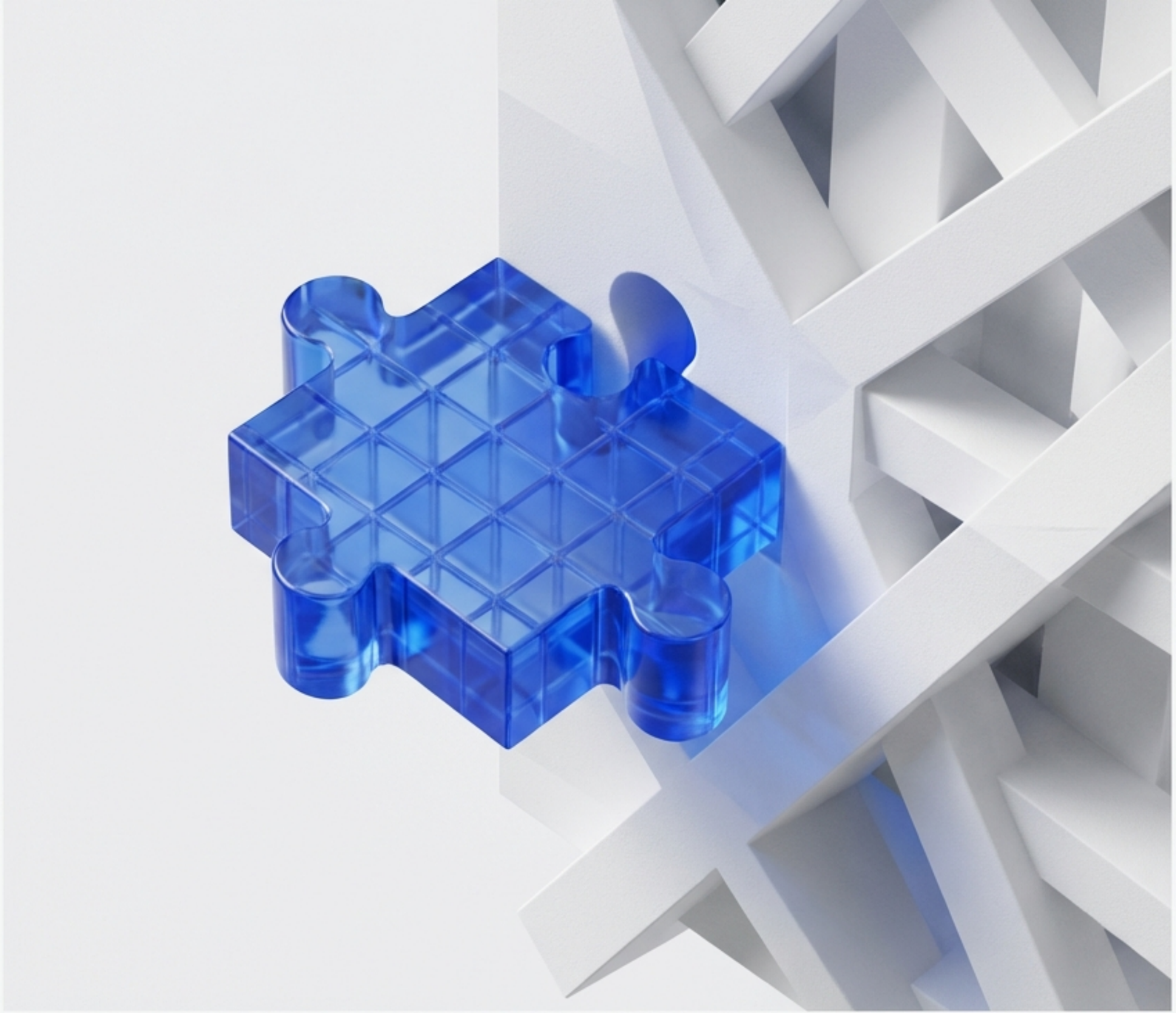


ARC-AGI-2とAI汎用性評価：流動性知能の壁への挑戦

Johan Land氏は、いかにして「多モデル反映的推論」で人間基準（72.9%）を突破したか

THE CONTEXT

Benchmark: ARC-AGI-2 (Abstraction and Reasoning Corpus)

