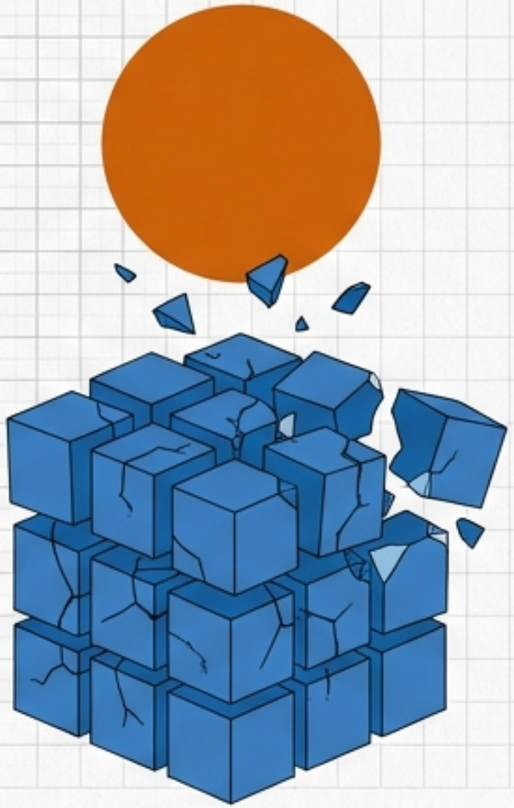
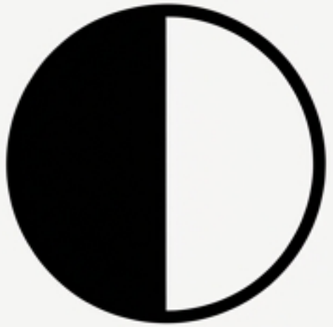
96%の断絶：なぜ中国製LLMは流動性知能で躓くのか

DeepSeek R1の「計算力」とARC-AGI-2の「推論力」——その構造的限界と次世代AIへの示唆



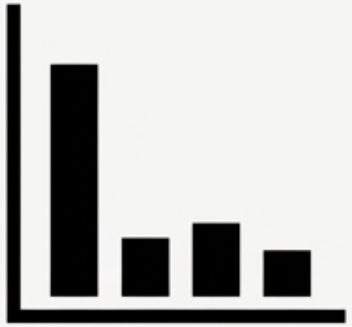
REPORT DATE: 2026年2月22日
SOURCE: ARC-AGI-2 Analysis Report v2

エグゼクティブ・サマリー



数学の天才、推論の幼児

DeepSeek R1はMATH-500で97.3%を記録する一方、ARC-AGI-2では1.3%に留まる。この乖離は、「知識の想起」と「真の推論」が別物であることを証明している。



西側の独走と中国の停滞

GoogleのGemini 3 Deep Think（84.6%）はベンチマーク解決目前。対してKimiやQwenなどの中国製モデルは70ポイント以上の大差をつけられている。

パラダイムの行き止まり

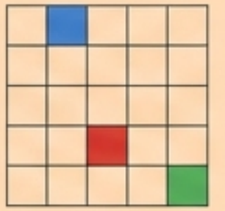
検証器なき強化学習（RL）と計算リソースの投入だけでは、流動性知能の壁を越えられない。「次のトークン予測」の限界が露呈した。

