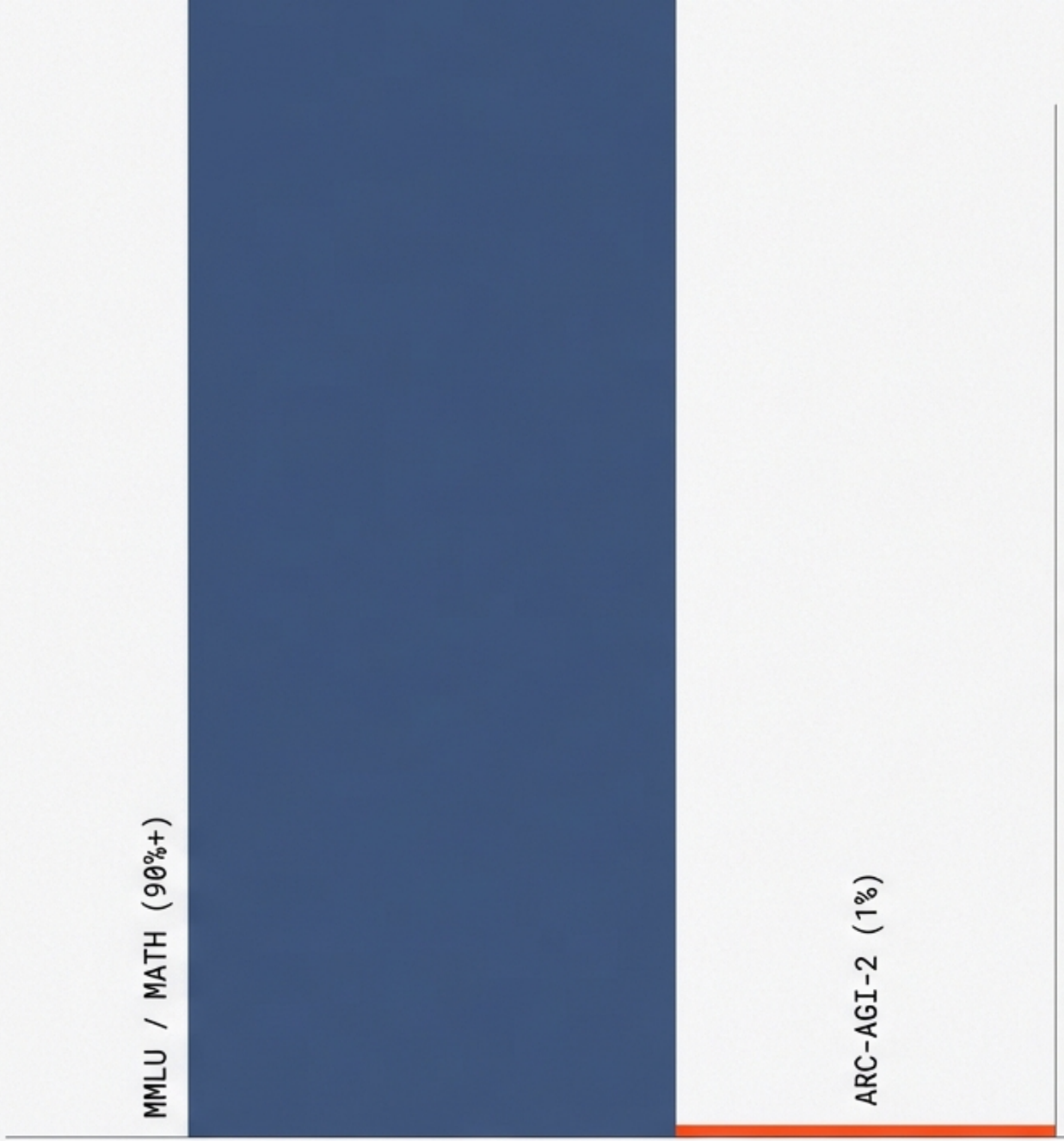


1%の特異点：中国製LLMとARC-AGI-2の壁

MMLUを制したDeepSeekとQwenが、なぜ「真の推論」で沈黙するのか。

Noto Serif JP Regular

Based on Manus AI Report - Feb 2026

2026年、知識集約型タスクの覇権

DeepSeekやQwenに代表される中国製LLMは、数学やコーディングなどの「スキルベース」の領域で驚異的な性能を証明した。

MMLUやGSM8Kにおいて、これらのモデルは人間を超えるスコアを常態化させている。

Key Insight: World Class Performance in Verifiable Tasks.

