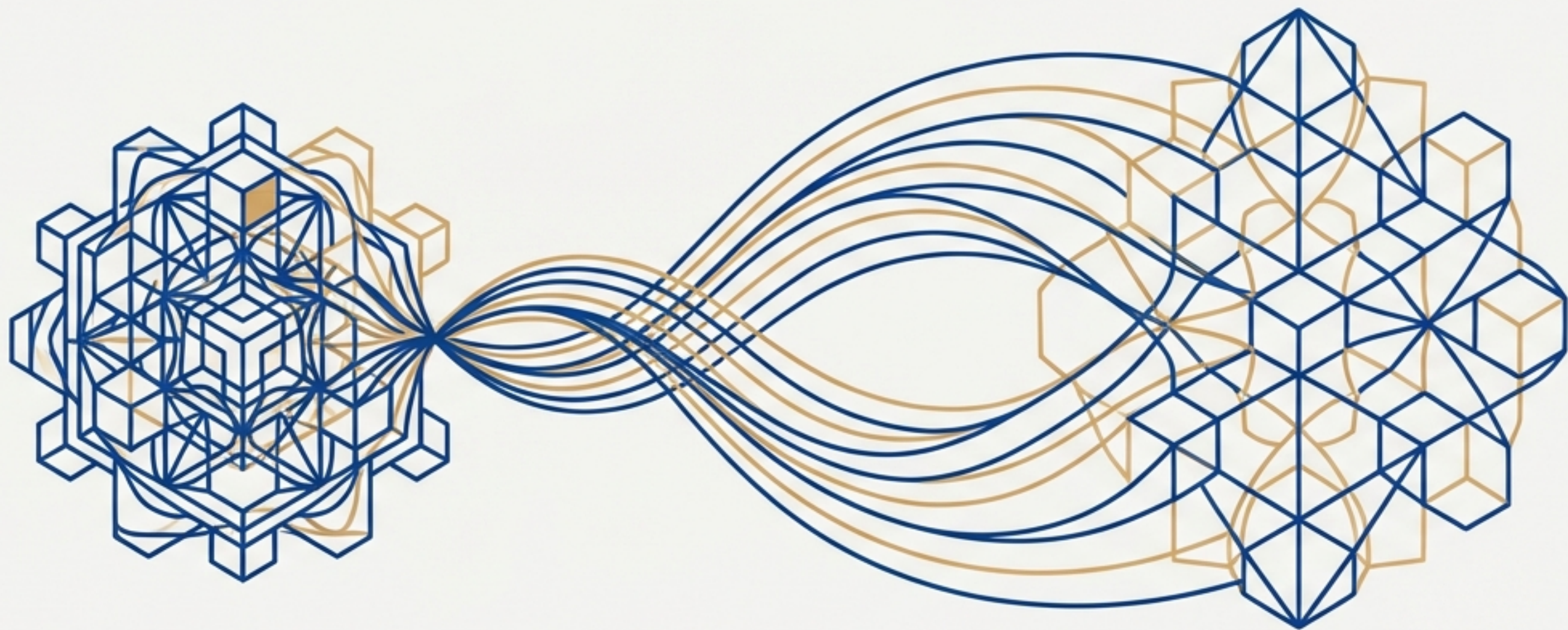


Poetiqのブレークスルー：AIの推論と経済性におけるパラダイムシフト

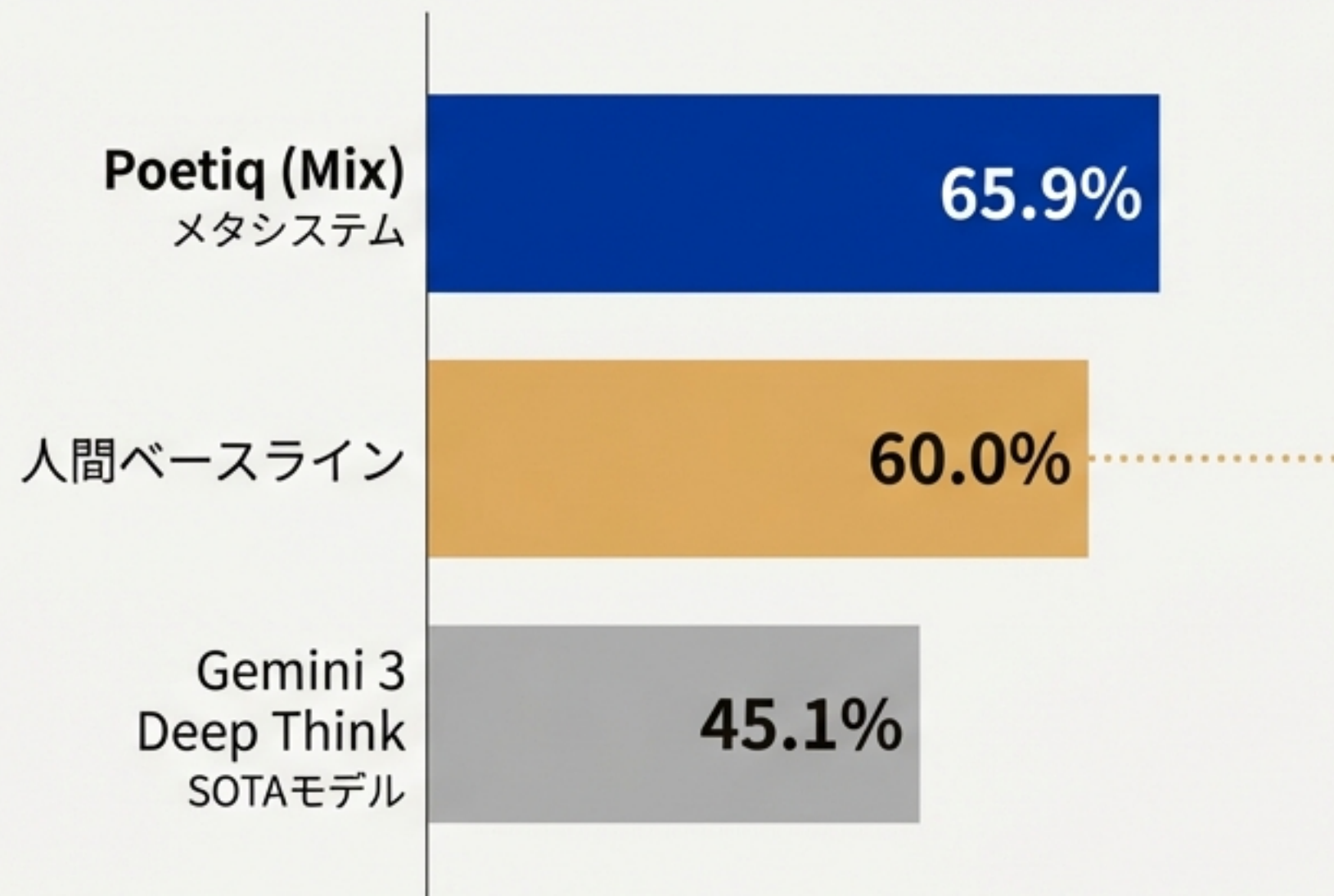
Poetiq (Mix)メタシステムとARC-AGI-2ベンチマークにおける人間超えの性能に関する分析



