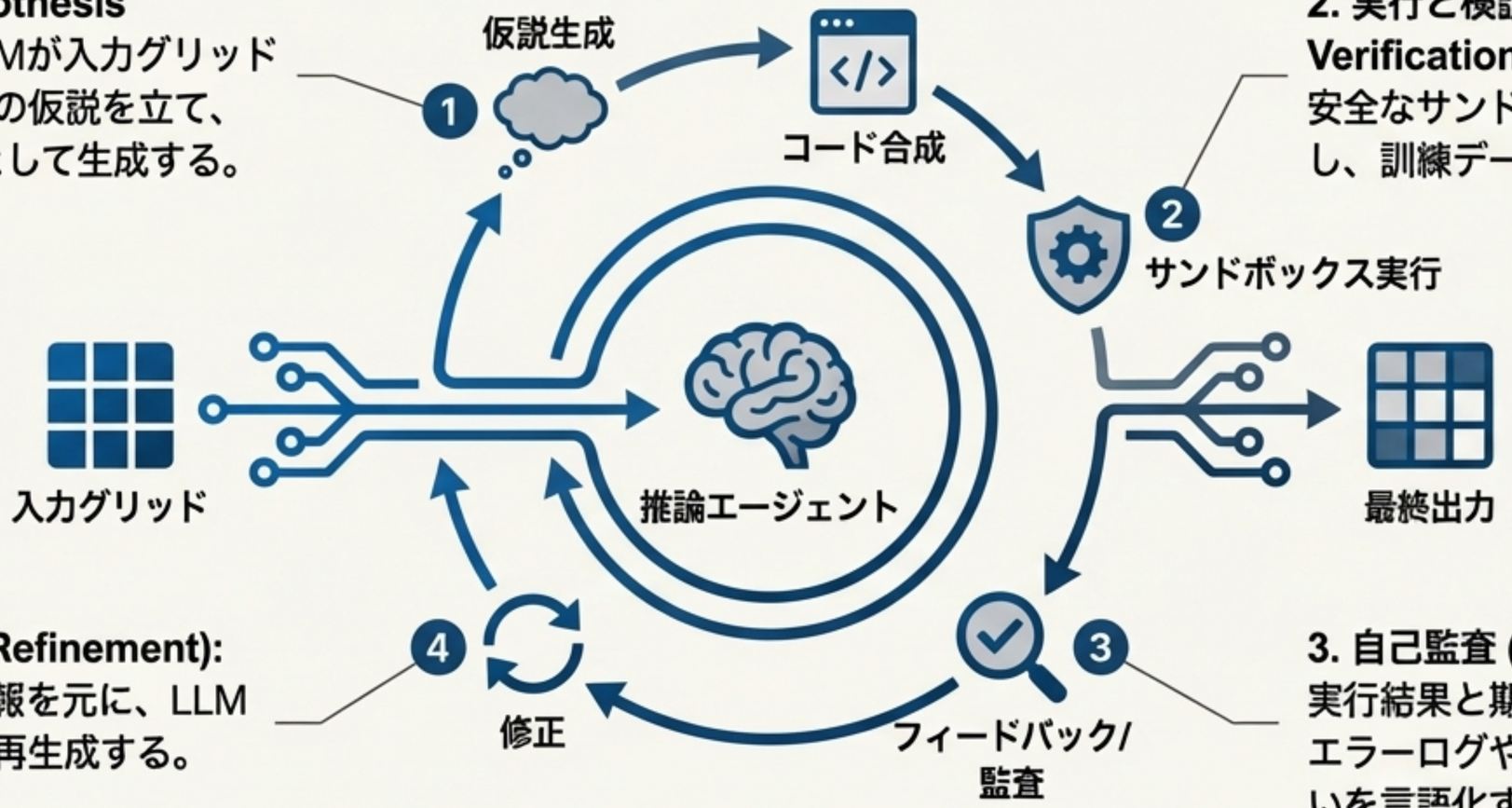
旧（Old Paradigm): プロンプトエンジニアリング。
「いかに上手くモデルに質問するか」
新（New Paradigm): インターフェースとしてのLLM。
「モデルから知識を引き出し、それをどう組み合わせるか」という高次の戦略を自動生成。

タスクの性質を分析し、最適な推論戦略を立案。
戦略に基づき、複数のモデル（Gemini 3, GPT-5.1等）を動的に呼び出す「専門家アンサンブル（Mixture of Experts similar approach)」的な挙動を示す。



Poetiq Meta-System: 再帰的推論と自己監査ループ

1. 仮説生成 (Hypothesis Generation): LLMが入力グリッドから変換ルールの仮説を立て、Pythonコードとして生成する。



2. 実行と検証 (Execution & Verification): 生成されたコードを安全なサンドボックス環境で実行し、訓練データに適用する。

3. 自己監査 (Self-Auditing): 実行結果と期待出力を比較検証。エラーログや差異を解析し、間違いを言語化する。

4. 修正と再生成 (Refinement): フィードバック情報を元に、LLMがコードを修正・再生成する。

Key Concept: このプロセスは「テスト時学習（Test-Time Training）」として機能し、モデルの重みを更新することなく、タスク専用のスキルを一時的に獲得させる。

