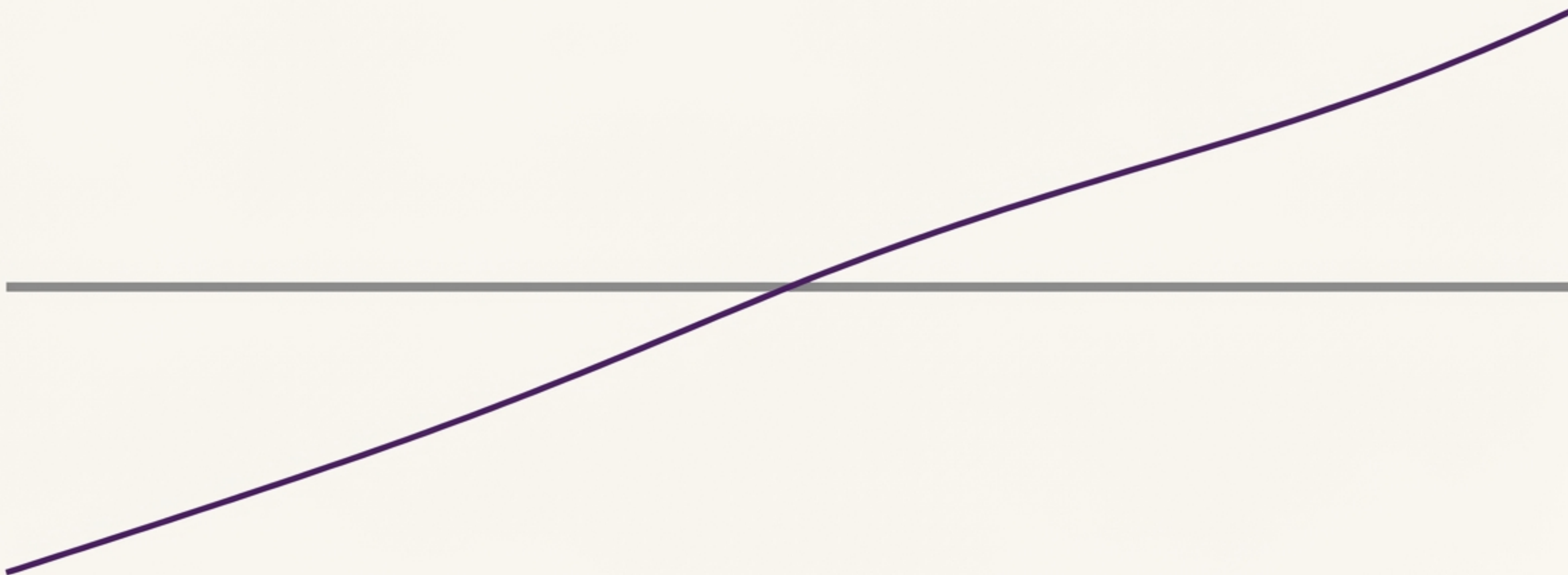


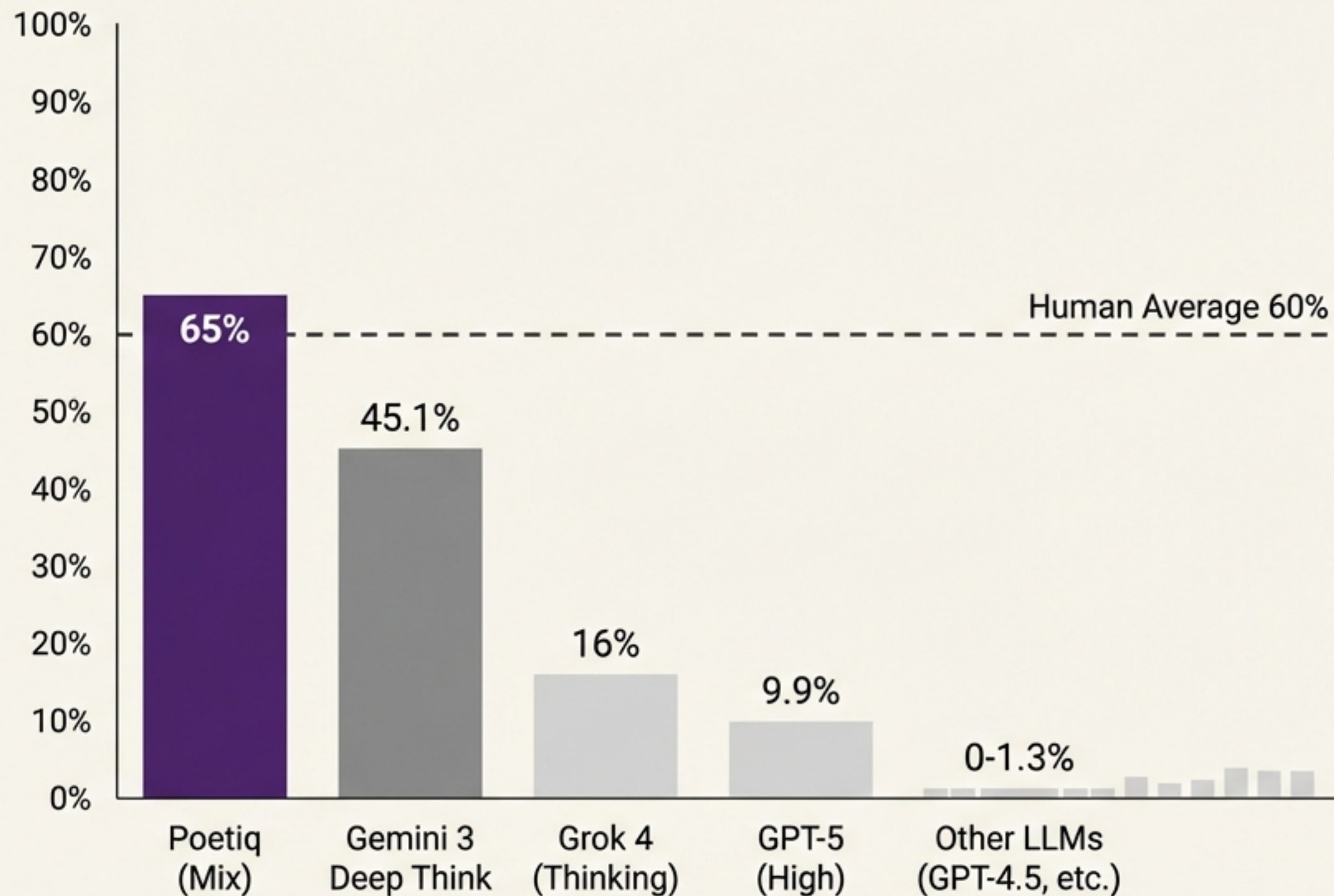
AI、抽象推論で人類を超える

Poetiqのメタシステムが、高難度ベンチマークARC-AGI-2で初めて人間ベースラインを突破



遂に超えた「人間の壁」

ARC-AGI-2ベンチマーク スコア比較



- **歴史的成果:** Poetiq (Mix) は、AI システムとして初めてARC-AGI-2 で人間の平均スコア（60%）を上回る約65%を達成。
- **圧倒的性能:** 最も近い競合である Gemini 3 Deep Think（45.1%）に約20ポイントの差をつけ、既存モデルを凌駕。
- **新アーキテクチャ:** この成果は、単一モデルの巨大化ではなく、複数の最先端LLMを統合する「メタシステム」によって実現された。

なぜこの成果が重要なのか？：ARC-AGI-2という「真の推論能力」を測る試金石

ベンチマークの概要

Name: ARC-AGI-2 (Abstraction and Reasoning Corpus for Artificial General Intelligence 2)