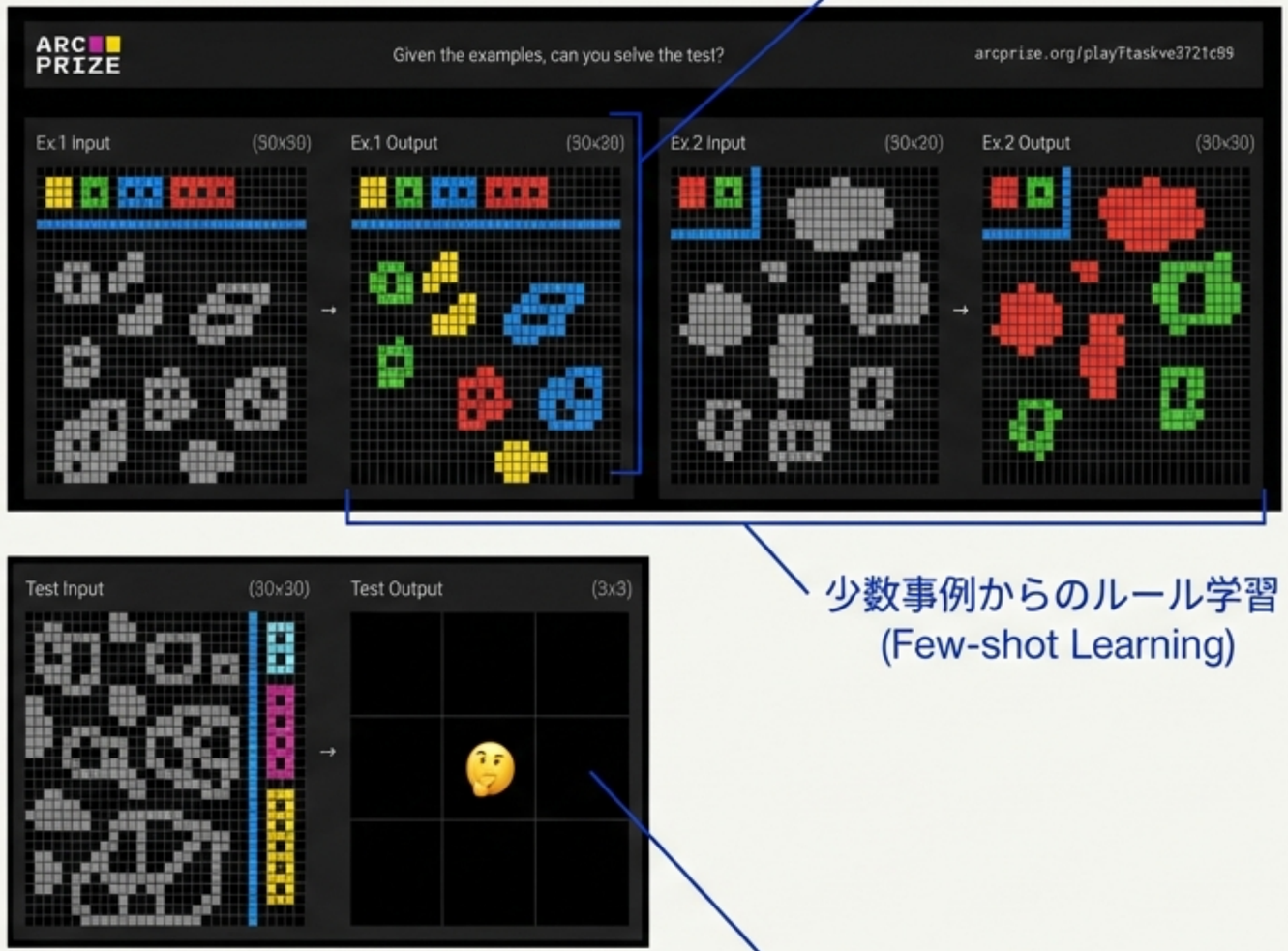
Challenge: 流動性知能 (Fluid Intelligence) - 未知のルールへの適応

Breakthrough: 72.9% Score (Beating Human Baseline)

「人間には易しく、AIには難しい」：ARC-AGI-2の正体

ARC-AGI-2 は知識量を測るテストではありません。「象徴的解釈 (Symbolic Interpretation)」イバターク」と呼ばれる、 抽象的なパターン認識とルール適応応能力を測定します。
これまでのAIは記憶に依存していましたが、このベンチマークは事前知識を排除し、「その場での推論 (On-the-spot reasoning)」を要求します。人間は直感的に解けますが、総当り攻撃や単純なパターンマッチングでは解けない設計になっています。

Helvetica Now Display, : International Klein Blue



少数事例からのルール学習
(Few-shot Learning)

未知の入力への適応
(Novel Adaptation)