GPT-5.2 X-High：圧倒的な推論能力を支える「思考トークン」と莫大なコスト

思考トークン（Thinking Tokens）

- ユーザーへの回答出力前に、モデル内部で数千〜数万トークンに及ぶ「思考の連鎖」を生成。
- 問題解決のプランニングや自己修正を行い、直感的な答えに飛びつくことを抑制する。
- 「X-High」設定は、この思考プロセスに割り当てる計算資源を極大化したモード。

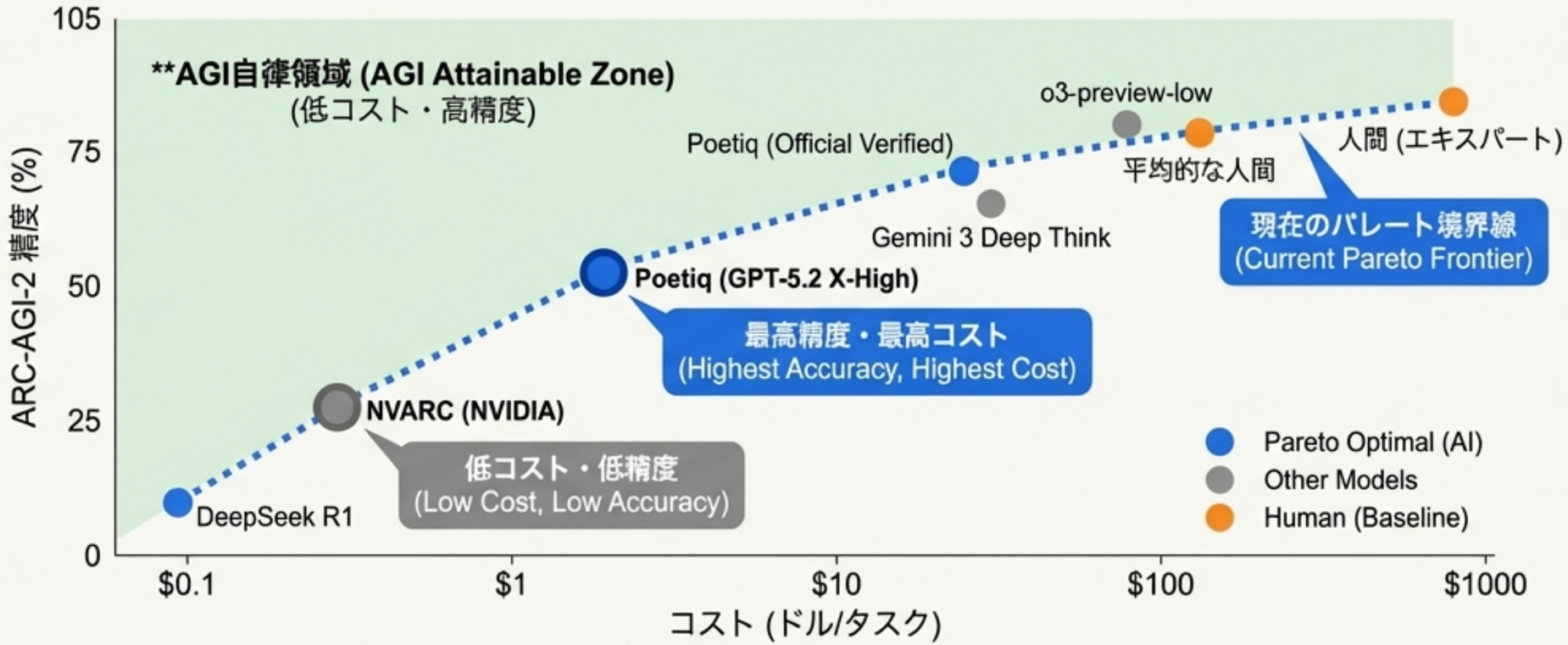
Synergy with Poetiq

- GPT-5.2はSWE-benchでSOTAを記録するなど、極めて高いコーディング能力を持つ。
- これは、Pythonコードを生成・修正するPoetiqのアプローチにとって最大の武器となる。

コスト（Cost）

- 内部の思考トークンも課金対象となるため、1回の回答生成に数ドルかかることも珍しくない。
- 「1タスクあたり$1.90」という過去の効率性アピールとは乖離する可能性がある「富豪的アプローチ」。