Purpose: 人間のような柔軟な抽象推論能力を測定するために設計された、色付きグリッド形式のパズル。

Key Feature: AIに未知の問題への適応能力を要求し、暗記やパターンマッチングでは解けない。

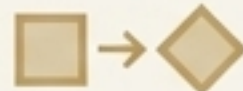
要求される3つの核心的推論



シンボリック推論: グリッド上の色や形が持つ意味や機能を理解する能力。



合成推論: 複数のルールを同時に適用し、相互作用を処理する能力。



文脈的ルール適用: 同一ルールを状況に応じて異なる方法で適用する柔軟性。

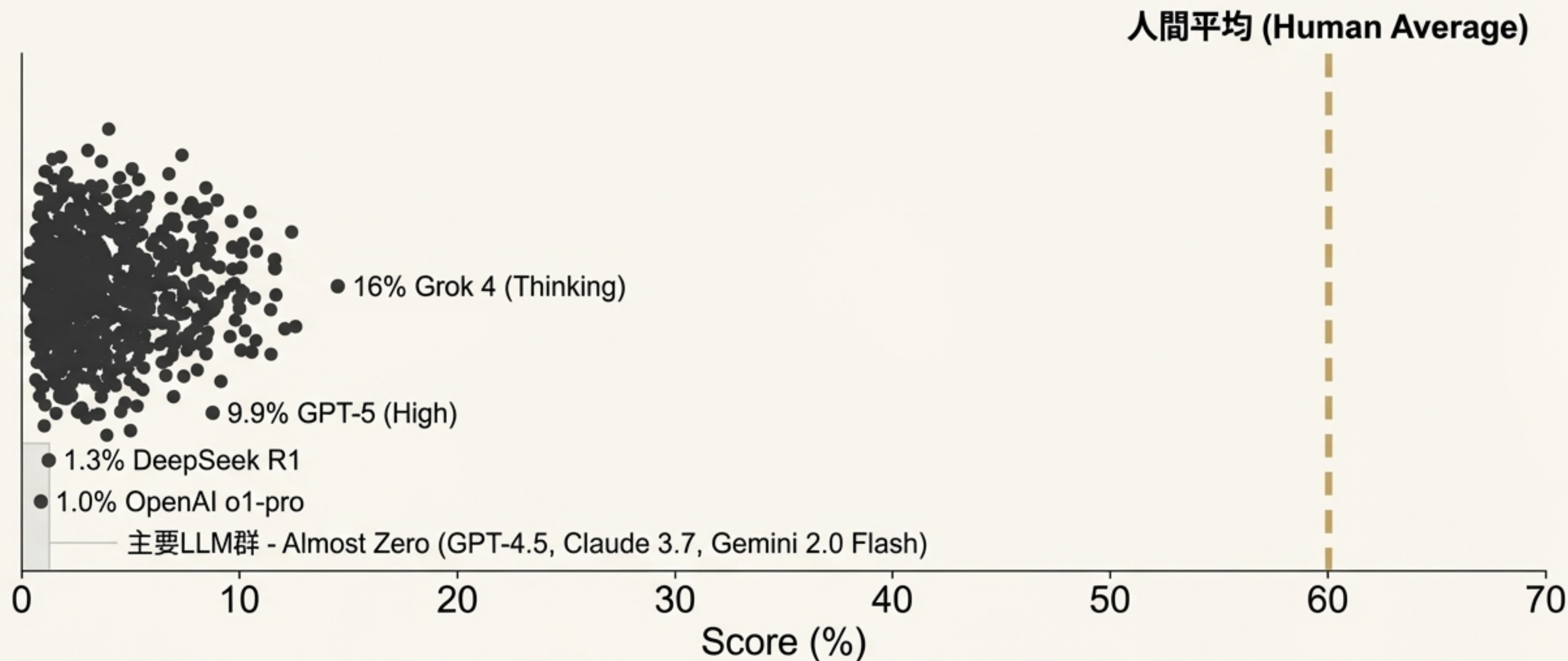
Training Examples

Input	Output	Input	Output