記憶 vs 推論：埋まらなかった「流動性知能ギャップ」

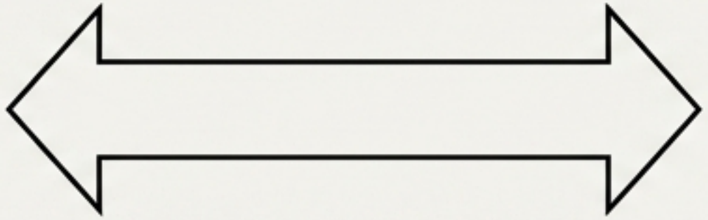
Standard LLM Baseline (2025)

0% - 17.6%

既存の言語モデル
(GPT-4.5 / GPT-5.1)

記憶に基づくタスクは得意だが、
未知の推論に失敗

The Fluid Intelligence Gap
（流動性知能の壁）



Human Average

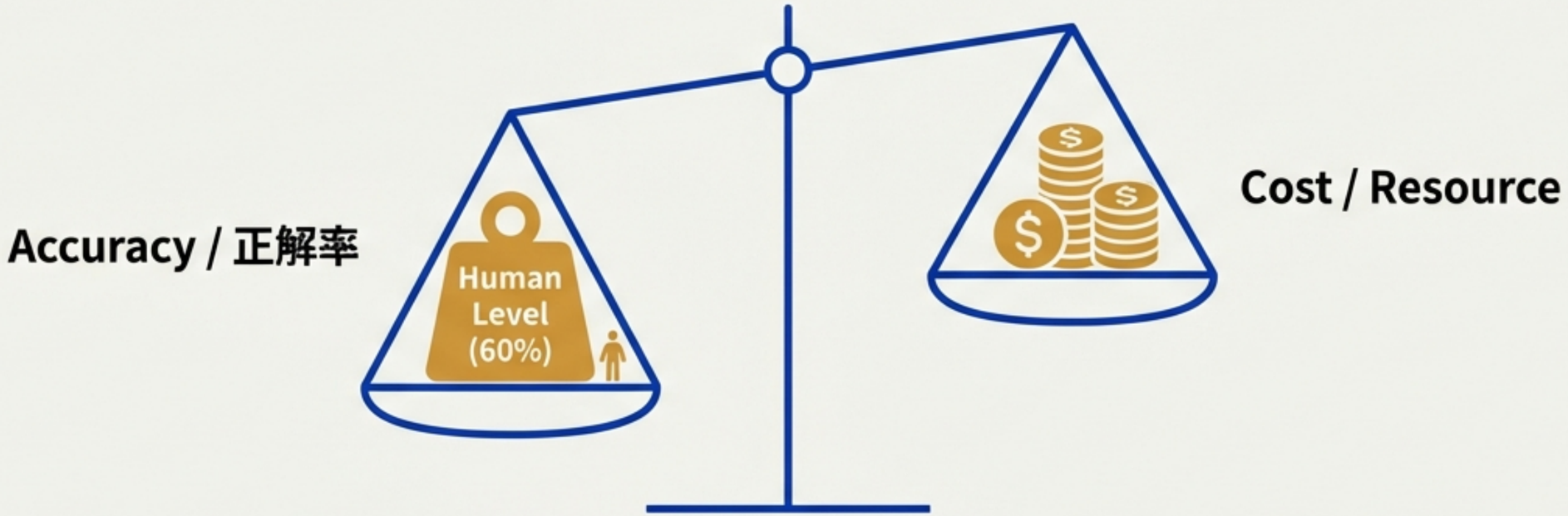
~60%

人間
(一般参加者400名)

直感的な流動性知能による解決

モデルサイズを拡大しても、このギャップは埋まらなかった。

知能のコスト：正確さと効率のトレードオフ



		Time	Cost	Efficiency
1	**Human (人間)**	2.3 min / task	$17 / task	High ↗
2	**Early AI Attempts**	Seconds (Instant)	High GPU Load	Low (<20%) ↘

真のAGIは、単に解けるだけでなく、効率的（Intelligence per $）でなければならない。

2025年の巨神たち：システムを構成する最強のモデル群

THE BRAIN

GPT-5.2 (OpenAI)

Planner / Logic

- 論理的推論と長文脈の分析
- 「Thinking Mode」による深い思考
- 推論の骨子（Chain of Thought）を構築

THE EYES

Gemini-3 (Google DeepMind)