精度とコストのパレート境界線：AI推論の現在地



- **Key Analysis**: Poetiqは自社のシステムが「パレート境界（コストと精度のトレードオフにおける最適ライン）」を押し上げたと主張。
- **Poetiq (GPT-5.2 X-High)** 最高精度を達成するが、コストも最高レベルにある。
- **NVARC (NVIDIA)**: 対照的に、小規模モデルのファインチューニングにより、1タスクあたり$0.20という低コストで27.6%のスコアを達成。
- **真のAGI（True AGI)**: 究極の目標は、グラフの左上に位置する「AGI自律領域（低コスト・高精度）」である。

スコア乖離の深層：なぜ公開スコアは信用できないのか？

The ‘Generalization Gap’: 公開評価セットでのスコアと、非公開の検証セットでのスコアの差。前回のPoetiqのシステムでは約11.3ポイント。

1. データ汚染（Data Contamination）：
   - フロンティアモデルは、インターネット上のARC過去問や解答コードを学習データとして「記憶」している可能性が高い。
   - 公開セットは記憶で解けるが、非公開セットでは純粋な推論能力が問われる。

2. メタシステムの過学習（Meta-System Overfitting）：
   - 開発チームが公開セットを使ってシステムのロジックやプロンプトを調整・最適化してしまう。
   - この調整が、分布の異なる非公開セットでは逆効果になる可能性がある。
   - ARC Prize主催者は、スコア差10%超を「過学習」と見なす基準を設けている。

予測される公式スコア：過去の乖離率から「補正後実力値」を試算する

⚖️ **数理的推計（Mathematical Estimation）：**

Based on the previous system's ~17.3% reduction rate (65.32% → 54.0%).

$$75.0\% \times (1 - 0.173) \approx 62.0\%$$

シナリオ	想定される減少幅	推定検証スコア	評価
楽観的（Optimistic）	-5% ～ -8%	69% ～ 71%	歴史的快挙。人間の平均を完全に凌駕。
現実的（Realistic）	**-10% ～ -15%**	**60% ～ 64%**	**SOTA更新。人間の平均と同等レベル。**
悲観的（Pessimistic）	-20% ～ -25%	50% ～ 55%	現行SOTAと同等。コスト増に見合わず。

最も現実的なシナリオでも、AIは人間の平均スコア（60%）に到達・凌駕する。
これは歴史的なマイルストーンである。

結論：AIはついに「人間レベルの流動性知能」の入り口に立った

- 75%という主張は、過去の事例から下方修正される可能性が極めて高い。
- しかし、本レポートの予測中央値である62%〜64%というスコア自体が、AI史における画期的な成果である。
- このスコアは、これまでAIにとって「絶望的な壁」であったARC-AGI-2において、人間の平均パフォーマンス（60%）に並び、あるいは超えることを意味する。
- Poetiqの成果は、AIが真の推論能力を獲得する上で重要な転換点となる。

Poetiqの成果が示す3つのパラダイムシフト

1. 推論は「計算」できる (Reasoning is Computable)

膨大な計算資源（Thinking Tokens）と適切なアルゴリズム（再帰的自己改善）を組み合わせることで、人間の直感やひらめきに近い推論プロセスをAIで再現できる可能性が示された。

2. システムアプローチの勝利 (Victory of the Systems Approach)

単一モデルの巨大化だけでは限界があり、コード実行、自己監査、外部ツール連携を含む統合「システム」としてAIを構築するアプローチの有効性が証明された。

3. AGIへの明確な道筋 (A Clearer Path to AGI)

「暗記不能」なベンチマークの攻略は、AIが特定の狭いタスク（Narrow AI）から、汎用的な問題解決能力（General Intelligence）　へと質的に変化し始めたことを示唆している。

75%という数字は号砲である

Poetiq AIによる75%という自己申告スコアは、たとえ下方修正されるとしても、
決して無視してはならない。それは、AIが「人間並みの推論」を手に入れる日が、もはやSFの世界の話
ではなく、エンジニアリングの射程圏内に入ったことを告げるファンファーレなのである。
世界の注目は、今後の公式検証結果に集まる。

ご清聴ありがとうございました

引用・参考文献（Citations & References）

Poetiq AI. 'ARC-AGI-2 SOTA at Half the Cost'. poetiq.ai

IntuitionLabs. 'GPT-5.2 & ARC-AGI-2: A Benchmark Analysis of AI Reasoning'. intuitionlabs.ai

ARC Prize. 'ARC-AGI-2 A New Challenge for Frontier AI Reasoning Systems'. arcprize.org

Chollet, F. et al. 'The Abstraction and Reasoning Corpus (ARC)'.

Reddit. r/singularity. 'Poetiq Achieves SOTA on ARC-AGI 2 Public Eval'.

NVIDIA Developer Blog. 'NVIDIA Kaggle Grandmasters Win Artificial General Intelligence...'.

LessWrong. 'ARC-AGI-2 human baseline surpassed (updated)'.

OpenAI API Documentation. 'Reasoning models'.

LLM Stats. 'GPT-5.2: Pricing, Context Window, Benchmarks, and More'.

arXiv:2412.04604v2 [cs.AI] 8 Jan 2025.