

AIが人間の抽象的推論を超えた日

Poetiq AI、難関ベンチマークARC-AGI-2で歴史的スコアを達成



75%の正解率。人間の平均を凌駕し、SOTAを15ポイント更新

75%

GPT-5.2 X-Highを搭載した
Poetiqシステムの公開評価ス
コア（2025年12月24日）

≈60%

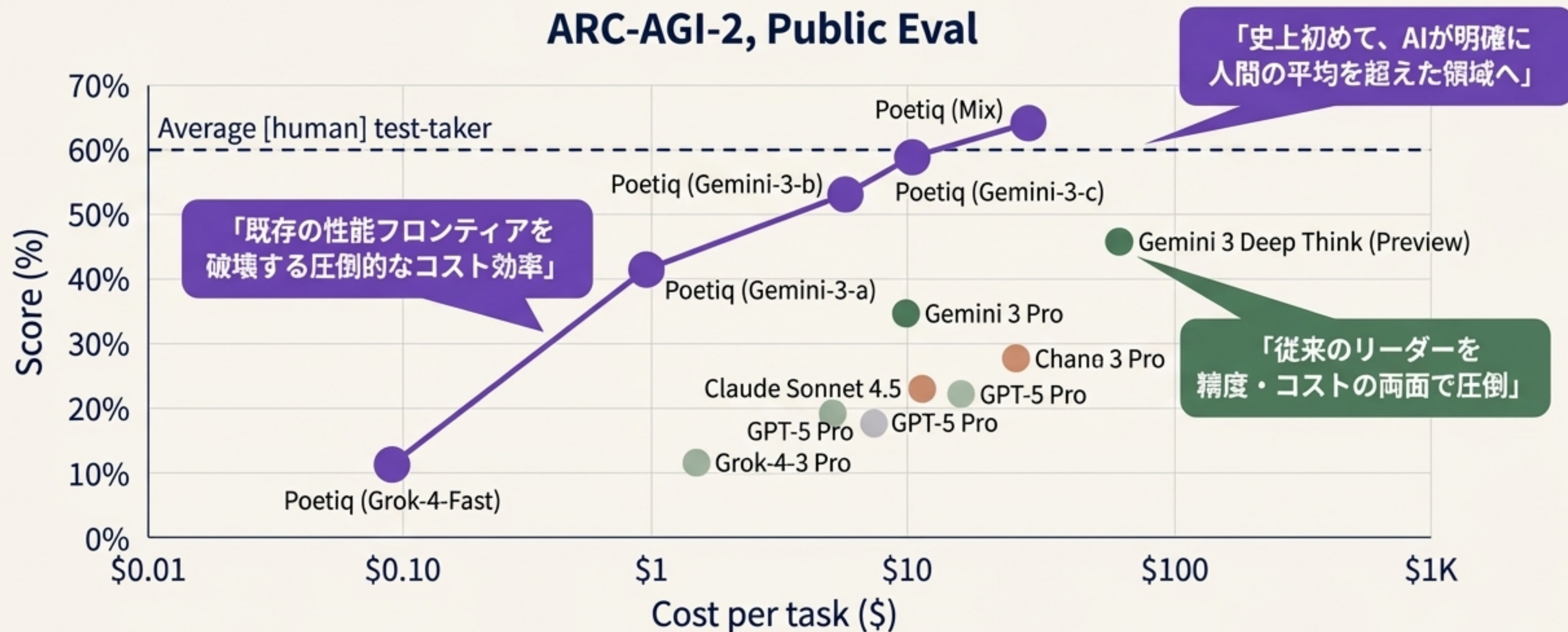
ARC-AGI-2における平均的
な人間の正解率

<\$8 / 問

75%のスコアを達成した際
の平均問題解決コスト

わずか1ヶ月前のSOTAを約15ポイント上回り、精度とコスト効率の両面で飛躍的な進歩を記録。Poetiqはモデルの追加学習やチューニングを一切行わず、既存の「ハーネス」に新モデルを組み込み込むだけでこの成果を達成したと報告。

性能・コスト曲線が示す、Poetiqによる「次元の違う」達成



Poetiqのシステムは、既存モデルを組み合わせるだけで性能とコストのトレードオフを根本から覆し、AI開発の新たなパラダイムを提示した。

公開評価から公式評価へ：現実的なスコア低下の予測

過去の実績（2025年11月）

公開評価スコア：
65.32%

公式評価スコア：
54.0%

約11ポイント
の低下

今回の予測

公開評価スコア：
75%

予測される公式スコア：
~64%?