Test Task

Input	Output

世界最高峰モデルたちの「惨敗」



「純粋な」大規模言語モデルは軒並み壊滅的なスコアを記録。これは、現在のAI技術が人間レベルの抽象推論能力を本質的に獲得していないことを明確に示している。

巨大化から、知的なシステム化へ：Poetiqのパラダイムシフト

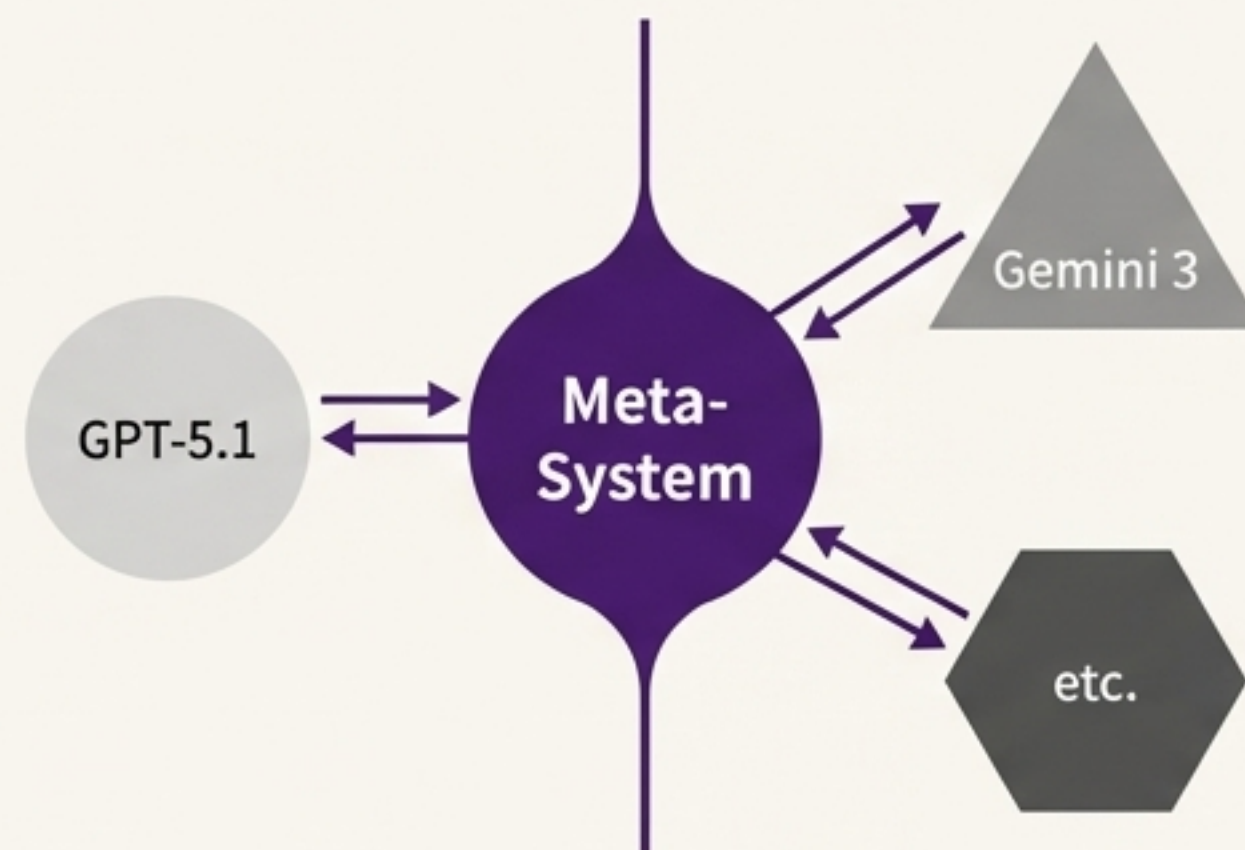
従来のアプローチ



単一巨大モデル

より多くのデータ、より多くのパラメータで
性能向上を目指す

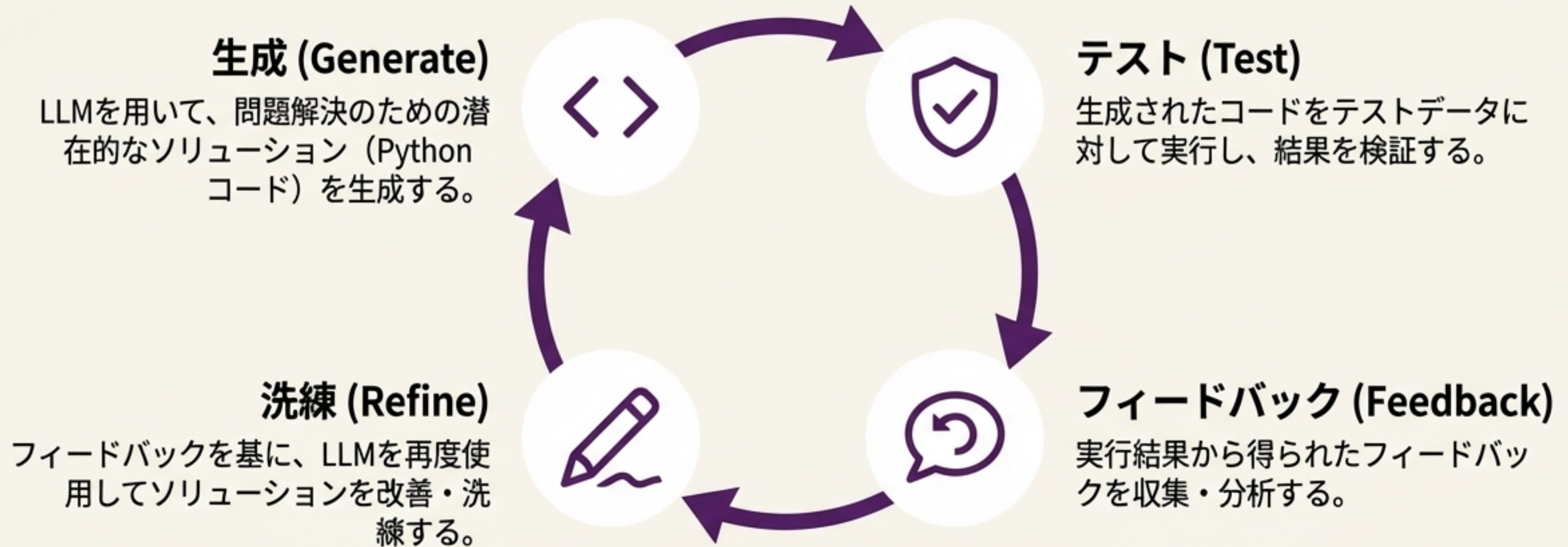
Poetiqのアプローチ



メタシステム

既存の最先端モデルを協調させ、相乗効果を生み出す
「LLM非依存的な再帰的自己改善メタシステム」