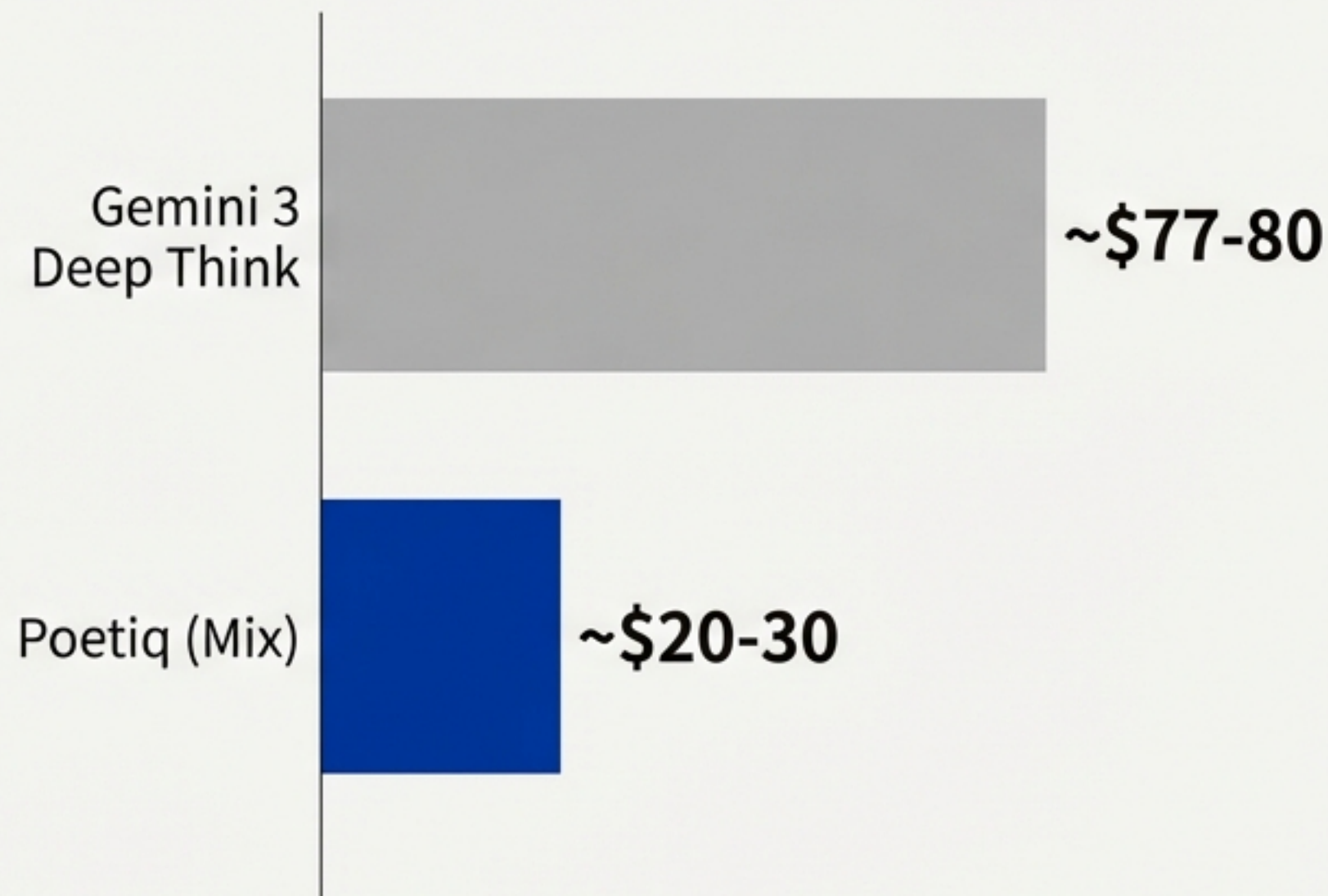
2025年11月28日

新システムが抽象的推論で人間を超越—しかも、わずかなコストで。

精度比較 (ARC-AGI-2)



コスト比較 (Taskあたり)



全データは2025年11月下旬時点の公開情報に基づく。Poetiqのスコアは自己報告値。

課題の核心：ARC-AGI-2はAIによる「攻略」が極めて困難なように設計された。



暗記への対抗 (Anti-Memorization) : MMLU等とは異なり、インターネット知識の暗記では解けない独自の視覚パズル。フューショット学習能力を測定。



流動性知能の測定 (Measures Fluid Intelligence) : 結晶性知能（知識の検索）ではなく、新たな状況に適応し、抽象的ルールを推論する能力を問う。



人間にとっての容易性 (Easy for Humans) : AIには難しいが、人間には直感的。「コア知識」のみで解けるよう設計されており、60%の平均スコアがタスクの妥当性を保証。