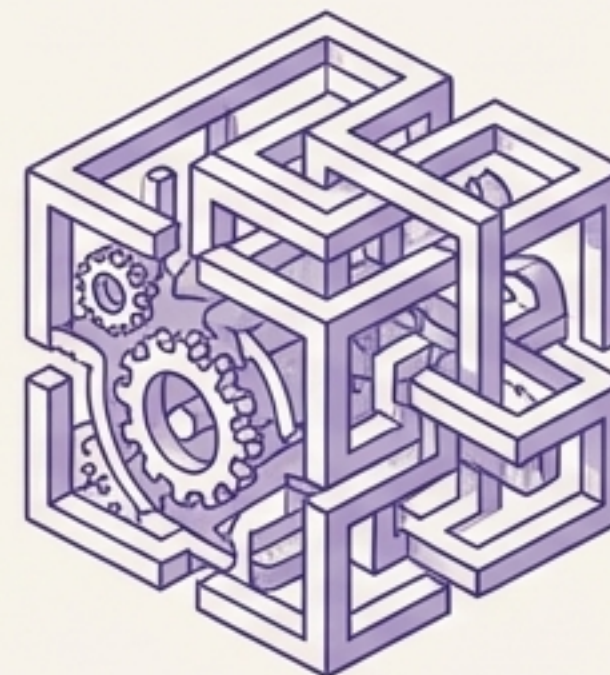
前回同様の低下が見られたとしても、公式スコアが60%台半ばに達する可能性は高い。これは依然として人間平均（~60%）を上回り、ARC-AGI-2における歴史的な最高記録となる。

ARC-AGI-2：AIの「真の推論能力」を測るために設計された難関

What is ARC-AGI-2?

- 人工一般知能（AGI）に近づくための抽象的推論力と適応的問題解決能力を測る評価基盤。
- 未知のパズル形式の課題で構成され、単なるパターンマッチングでは解けない。