推論のエンジン：再帰的自己改善ループ



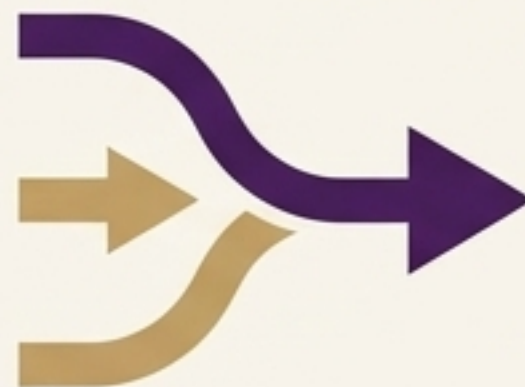
このアプローチは「プログラム合成（Program Synthesis）」と「テスト時適応（Test-Time Adaptation）」を組み合わせたものであり、単一の回答を求めるのではなく、反復的なプロセスを通じて段階的に正解を構築する。

成功を支える3つの技術的支柱



プログラム合成 (Program Synthesis)

ARC-AGIをコード生成タスクとして捉える。LLMが直接グリッドを予測するのではなく、実行・検証可能なコードを生成することで、具体的フィードバックを通じた精度向上を可能にする。



アンサンブルと投票 (Ensemble & Voting)