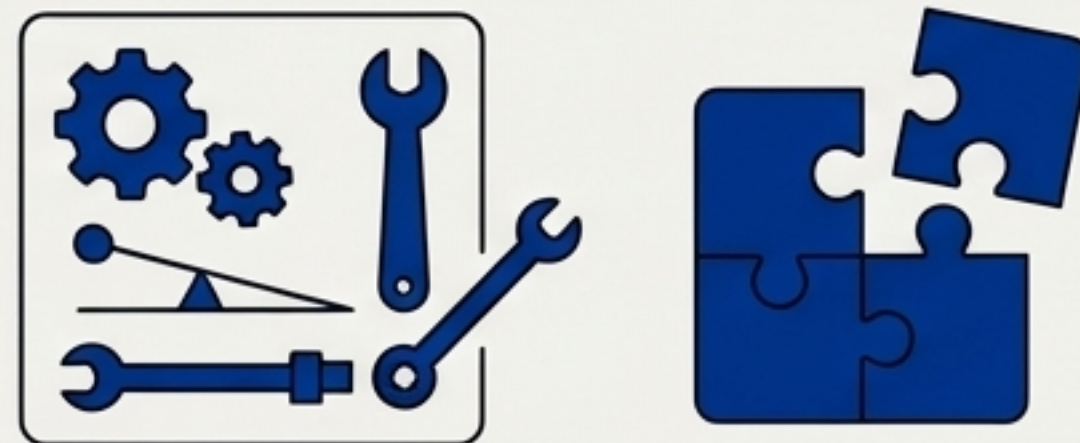


60%の分水嶺 (The 60% Watershed) : この閾値を超えることは、AIが真の汎用推論能力において人間に並んだことを示す強力なシグナルと見なされていた。

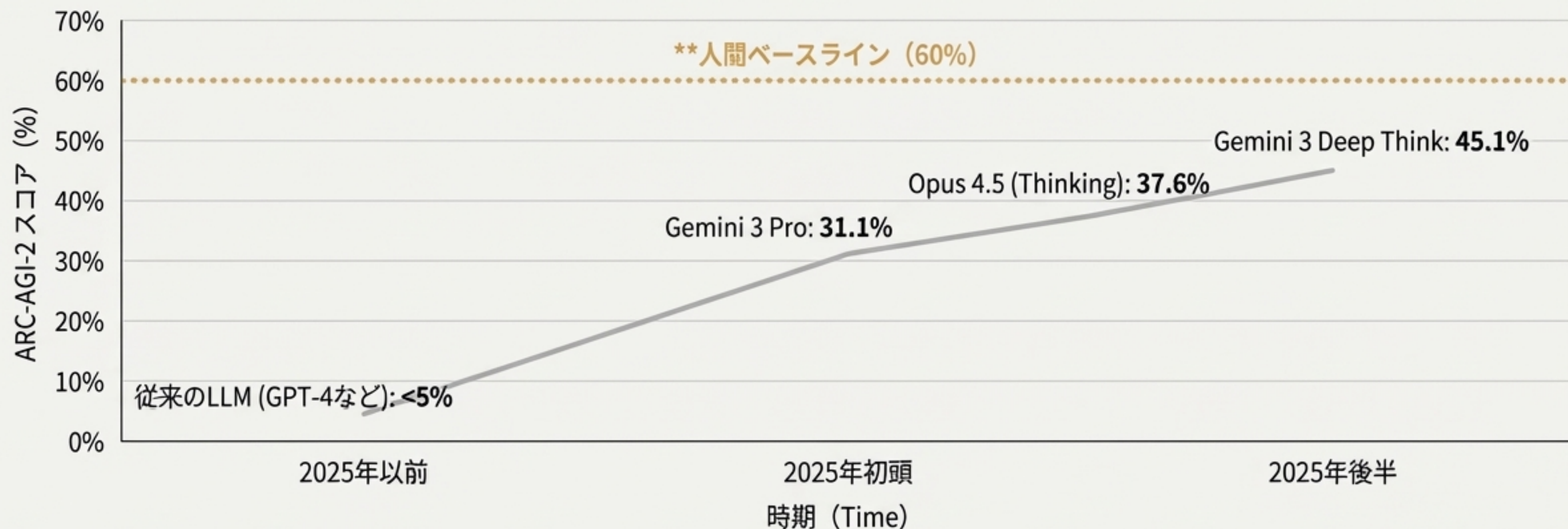
結晶性知能 (Crystallized Intelligence)



流動性知能 (Fluid Intelligence)