The Key Challenge

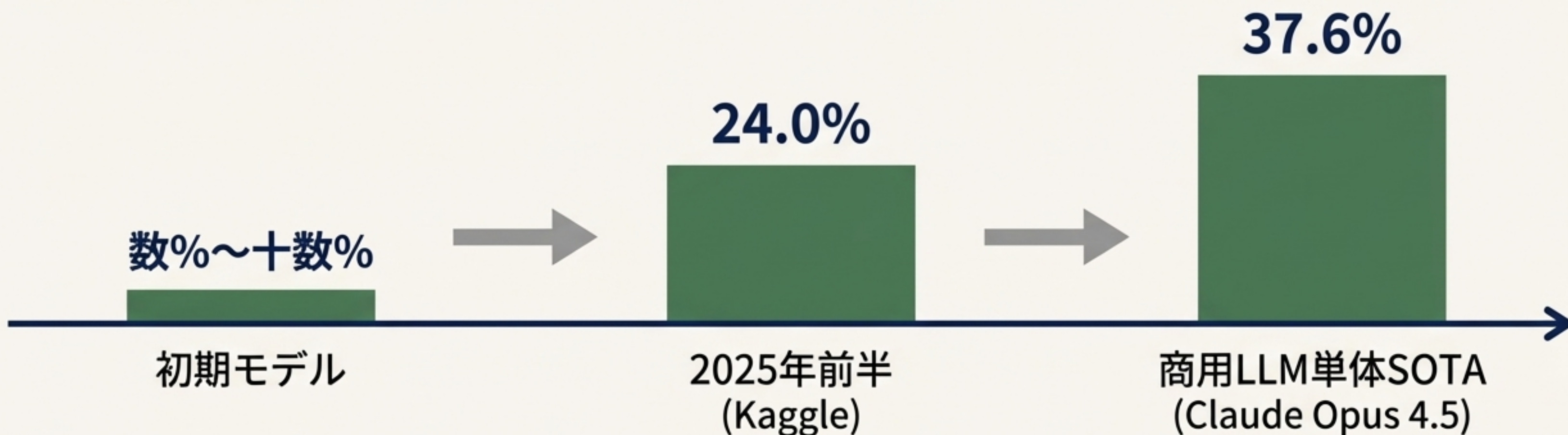


- 「訓練による上達」がほぼ不可能: 大量データによる過学習（ベンチマックス）を避け、初見の問題をその場で解く汎用性が試される。
- AIの純粋な推論能力と試行錯誤の効率がスコアを直接左右する。

Poetiq登場以前：「人間の壁」に阻まれていた最先端AI

人間の壁

人間平均 (~60%)