Forensic Editorial

97.3% vs 1.3%：パラドックスの正体

97.3%

MATH-500
結晶性知能 / 知識の想起



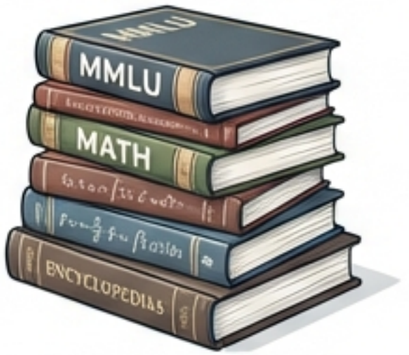
1.3%

ARC-AGI-2
流動性知能 / 未知への適応

現行のLLM評価指標は、暗記能力と知能を混同している。

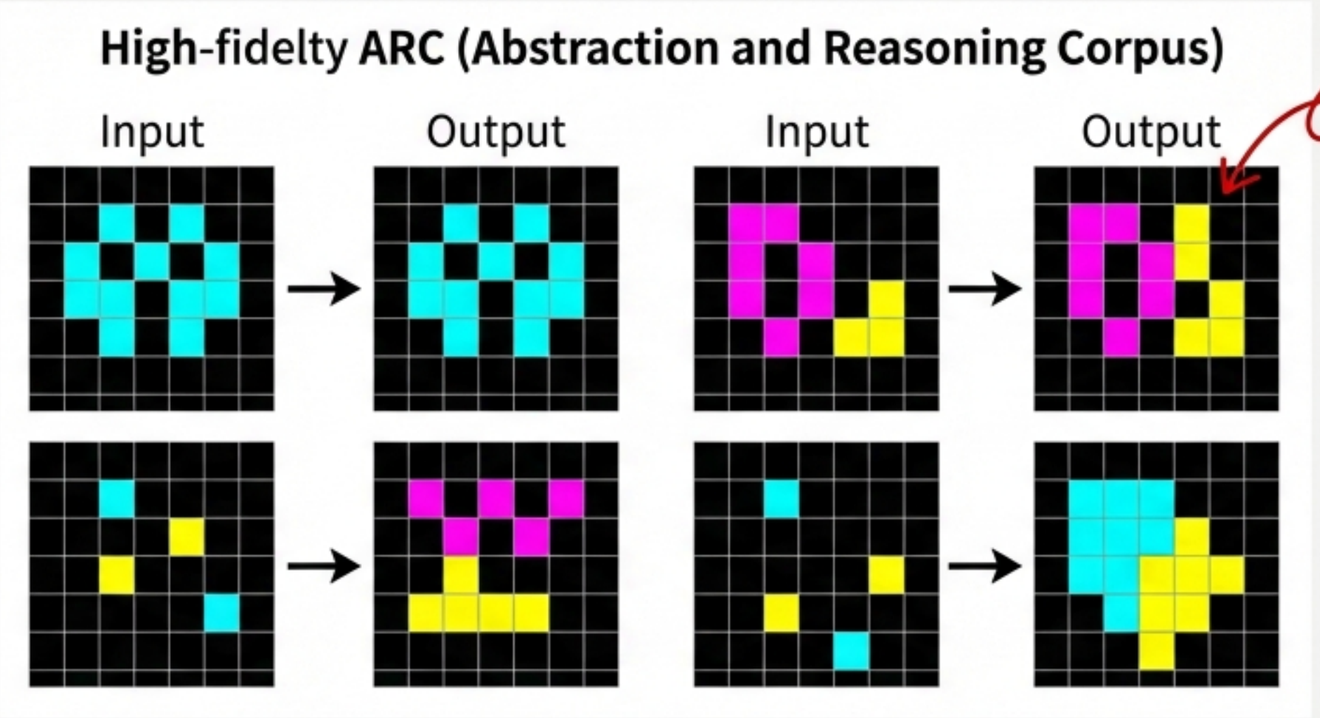
新たな物差し「ARC-AGI-2」とは何か

開発者: François Chollet & ARC Prize Foundation（2025年3月公開）
目的: 「流動性知能（Gf）」の測定 —— 訓練データにない未知の問題を解く能力。

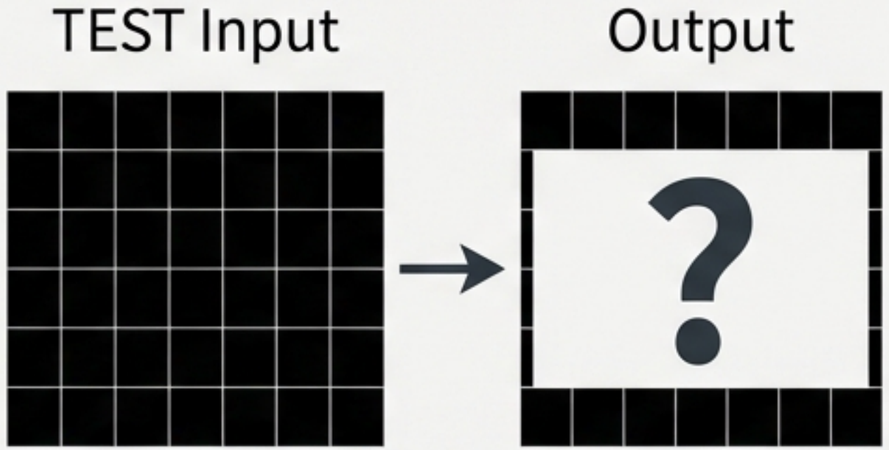
	
従来の指標（MMLU, MATH）	**ARC-AGI-2**
• 特定のスキルと暗記パターンを測定 • データ汚染（Data Contamination）の影響大 • 結晶性知能	• 適応力とルールの発見を測定 • 言語・文化的背景を排除 • 流動性知能