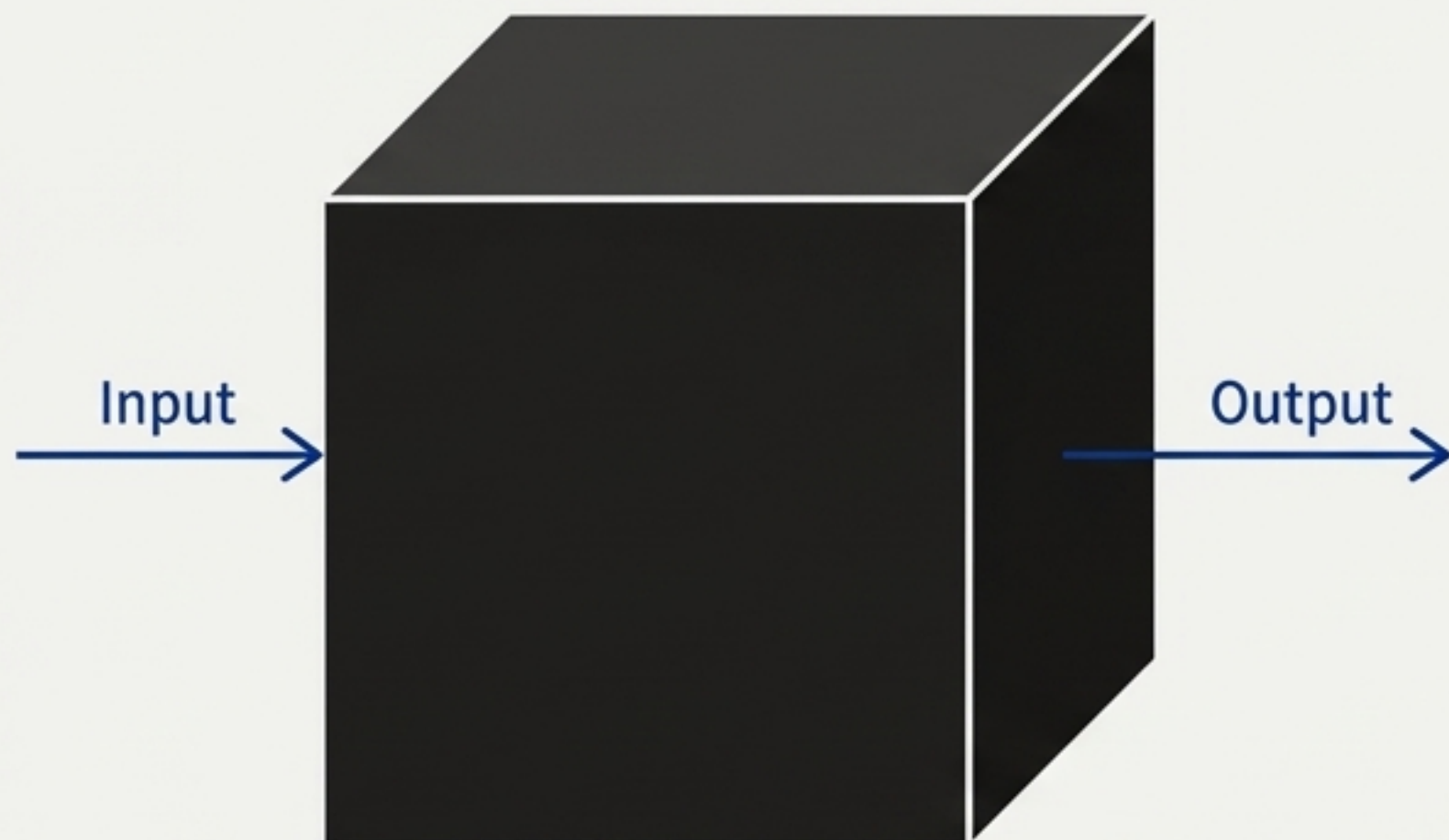
Poetiq以前、進歩は漸進的でありフロンティアモデルでさえ人間の閾値には遠く及ばなかった。



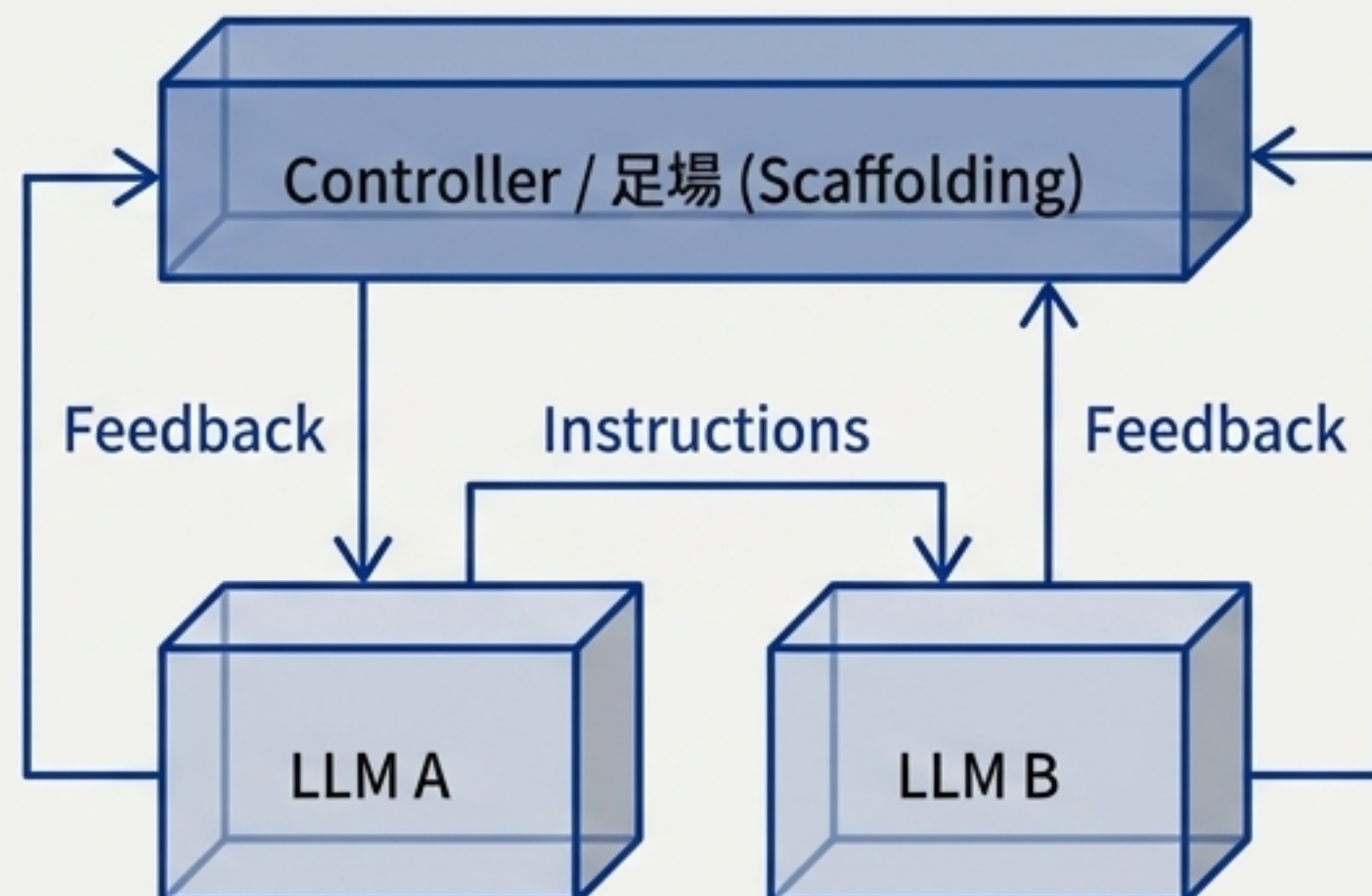
2025年11月以前、スコアは一桁台の改善が標準であり、45%の壁を越えることは大きな課題とされていた。