Vision / Multimodal

- ネイティブなマルチモーダル理解
- 視覚的パターンの認識と計画
- 画像、テキスト、コードの統合

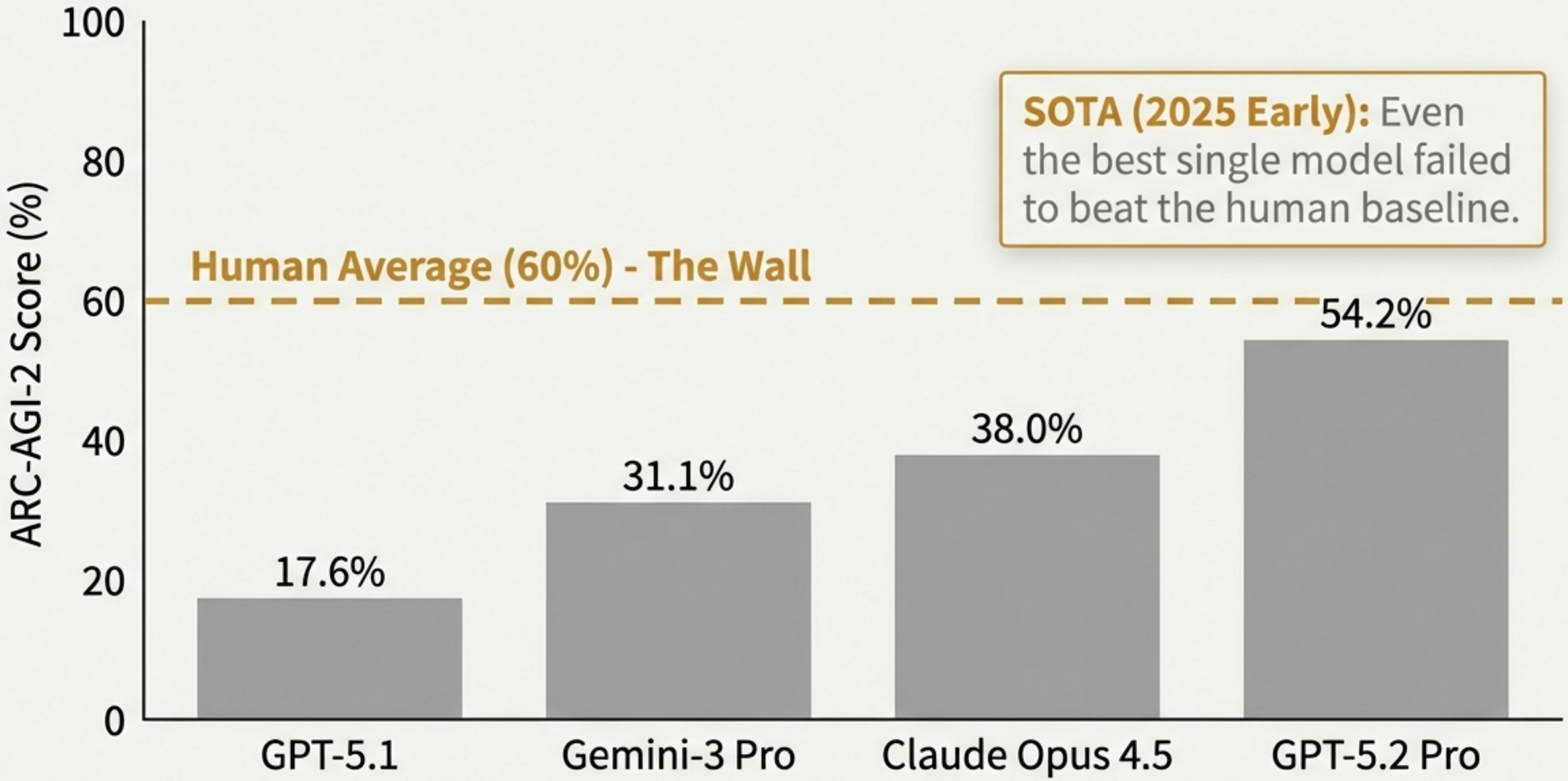
THE HANDS

Claude 4.5 Opus (Anthropic)