基準値: 人間 = 100%（平均的な人間でも約60%を獲得、認知プリミティブを測定）

言語を持たないIQテスト



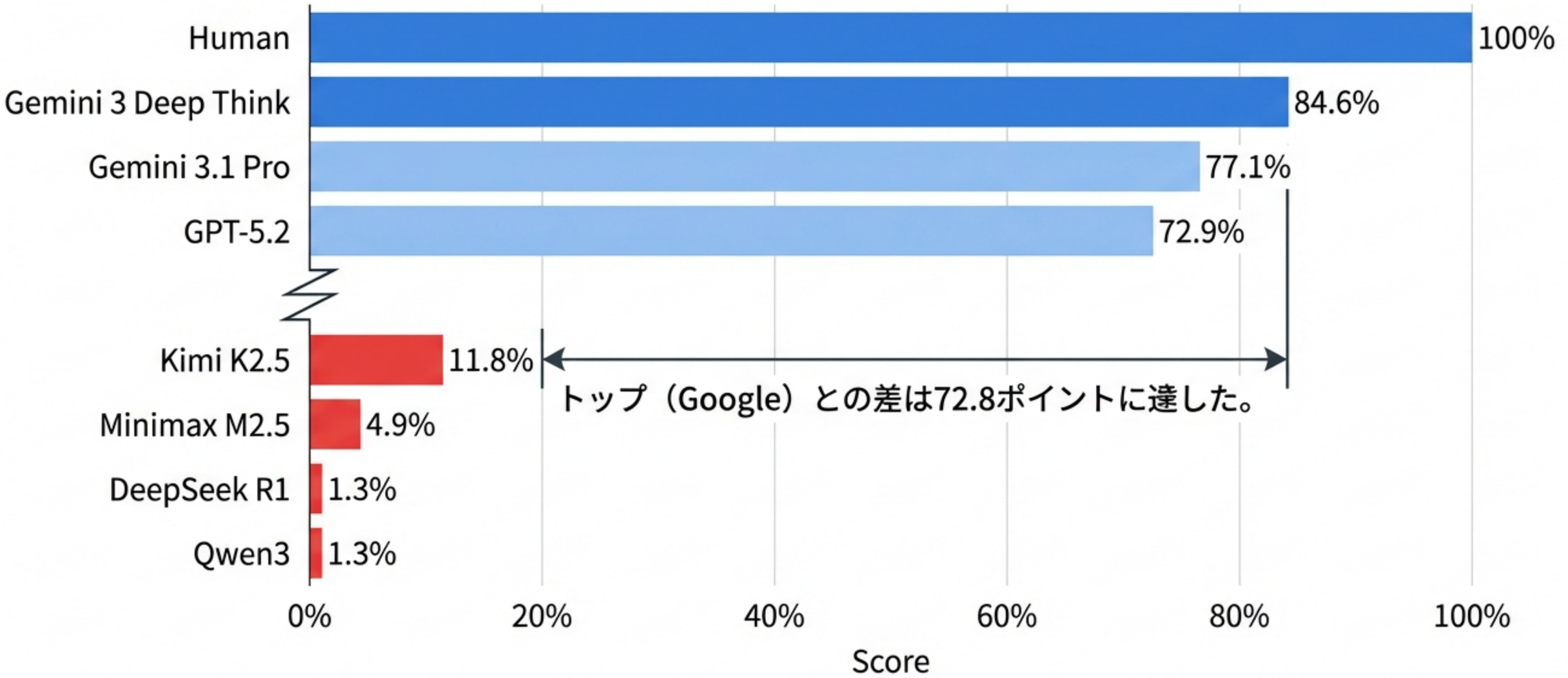
必要なのは知識ではなく「プリミティブ」



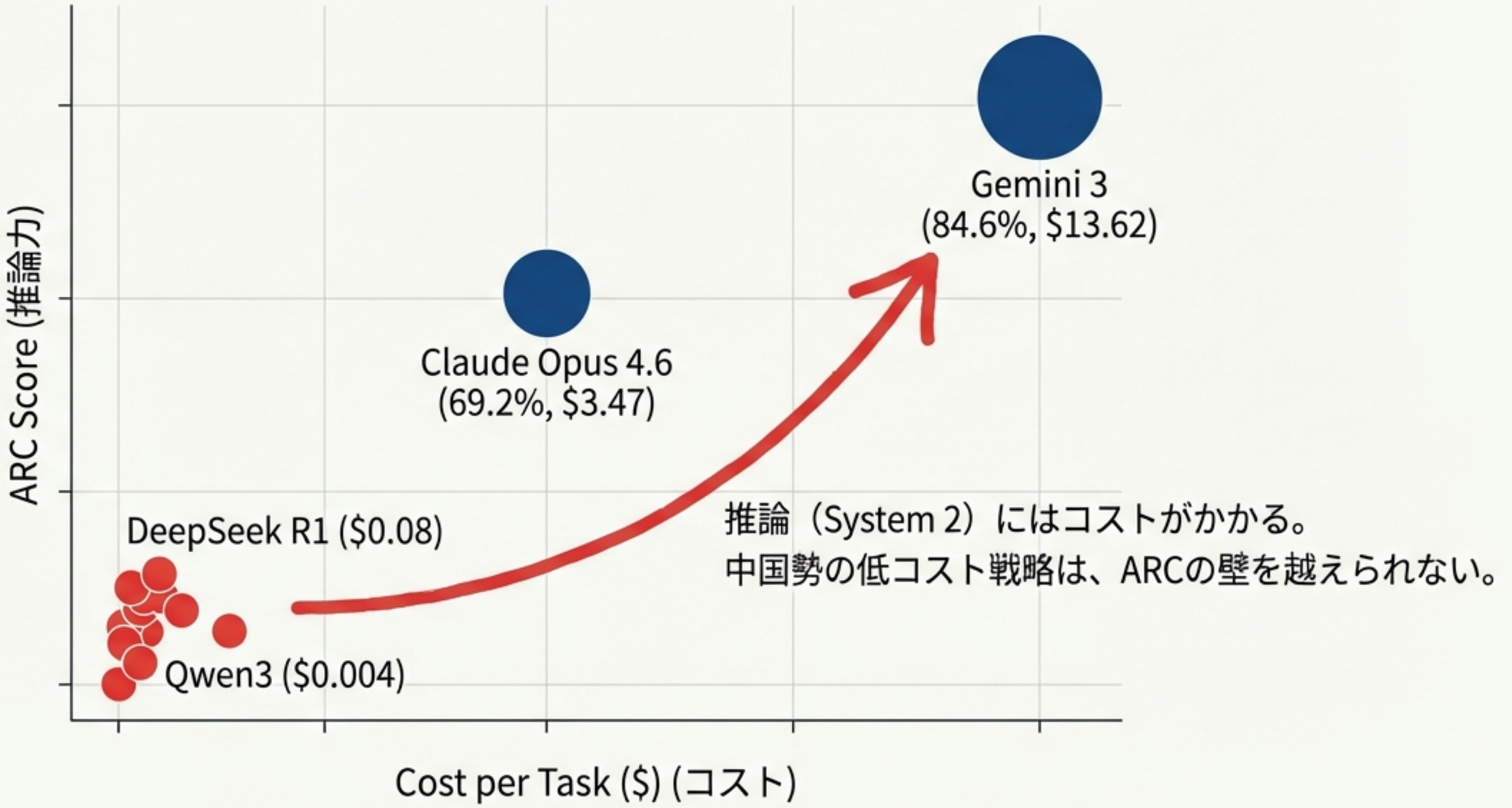
対象の永続性、幾何学、対称性、因果関係

制約：Pass@2（回答権は2回のみ。総当たり攻撃不可）