複数のエキスパート（コード生成プロセス）が独立してソリューションを生成。高度な投票技術を用いて結果を統合し、単一モデルの誤りを補正し、堅牢性を高める。

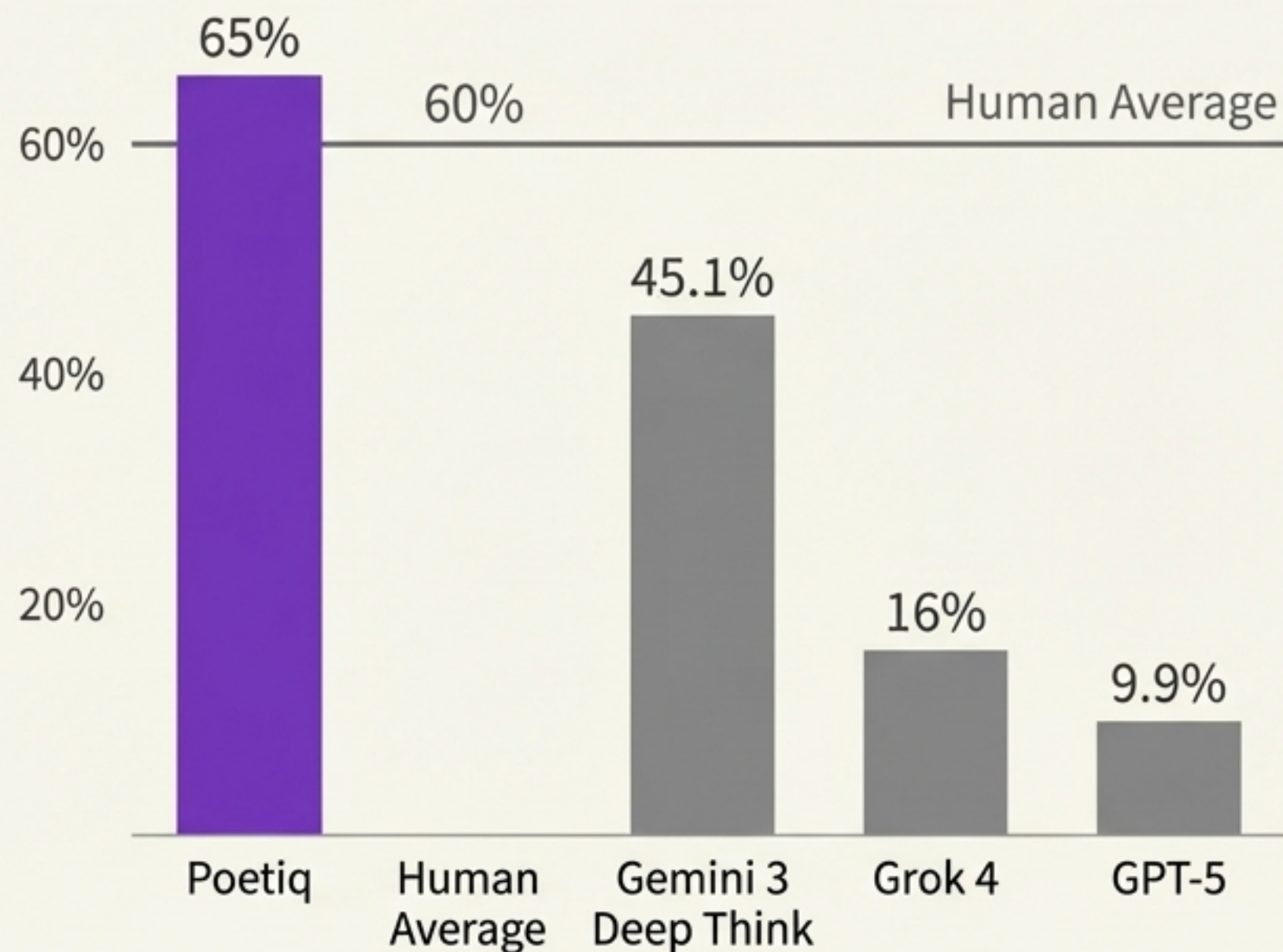


自己監査 (Self-Auditing)