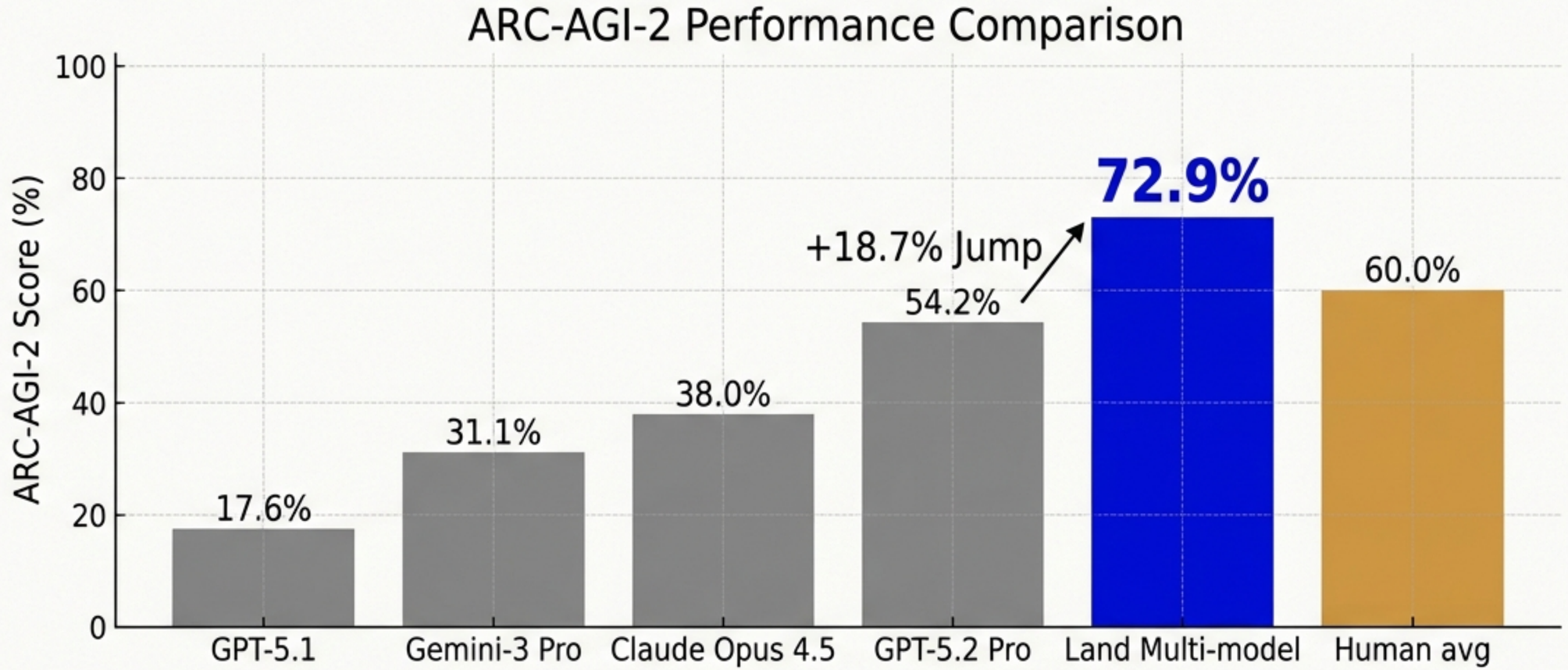
Coder / Engineer

- コーディングとデバッグ能力
- 長時間の自律的エージェント動作
- 実装とエラー修正のループを担当

限界点：単体モデルが直面した「60%の壁」

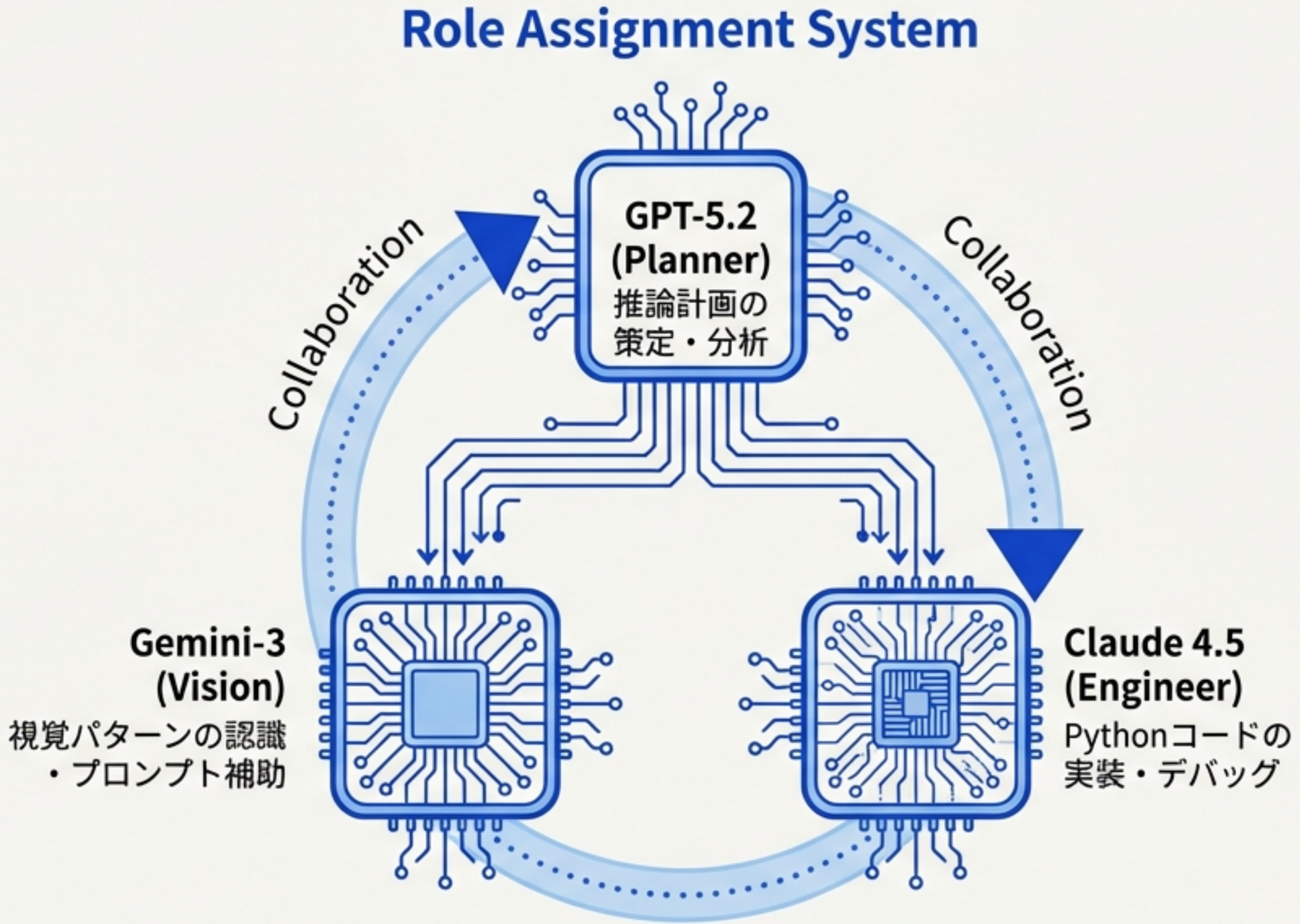