グローバル・リーダーボード：拡大する格差



知能のコスト：安価な計算力は推論を代替しない



透明性の欠如と「沈黙」

Western Labs

BENCHMARK	SCORE
ARC-AGI-2	84.6%

Chinese Labs

Qwen Case Study

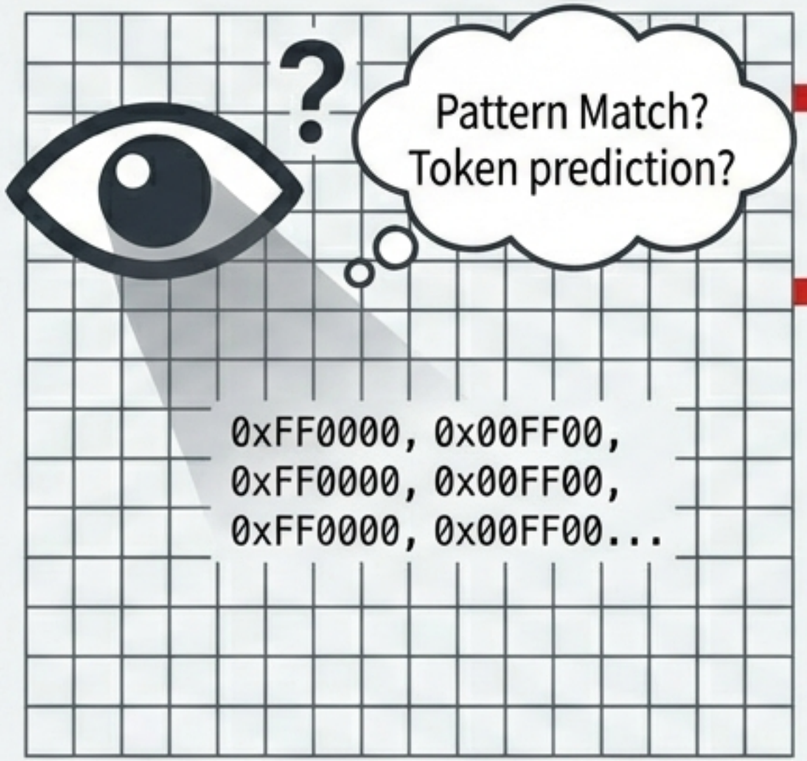
Self-Reported: 41.8%