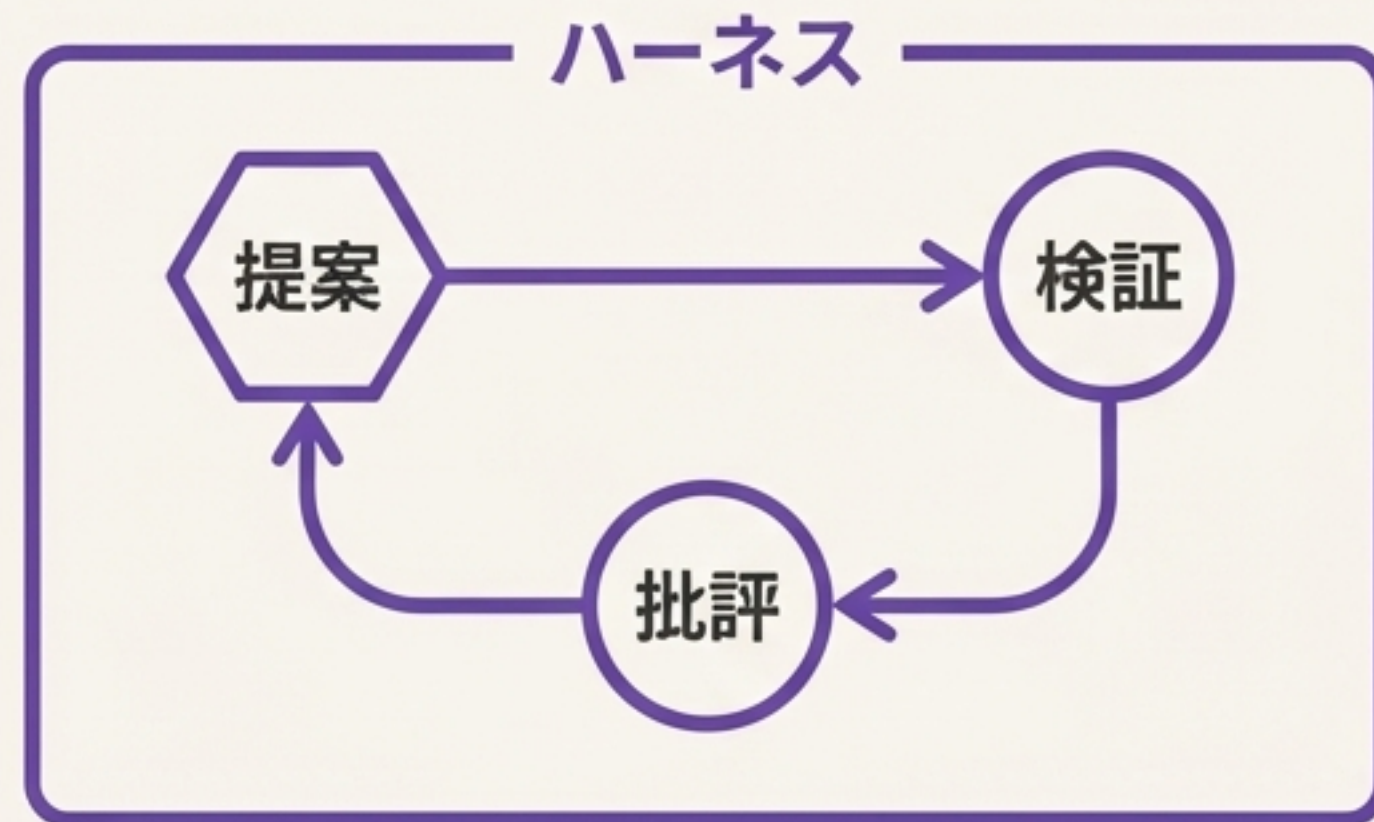


長らく、ARC-AGI-2は「AIが人間には到底及ばない領域」と見なされてきた。

Poetiqの秘密兵器「ハーネス」：モデルではなく、知性を構築するメタシステム

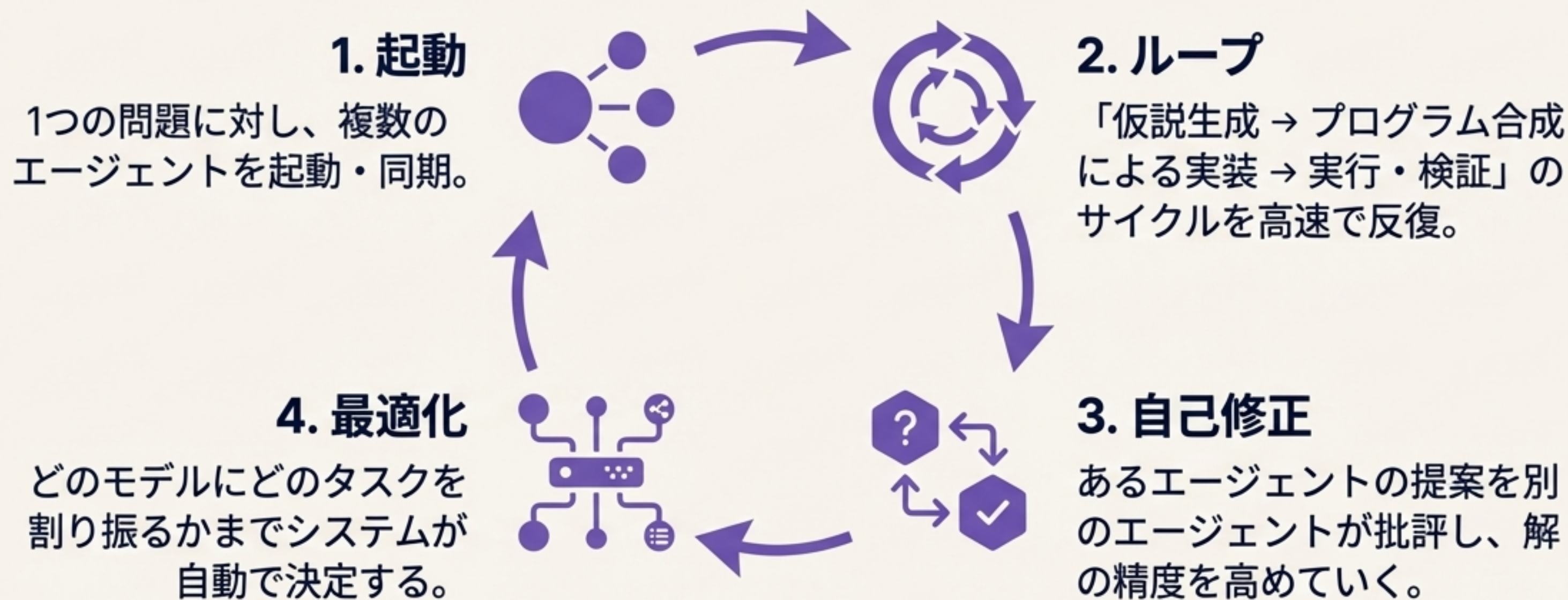
Poetiqの成果は新モデル開発ではなく、既存のLLMを部品として活用する特殊な解法フレームワーク「メタシステム（ハーネス）」によるもの。