ブレイクスルーは新モデルではなく、既存のAIを統合管理する「メタシステム」である。

モノリシックモデル / Monolithic Model

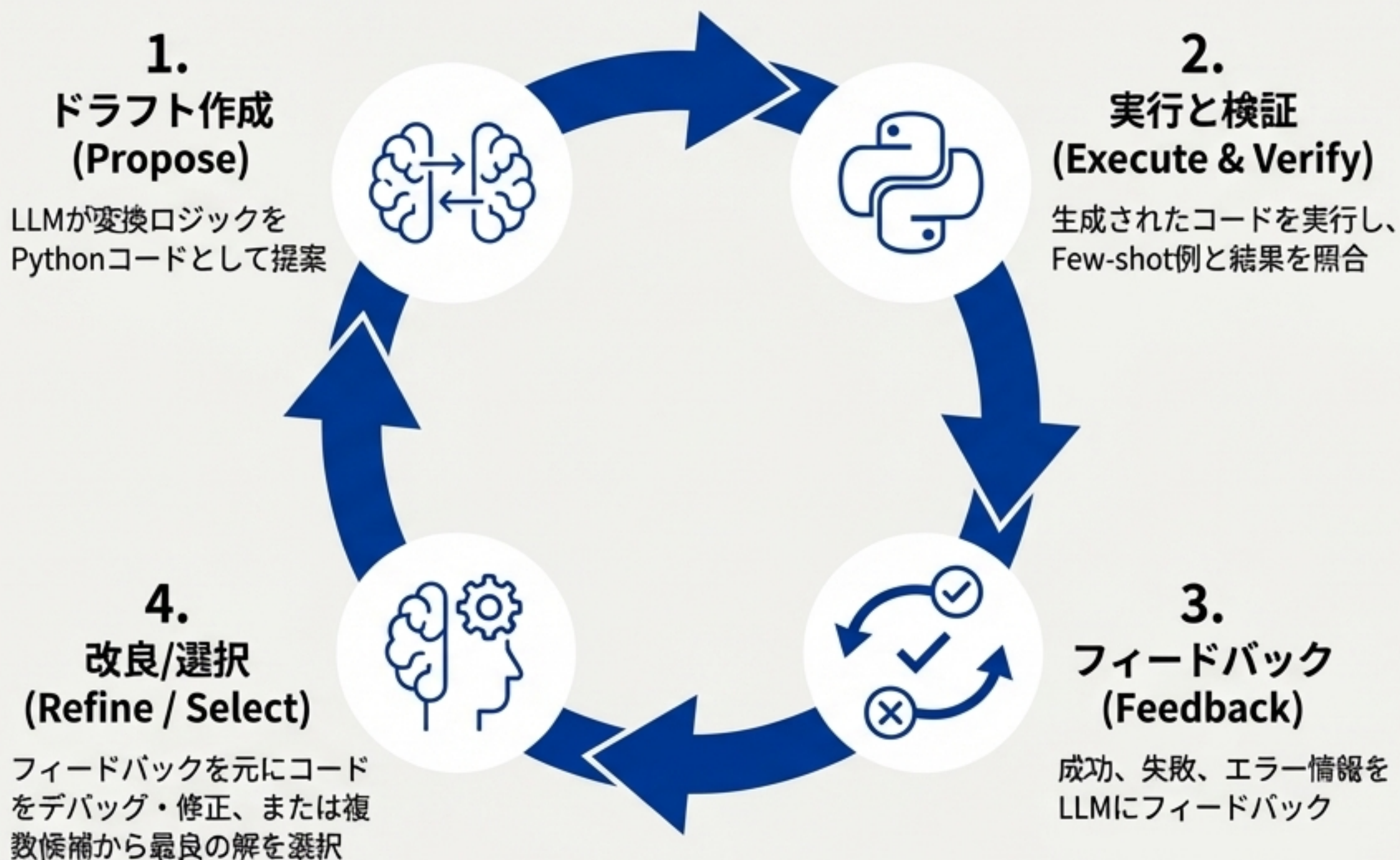


メタシステム / Meta-System



- Poetiqはゼロから訓練されたモデルではない。
- 既存のフロンティアモデル（Gemini 3, GPT-5.1）を、交換可能な「推論エンジン」として活用する認知アーキテクチャである。
- この「スキャフォールディング（足場）」アプローチが、モデル単体では不可能な能力を引き出す。

Poetiqのエンジン：コード生成、実行、自己修正の再帰的ループ。



KEY INSIGHT

このプロセスは、推論を曖昧な自然言語から、検証可能なPythonコードへと変換する。「思考の連鎖」を外部化し、決定論的なテストを可能にしている。

「Mix」アプローチは複数モデルの相補的な強みを活用し、堅牢性を高める。



Google Gemini 3

強み (Strengths)

高度なマルチモーダル能力、視覚的推論。「Deep Think」モードによる深い思考。

ARC-AGI-2単体スコア

45.1%



- **LLM非依存 (LLM-Agnostic)**
システムは下層エンジンを自由に交換可能。
- **アンサンブル効果 (Ensemble Effect)**
異なるモデルのバイアスを相殺し、多様な解候補を生成。「洗練された投票メカニズム」で最も堅牢な解を選択し、単一モデルの限界を超える。