Independent Verificati

自己申告と検証結果に4倍の乖離。ベンチマークの信頼性が問われている。

失敗の解剖①：記号解釈の盲点

Symbolic Blindness

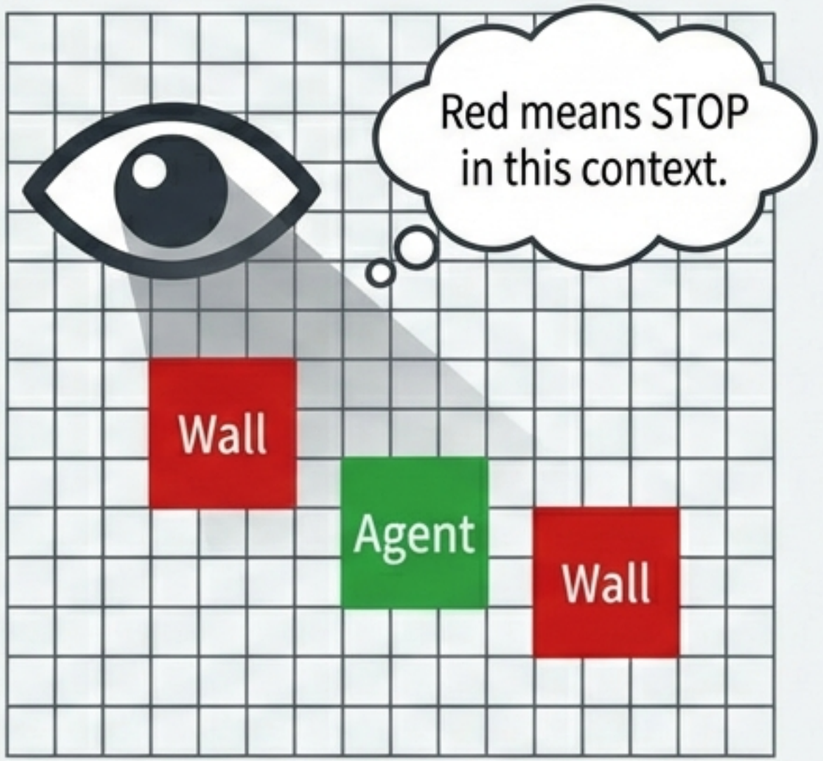


AIの視点
(The LLM View)

文脈依存的ルール適用
(Context-Dependent Rule Application)