システムが自身の進捗を自律的に評価。解決策が十分な品質に達したと判断すると自動的にプロセスを終了させ、無駄な計算を回避しコストを最小化する。

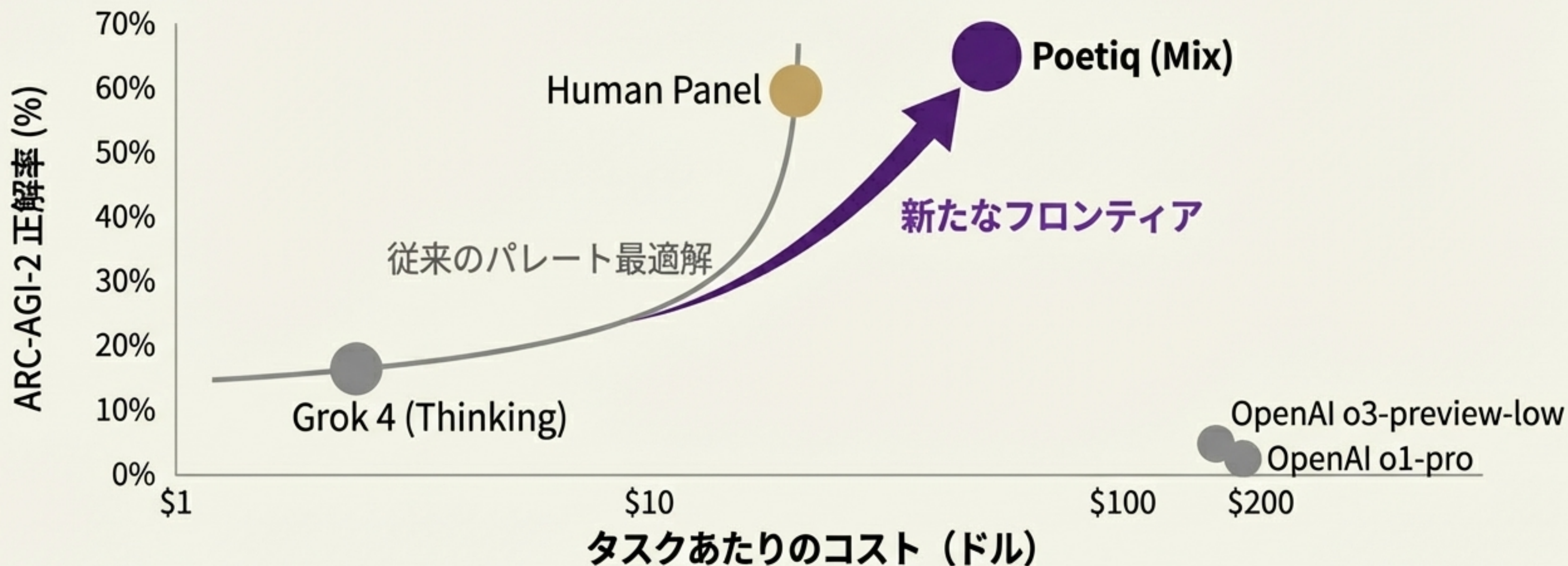
データで見るPoetiq (Mix) の圧倒的優位性

ARC-AGI-2ベンチマーク 最終スコア



- **人間超え:** 人間の平均スコア60%を5ポイント上回る歴史的マイルストーンを達成。
- **競合を圧倒:** Gemini 3 Deep Think (45.1%)との性能差は、メタシステムの相乗効果が単純なモデル性能の加算以上であることを示唆。
- **効率的な実行:** 自己監査機能により、平均2回未満のリクエストで試行を完了。ARC-AGI-2が許容する2回の試行枠内で高精度を実現する鍵となった。

性能とコストのフロンティアを再定義



Poetiqは、コストと性能のトレードオフを打破。人間の約3倍のコストで5ポイント高い性能を達成し、AI推論におけるコスト効率の基準を大幅に更新した。

特定モデルに依存しない、汎用性と転移可能性

- ✓ **事前適応:** メタシステムは、GPT-5.1とGemini 3がリリースされる前に、オープンソースモデルのみで開発・適応された。
- ✓ **ゼロショット実行:** 開発プロセスにおいて、ARC-AGI-2の問題は一切使用されていない。
- ✓ **広範な互換性:** アプローチは、GPT、Claude、Gemini、Grokなど12種類以上のLLMファミリーに対して有効であることが確認済み。

Poetiqのメタシステムは、単一のモデルバージョン、ファミリー、サイズに縛られない。これは、将来登場するあらゆるLLMの能力を最大限に引き出す、汎用的な推論能力向上フレームワークである。



ブレークスルーを可能にした世界トップクラスのチーム



構成: Google DeepMind出身の研究者・エンジニア6名からなる少数精鋭チーム。

経験: DeepMindでの経験年数は合計53年。AlphaGo、Transformer、Geminiなどの開発で培われた専門知識。

開発速度: Gemini 3とGPT-5.1のリリースからわずか数時間で最新モデルをシステムに統合し、最高性能を達成。アーキテクチャの柔軟性と強力さの証明。

今後の課題と透明性のある評価



1. Public vs. Semi-Private評価

- 現状：今回の65%というスコアは「公開評価セット」での結果。
- 課題：ARC-AGI-1では、公開セットと非公開セットで性能が低下するモデルも存在。Poetiqのシステムもベースモデルの性能低下の影響を受ける可能性がある。
- 対応：公式評価が投稿され次第、更新情報を提供する予定。



2. コストと実用性のバランス

- 現状：Poetiqのコスト（約\$50/タスク）は人間（\$17/タスク）の約3倍。
- 目標：ARC Prize 2025では、\$0.42/タスクというコスト目標が設定されており、商業的実用化にはさらなる効率化が求められる。

AGI研究の新たなマイルストーン：単一モデルの時代から、システム協調の時代へ

**AGI達成への道は、必ずしも単一の超巨大モデルの開発にある
るのではない。Poetiqの成果は、複数の専門的なシステムを
協調させる「知的なアーキテクチャ」こそが、これまで不可能と
された抽象推論能力を解き放つ鍵であることを実証した。**

AGIはまだ完全に解決されたわけではない。

これは新たな、有望な道である。

「規模のみ」のアプローチから「システムとアーキテクチャ」
のアプローチへとパラダイムが移行している。