ブレイクスルー：人間基準を粉砕した72.9%



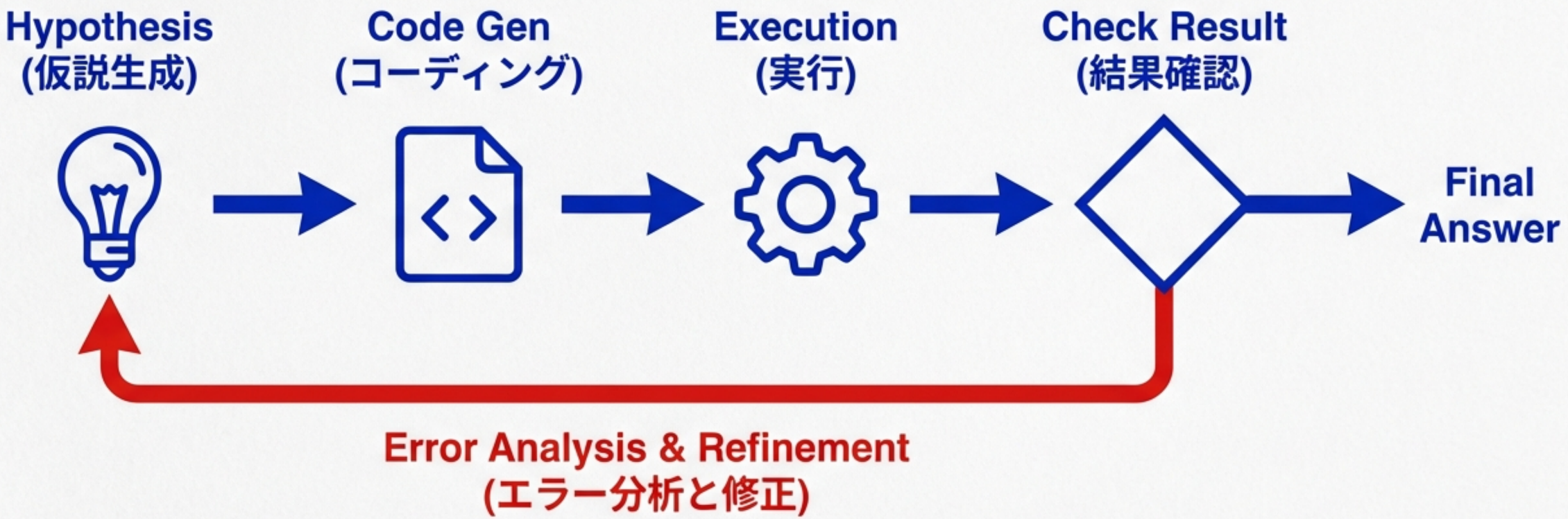
単一の「スーパーモデル」ではなく、「多モデル反映的推論 (Multi-Model Reflective Reasoning)」が壁を突破した。