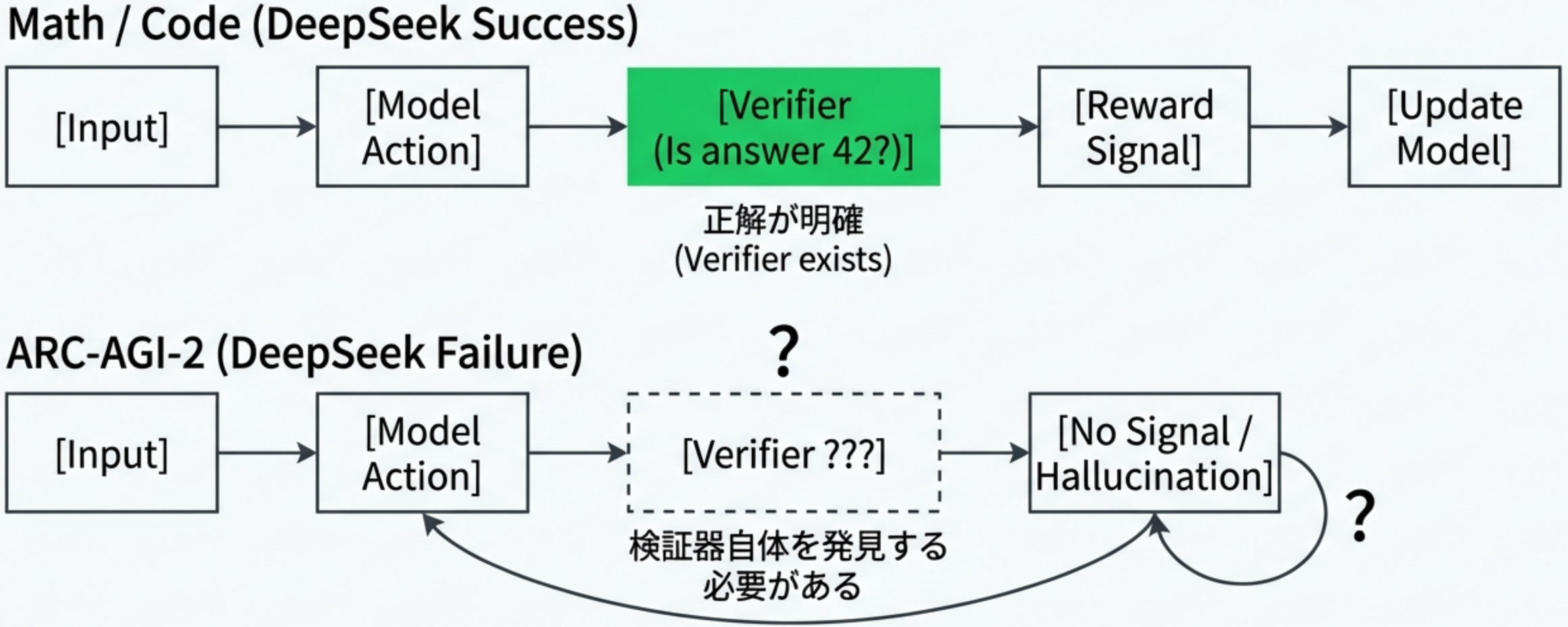
LLMは記号を表層的なパターンとして処理するが、文脈によって変化する『意味』を理解できない。

真の理解 (Contextual Understanding)



真の理解
(Contextual Understanding)

失敗の解剖②：強化学習（RL）の罠



「純粋なRLの成功は、信頼性の高い報酬信号に依存する」(DeepSeek R1 Paper)

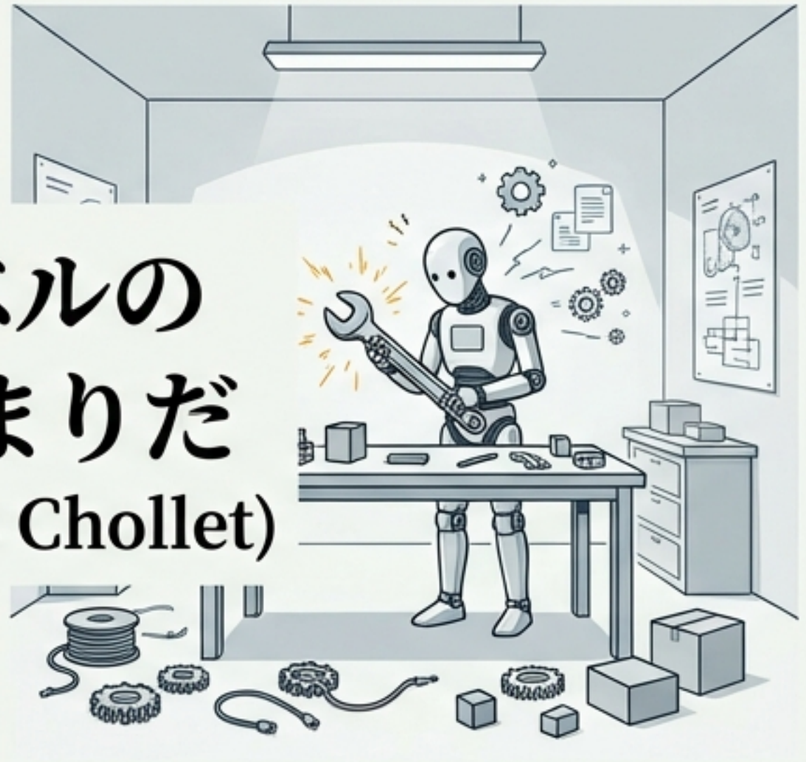
「スキル」と「知能」は別物である

LLM (DeepSeek/GPT-4)