「箱詰めのAI専門家チーム」



- 複数のAIエージェントが協調し、人間がチャットボットと対話するように、AI同士で提案・検証・批評のループを自律的に実行する。
- 単一LLMの限界を、自己修正を重ねることで突破する。

エージェント群のオーケストレーション： 仮説→実装→検証の自律的ループ



モデル非依存という強み：あらゆるLLMの潜在能力を解放



Poetiqハーネスは特定のモデルに依存しない。Gemini 3, GPT-5シリーズ, Claude, オープンソースモデル（12Bパラメータ）など、十数種類のモデルで性能向上を実証済み。

Efficiency Proof

- ClaudeやGrok等のモデルに適用した結果、多くの問題が平均「1回未満のリクエスト」で解決。
- これは、従来の「思考2回分を1回に凝縮する」ような高度な最適化戦略が機能していることを示唆する。

業界の覚醒：巨大モデル一辺倒から「賢い仕組み」の探求へ

小さなスタートアップの成果が、AGIへの道はモデルの巨大化だけでなく、アルゴリズム的工夫から生まれる可能性を証明した。

OpenAI

PoetiqにGPT-5.2を提供協力。自社モデルの能力を外部ハーンネスで引き出す実験に前向きな姿勢。

Google DeepMind

合議制エージェント開発を推進。マルチステップ推論を行う「Deep Research」APIを公開。

Anthropic

モデルをラップする工夫の重要性を認識し、合成データでの自己改善や外部ツール統合を模索。

新たな競争の構図：「巨大モデル」vs「賢いハーネス」

ARC-AGI-2をめぐる競争は、AIの能力をどう引き出すかという方法論の戦いへ移行した。

巨大モデル



賢いハーネス



← **Leader: Poetiq** AIのアプローチが明確に一步リード。

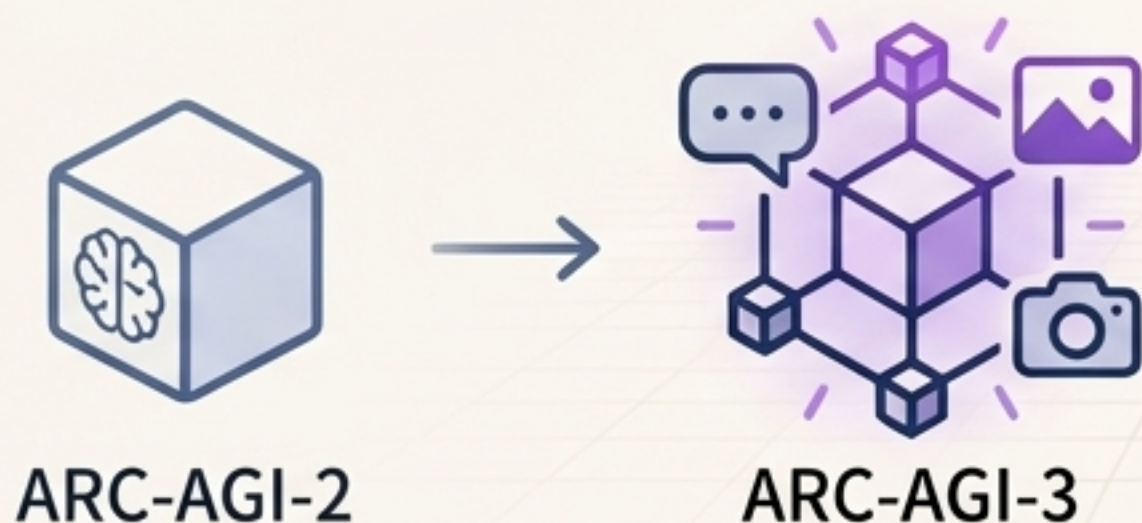
← **Challengers:** Google, OpenAI, Anthropicなどが類似のアプローチで追随。xAIのJeremy Berman氏のような個人研究者も奮闘するが、最強クラスのモデルを取り込めるPoetiqが大きく引き離す。