アーキテクチャ：AIによる「組織的」問題解決



天才的な「脳」がすべてを行うのではなく、適材適所のスペシャリスト（視覚、論理、実装）にタスクを配分するアーキテクチャ。

プロセス：試行錯誤のループ (Reflective Reasoning)



「試行 → 失敗 → 分析 → 再試行」の科学的プロセスを自動化。
1問あたり最大10万回のコード実行を行う。

「考える時間」の拡張：秒から時間へのシフト

Standard Inference (通常推論)

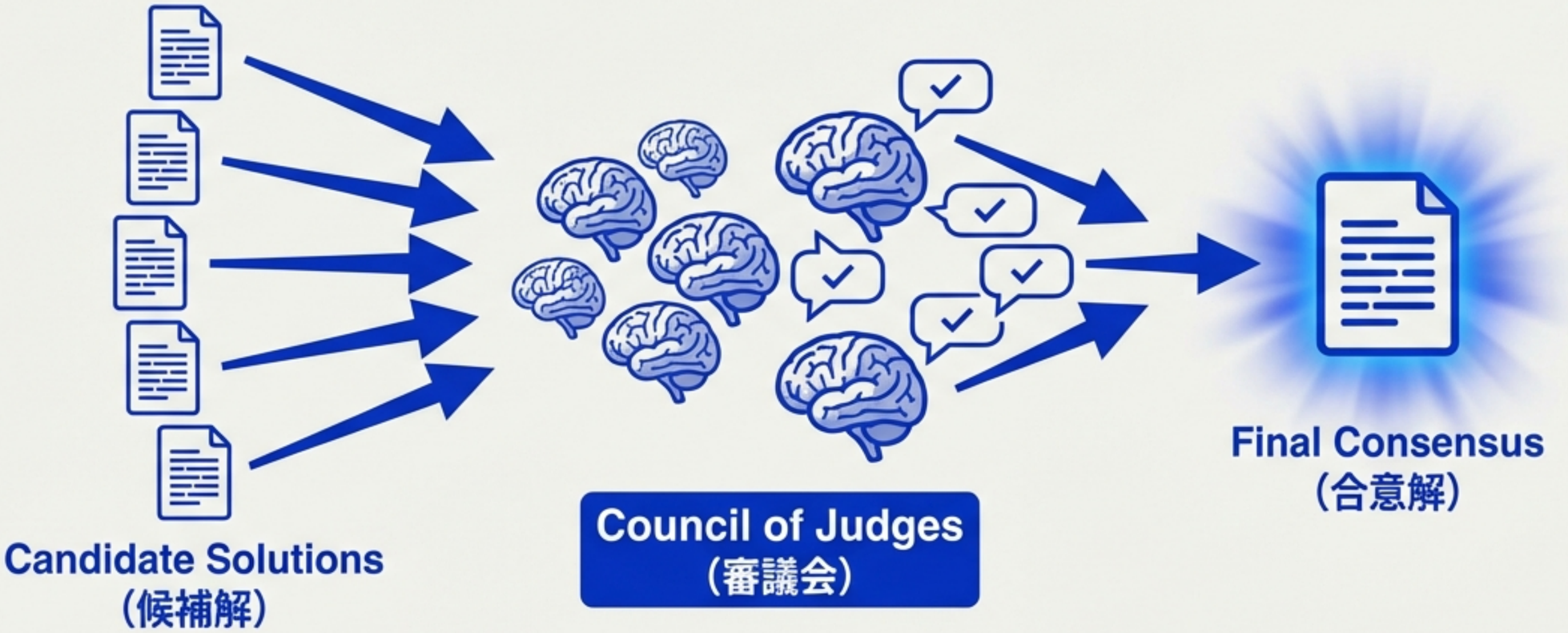
~10-30 Seconds

Land’s Multi-Model System

~6 Hours / Task

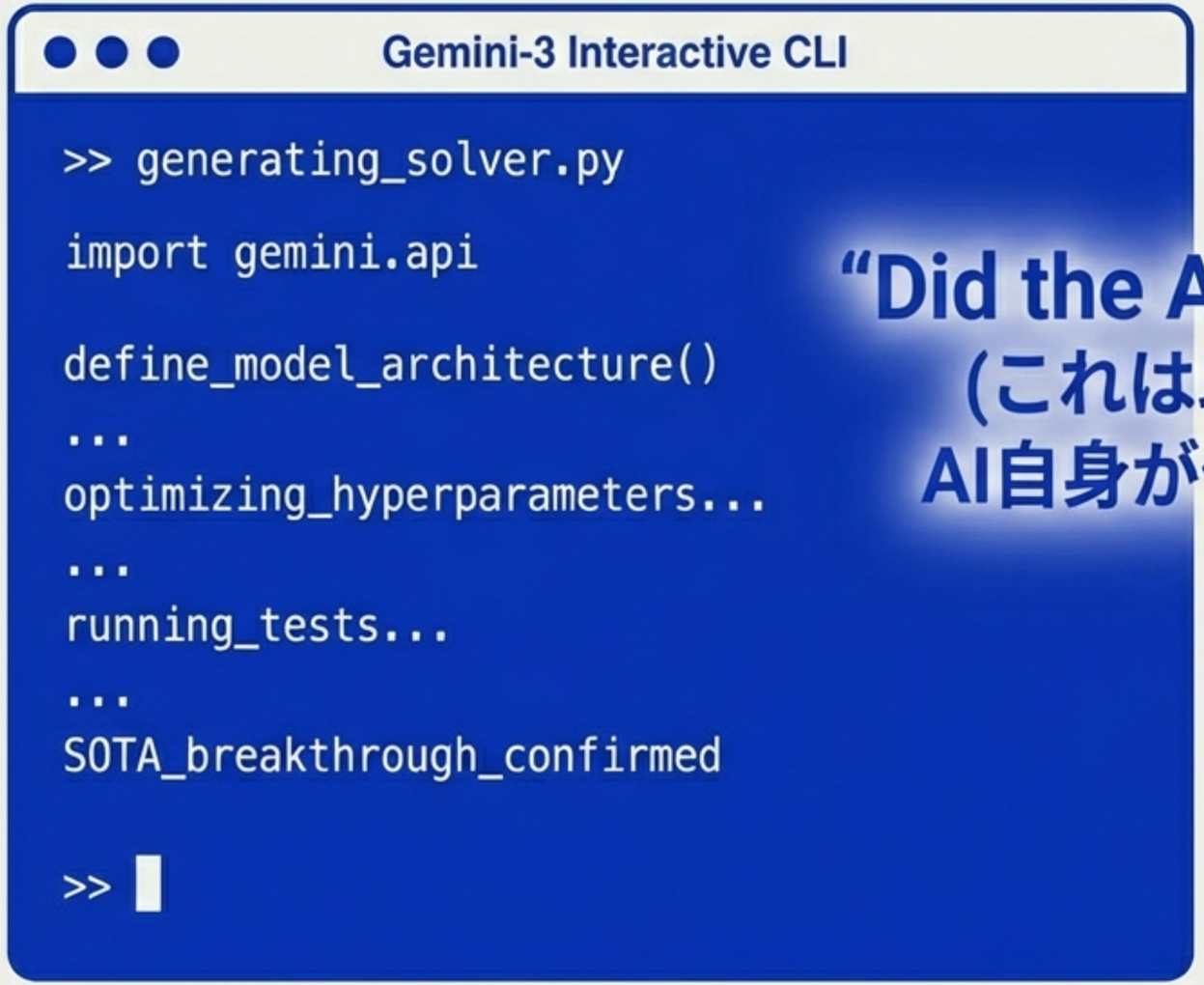
計算リソース（時間）を正確さに変換する戦略。
「疑似的な忍耐力 (Pseudo-Patience) 」により、
インスタントな推論では到達できない解探索を行う。

審議会（Council of Judges）：AIによるAIの相互評価



幻覚（**Hallucination**）を防ぐため、生成された複数の解法を別のモデル群が審査・投票。
多数決と相互検証により、まぐれ当たりではない信頼性の高い答えを選出する。

メタ発見：AIが作り出した「AIソルバー」



“Did the AI build its own successor?”
(これは以前のSOTAを打ち破るAIを、AI自身が作り出したと言えるだろうか？)

Johan Land氏のシステムのコード自体は、
人間が手書きしたものではなく、
Gemini-3との対話を通じて生成された。
人間はアーキテクト（設計者）となり、
実装はAIが行った。

知能の代償：39ドルの意味

$39.00

Cost per Task (ARC-AGI-2)

Inefficiency Gap
(非効率ギャップ)

Human Labor: ~$17.00

ARC-AGI-1 Solver: ~$11.00

現在の方法は「力技（Brute Force）」に近い。
正確さは達成したが、コストと効率は後退した。
実用化には、このコストを高効性能を劇的に下げる必要がある。

未来への展望：力技から「エレガントな推論」へ

- 60%の壁は、モデルの協調（オーケストレーション）によって突破された。
- 次の課題は、6時間の思考時間を短縮し、汎用的な「自己反省アルゴリズム」を確立すること。
- ARC-AGI-3への道は、計算量による解決から、より洗練された知能への進化にある。