OpenAI GPT-5.1

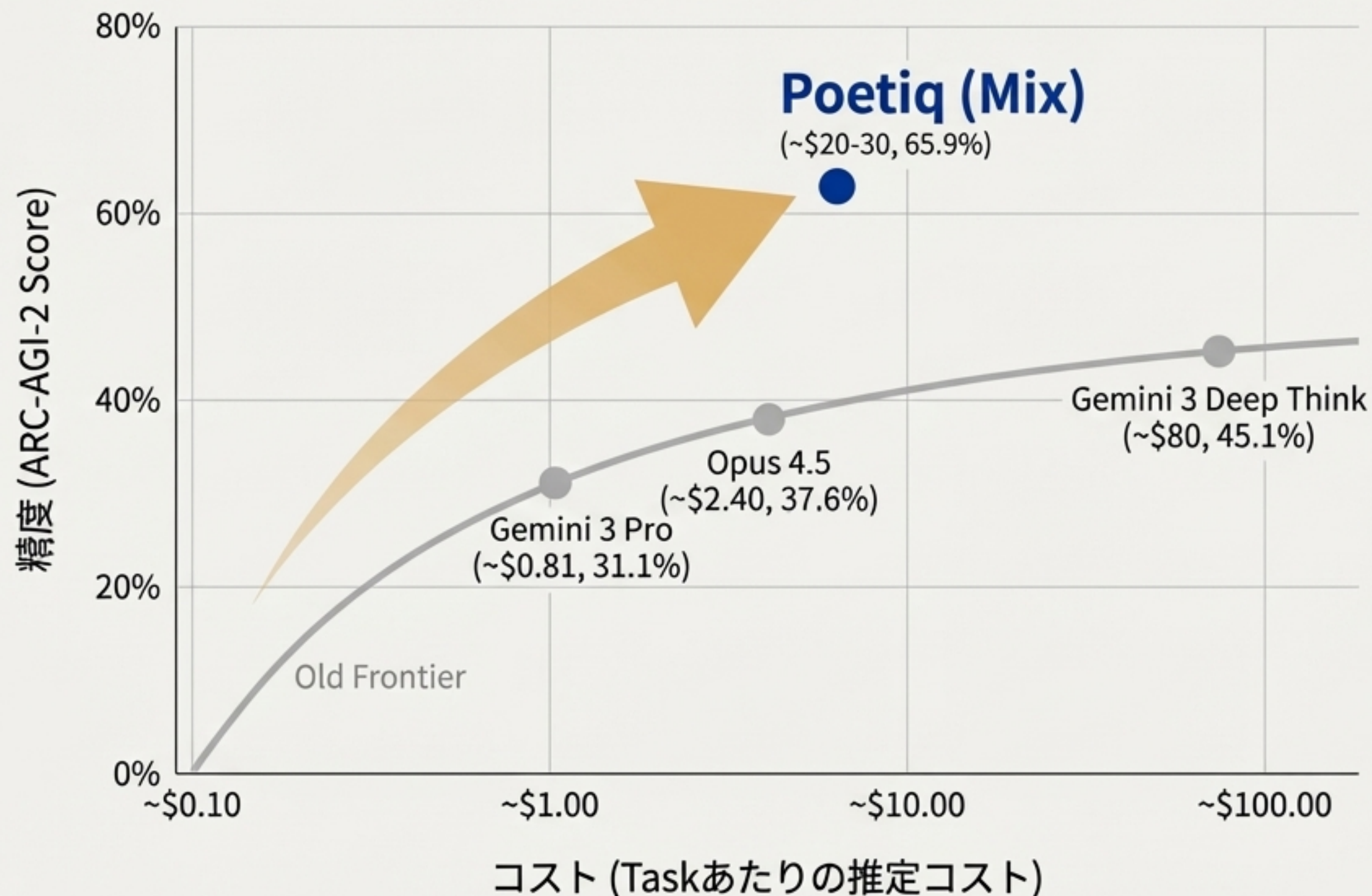
強み (Strengths)

優れた指示追従能力とコーディング能力。Poetiqのプログラム合成パラダイムの中核を担う。

ARC-AGI-2単体スコア

17.6%

Poetiqは新たなパレートフロンティアを創造： 圧倒的に低いコストで、より高い精度を達成。



Poetiqは単に曲線上で移動したのではない。曲線全体を「より安く、より良く」という方向にシフトさせ、「スキャフォールディング」が付加する価値がAPIコストを上回ることを証明した。

効率性の秘訣は「自己監査」：有効な解を発見した瞬間に計算を停止する。

思考時間予算 (Thinking Time Budget)

Gemini 3
Deep Think
(固定的な計算)

スコア最大化のため、固定された長時間の思考を実行。コスト: ~\$80

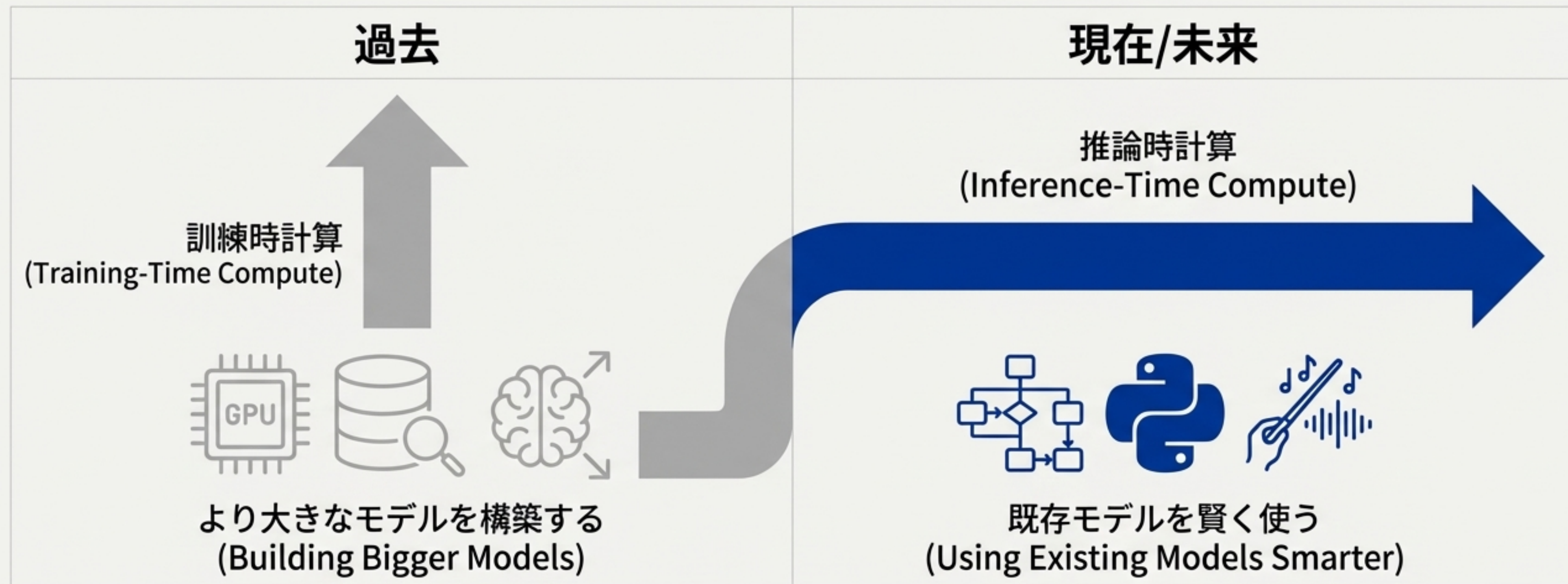
Key Insight: Poetiqのコスト優位性は、推論リソースを動的かつ目的に応じて管理する能力に由来する。