ARC-AGI-2は単なる競技会を超え、次世代AI開発の方向性を示す羅針盤となっている。

次なるフロンティア：ARC-AGI-3と「ハーネスの時代」の到来

ベンチマークの進化

ARC-AGI-2は想定より1年という速さで人間水準が突破された。



ARC-AGI-3が2026年第1四半期にリリース予定。対話型の長期課題やマルチモーダル推論など、より複雑な問題が想定される。

到来するトレンド

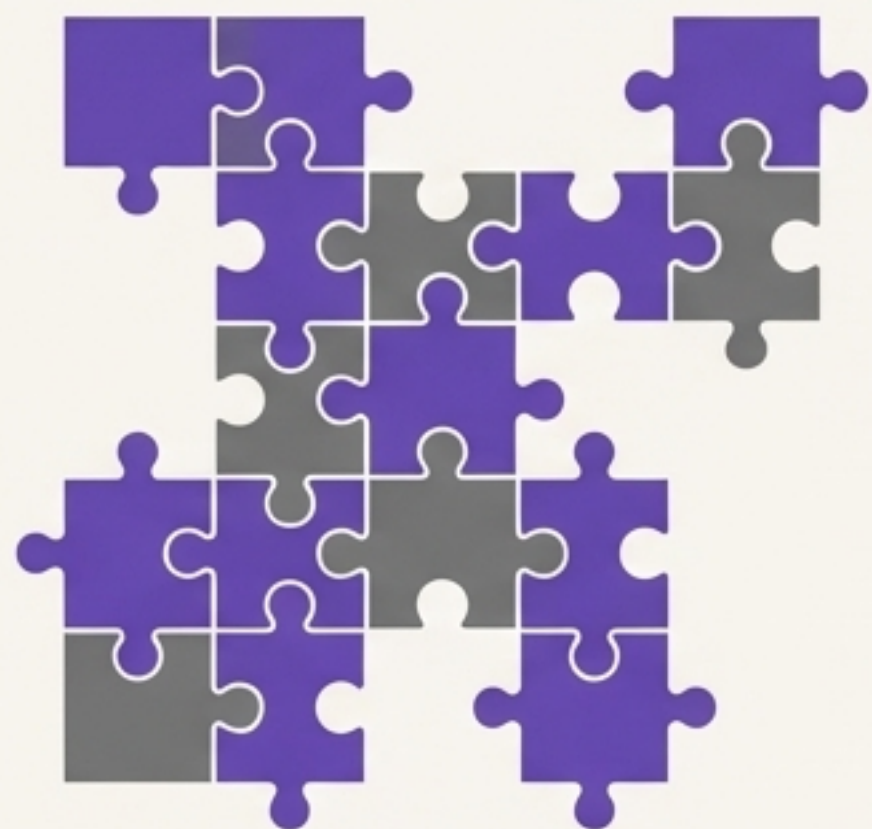
業界では「来年はハーネスの年になる」との声。



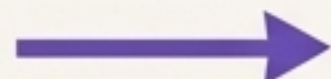
2026年には、各社から推論オーケストレーション技術が相次いで発表され、AI同士が対話しながら問題解決するシステムが標準になる可能性がある。

抽象パズルから現実世界の課題解決へ

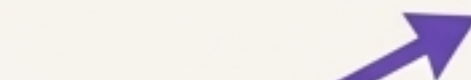
ARC-AGIで培われた技術は、現実世界の複雑な課題解決に応用可能。



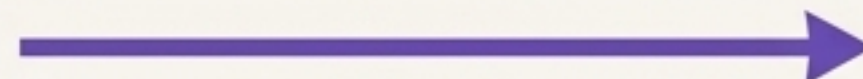
ARC-AGI の抽象的な課題



Poetiqのメタシステム
(ハーネス)



科学的発見



創造的設計



複雑な計画立案

複雑な計画立案、学術的発見、創造的設計など、これまでAIには困難だった領域でのブレークスルー。

Poetiqは「我々のメタシステムは複雑な実世界の問題にも適用可能」と述べ、応用研究を既に開始している。
わずか6名の精鋭チームが示した新たな道筋は、AGI開発競争の未来を大きく左右するだろう。