ベクトルプログラムのリポジトリ

既存の解決策を検索・想起する (Math/MMLU)

AGI (ARC Ideal)



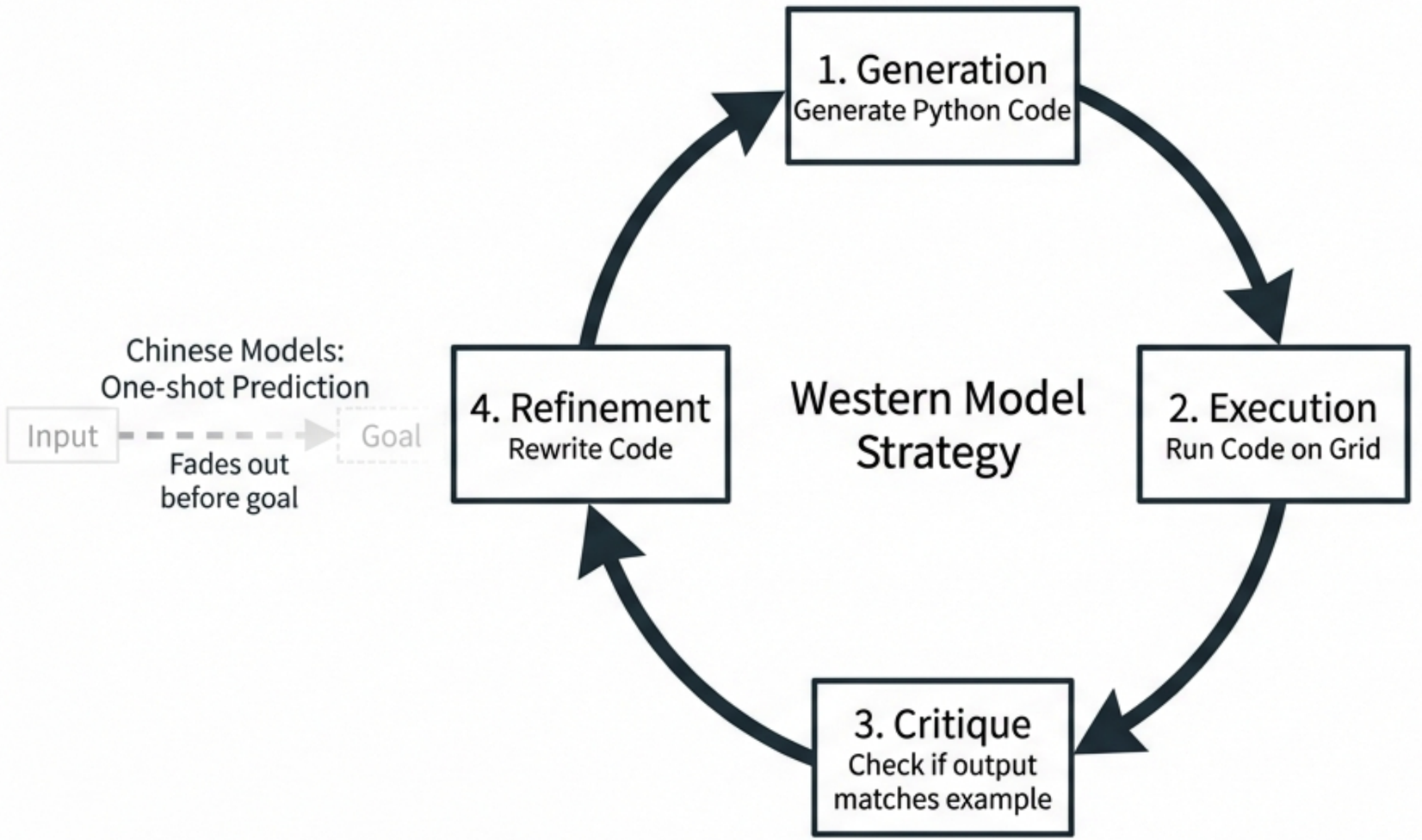
プログラム合成機

その場で新しい解決策を創り出す (ARC)

LLMは人間レベルの知能への行き止まりだ

(Yann LeCun / François Chollet)