Poetiq (Mix)
(動的な計算)



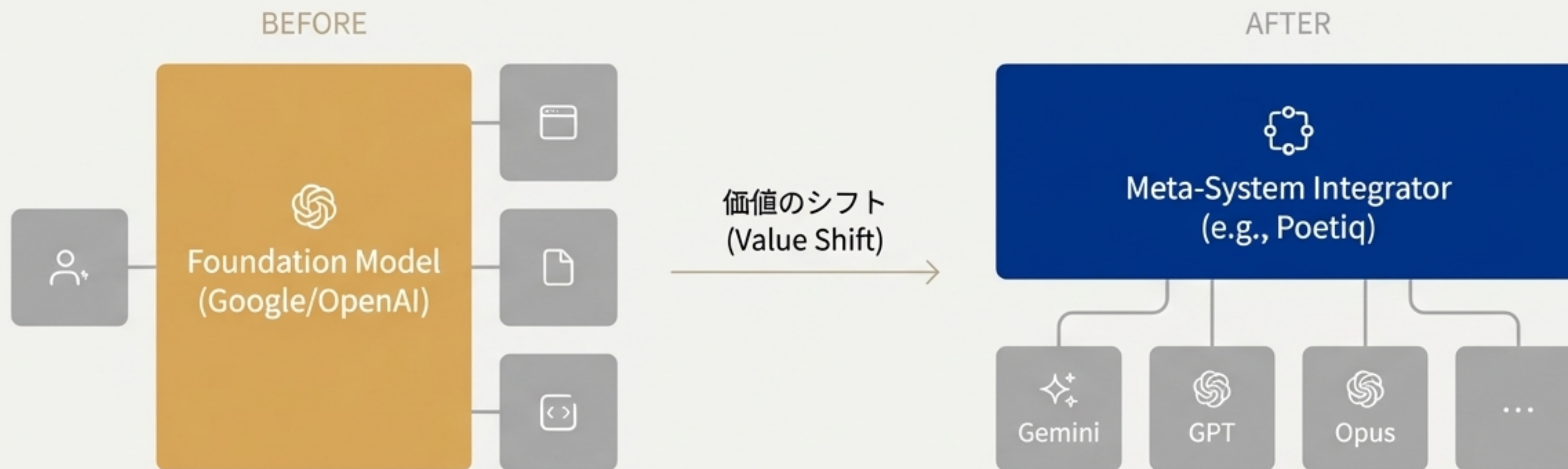
Python実行により正解を検証。解が見つかった瞬間に計算を打ち切り、無駄なAPIコールを削減。コスト: ~\$20-30

業界の焦点は「モデルのスケーリング」から 「システムのスケーリング」へ移行している。



2025年は「推論時計算」の時代の幕開けである。Poetiqは、GPT-6を待たずとも、既存のモデルに適切な「思考時間」と「道具 (Python)」を与えることで、AGIレベルの問題を解決できることを示した。

メタシステムの台頭により、基盤モデルは交換可能な「認知的コモディティ」と化す可能性がある。



- Poetiqのようなシステムが下層モデルを自由に組み合わせられるなら、モデル自体は「部品」になる。
- AI業界の価値の源泉は、モデル作成者から、それらを統合して最高の結果を出す認知アーキテクチャ設計者へとシフトする可能性がある。
- This empowers a new ecosystem of innovation on top of the foundational layer.

Poetiqの外部的で検証可能なループは、Geminiの内部的で不透明な推論を凌駕する。

特徴 (Feature)	Poetiq (Mix)	Gemini 3 Deep Think
タイプ (Type)	メタシステム (ソフトウェアラッパー)	統合モデルモード (End-to-End)
推論媒体 (Medium)	Pythonコード (検証可能)	内部トークン (不透明)
エラー修正 (Error Fix)	明示的 (Pythonエラーを検知し再試行)	暗黙的 (内部アテンション)
停止条件 (Stop)	動的 (自己監査による早期終了)	固定的 (計算予算に基づく)
コスト (Cost Profile)	~\$20-30 (効率的)	~\$77-80 (計算集約的)
スコア (Score)	~65.9%	45.1%

Concluding Insight: Poetiqの優位性は**検証問題 (Verification Problem)**の解決にある。Pythonインタプリタを「敵対的な検証者」として利用することで、思考の誤りを提出前に検知し、幻覚を排除している。

これは真のブレークスルーか？ 注意点と反論を検証する。

懐疑論①: パブリック評価セットへの過学習では？ (Skepticism 1: Overfitting to the Public Set?)

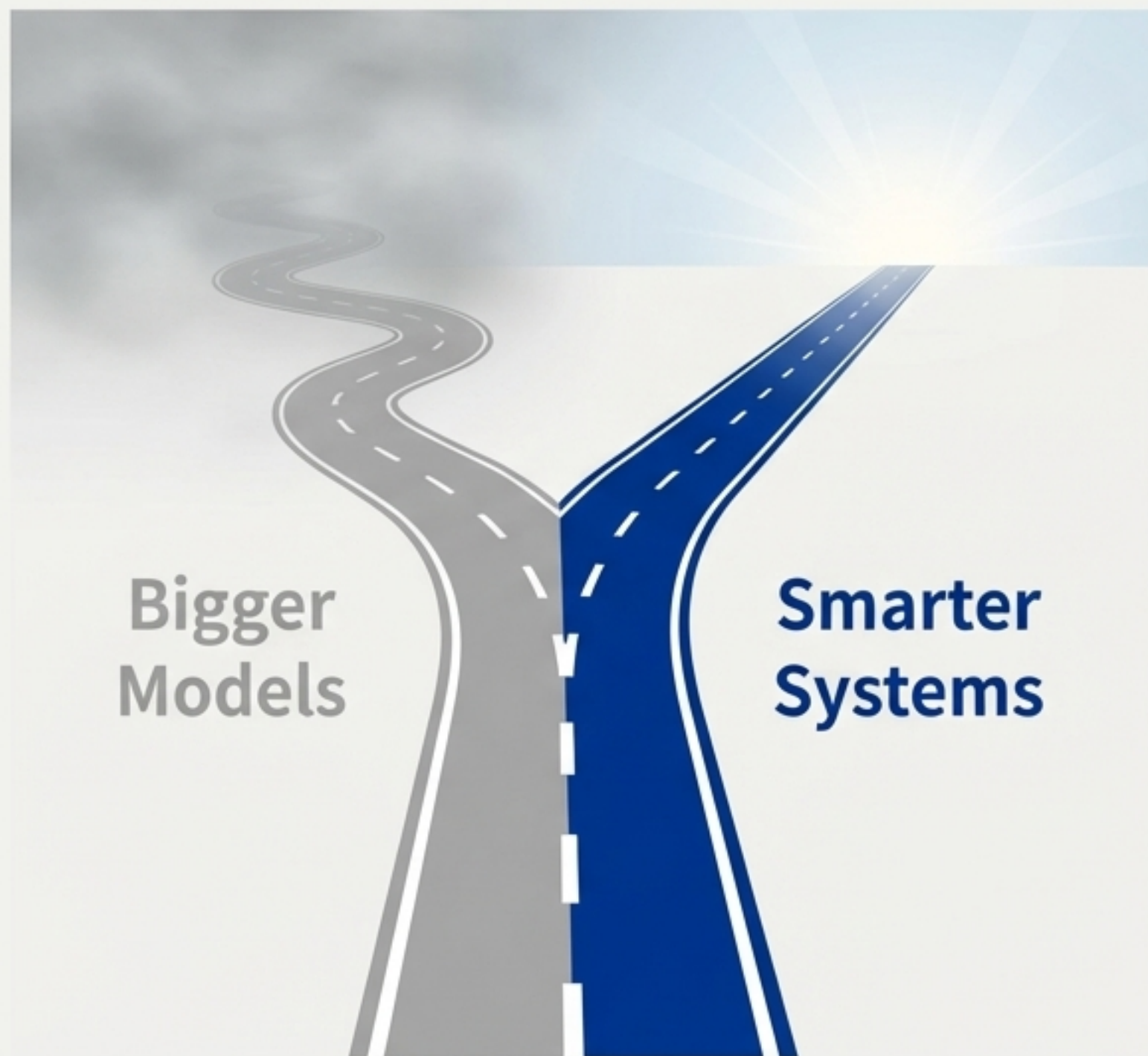
- 主張 (Argument): 報告された65%は公開データに対するものであり、非公開のプライベートセットでの真の汎用性は未知数。「ベンチマキシシング」のリスク。
- Poetiqの反論 (Counterargument): システムの適応はGemini 3/GPT-5.1リリース前に行われた。ARC-AGI-2の問題自体は^{事前}学習に使用しておらず、汎用的な推論メカニズムが機能した結果である（ゼロショット適応）。



懐疑論②: スコアの信頼性は？ (Skepticism 2: Reliability of the Score?)

- 主張 (Argument): スコアは自己報告であり、第三者による完全な検証が待たれる。
- Poetiqの反論 (Counterargument): 透明性を確保するため、ソルバーのコードをオープンソース化し、第三者による検証を可能にしている。

AGIは「解決」されたのか？—否、 しかし、未来への道筋は根本的に変わった。



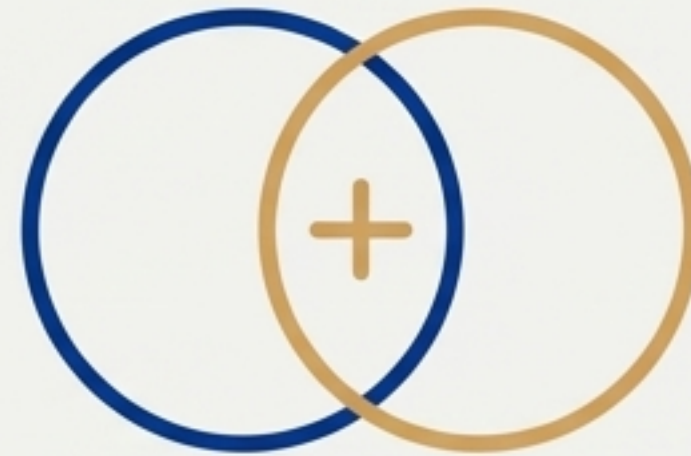
- 「解決」の定義 (Defining "Solved") :
「人間平均を超える」という意味ではYes。しかし、真のAGIは視覚パズルから小説執筆まで、ドメイン間をシームレスに移動する必要がある。
- 残されたギャップ (The Remaining Gap) :
65%は人間平均を超えたが、人間のエキスパート (90-100%) には及ばない。また、35%は依然として誤答している。
- 根本的な変化 (The Fundamental Change) :
Poetiqは、特定の複雑なドメインにおいて人間超えの性能を達成するための、強力で検証済みの**青写真 (blueprint)**を提供した。これが最大の貢献である。

Poetiqのブレークスルーが示す3つの重要な示唆。



1. システムが知能を解き放つ (Systems Unlock Intelligence)

既存のモデルでも、優れたシステムアーキテクチャ（足場）があれば、人間超えの推論能力を発揮できる。



2. 未来はハイブリッドである (The Future is Hybrid)

複数モデルと外部ツール（コード実行など）の統合は、もはやニッチな技術ではなく、SOTAを達成するための新たな標準となる。



3. 経済性が逆転する (Economics Are Inverted)

より高い知能が、より低いコストで実現可能になった。「推論時計算」の賢い管理が、新たな競争優位性の源泉となる。