壁を越えるアプローチ：推論時の学習

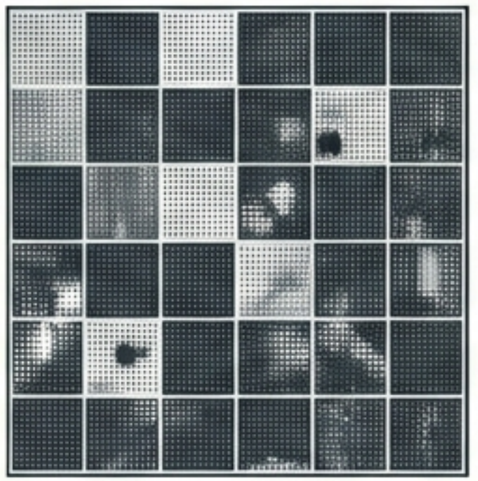


- Refinement Loops:
  自らの回答を批判・修正
  (GPT-5.2)
- System 2 Thinking:
  出力前の『熟考』プロセス
  (Gemini 3)
- Program Synthesis:
  画素予測ではなく、ロジック（コード）を生成

NVIDIA NVARC Team: カスタムアーキテクチャで$0.20という低コストで27.6%を達成。

次なるフロンティア：ARC-AGI-3へ

2025: ARC-AGI-2



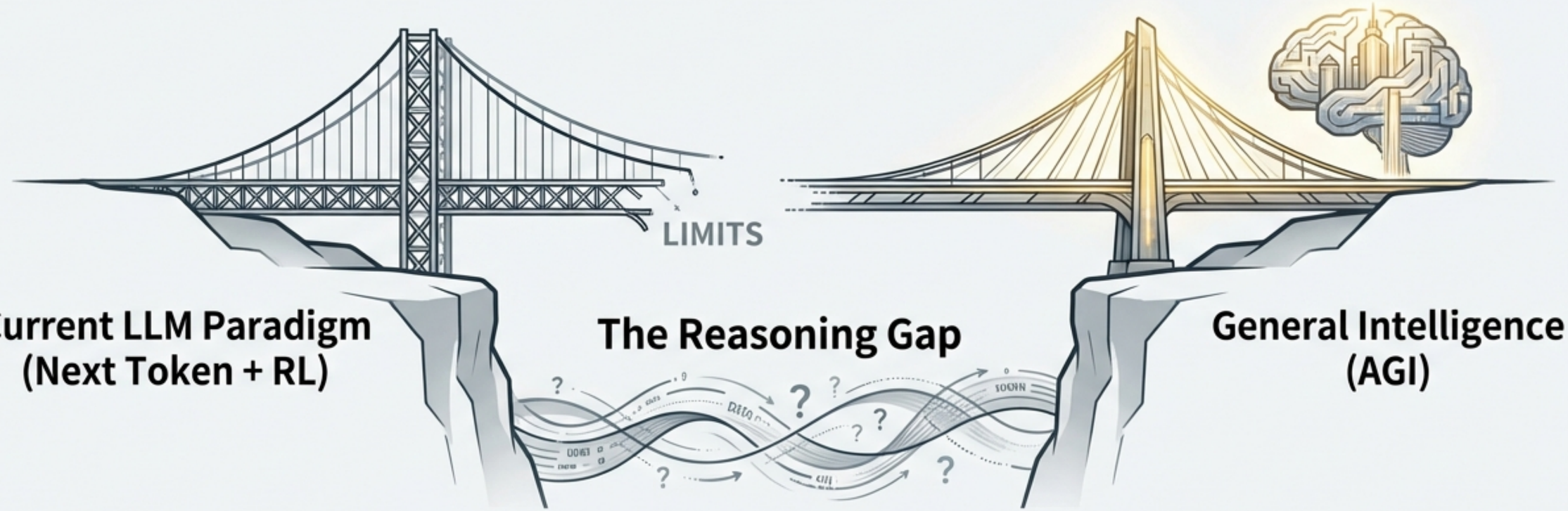
静的推論

2026: ARC-AGI-3

インタラクティブ推論

静的なグリッドから、実験とフィードバックが必要な動的環境へ。
ARC-AGI-2（静的）を解けないモデルに、ARC-AGI-3（動的）を解くチャンスはない。

結論：スケーリング則の向こう側



Paradigm Dead End

単純なトークン予測とRLは限界に達した。

The China Challenge

最適化（Optimization）はツールを作るが、知能（Mind）は作らない。

The Path Forward

想起ではなく、動的に推論するアーキテクチャへの転換が必須。