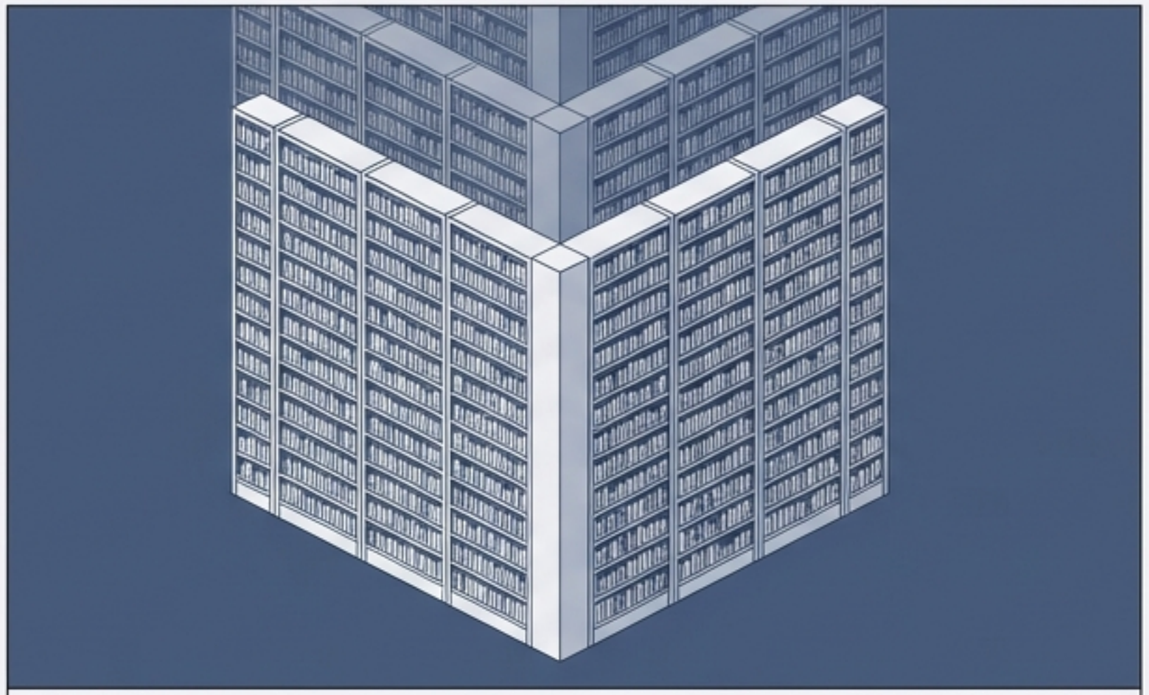
Model	MMLU	MATH	Rank
DeepSeek-V3	88.5 ↑	90.2 ↑	#1
Qwen-2.5-Math	87.1 ↑	89.8 ↑	#2
Claude 3.5 Sonnet	86.8	88.0	#3

80%と1%の断絶：ベンチマーク間のパラドックス

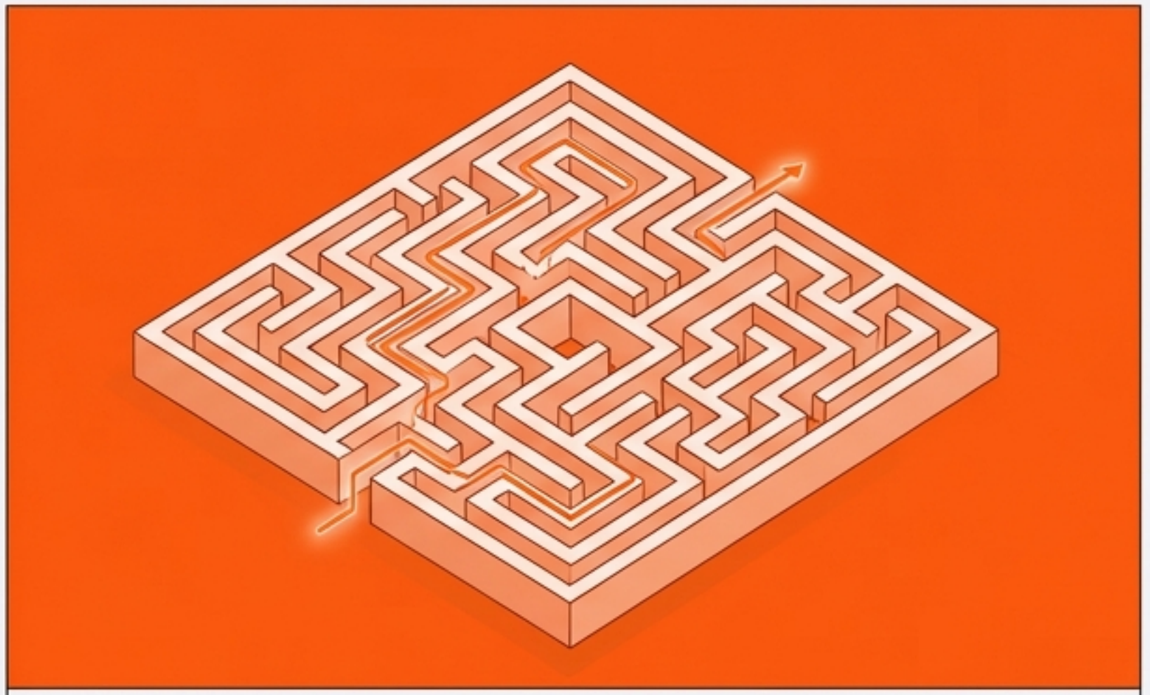
既存のテストで頂点に立つモデルが、ARC-AGI-2ではほぼ0%台（1%前半）に転落する。

これは計算資源の不足ではない。「知能の定義」における構造的な乖離である。

記憶か、発見か。



結晶性知能 (Crystallized)

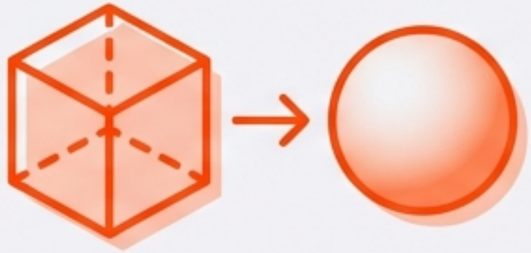


流動性知能 (Fluid)

このスコアの乖離は、AIが何を「学習」したかの違いを浮き彫りにする。

- **結晶性知能 (Crystallized):** 既存の知識を記憶し、適用する力（MMLU）。
- **流動性知能** (Fluid): 未知の課題に対し、その場でルールを発見する力（ARC-AGI-2）。

ARC-AGI-2：流動性知能を測る試金石



未知への適応

事前知識に依存せず、少数の例（Few-shot）から法則を導き出す。

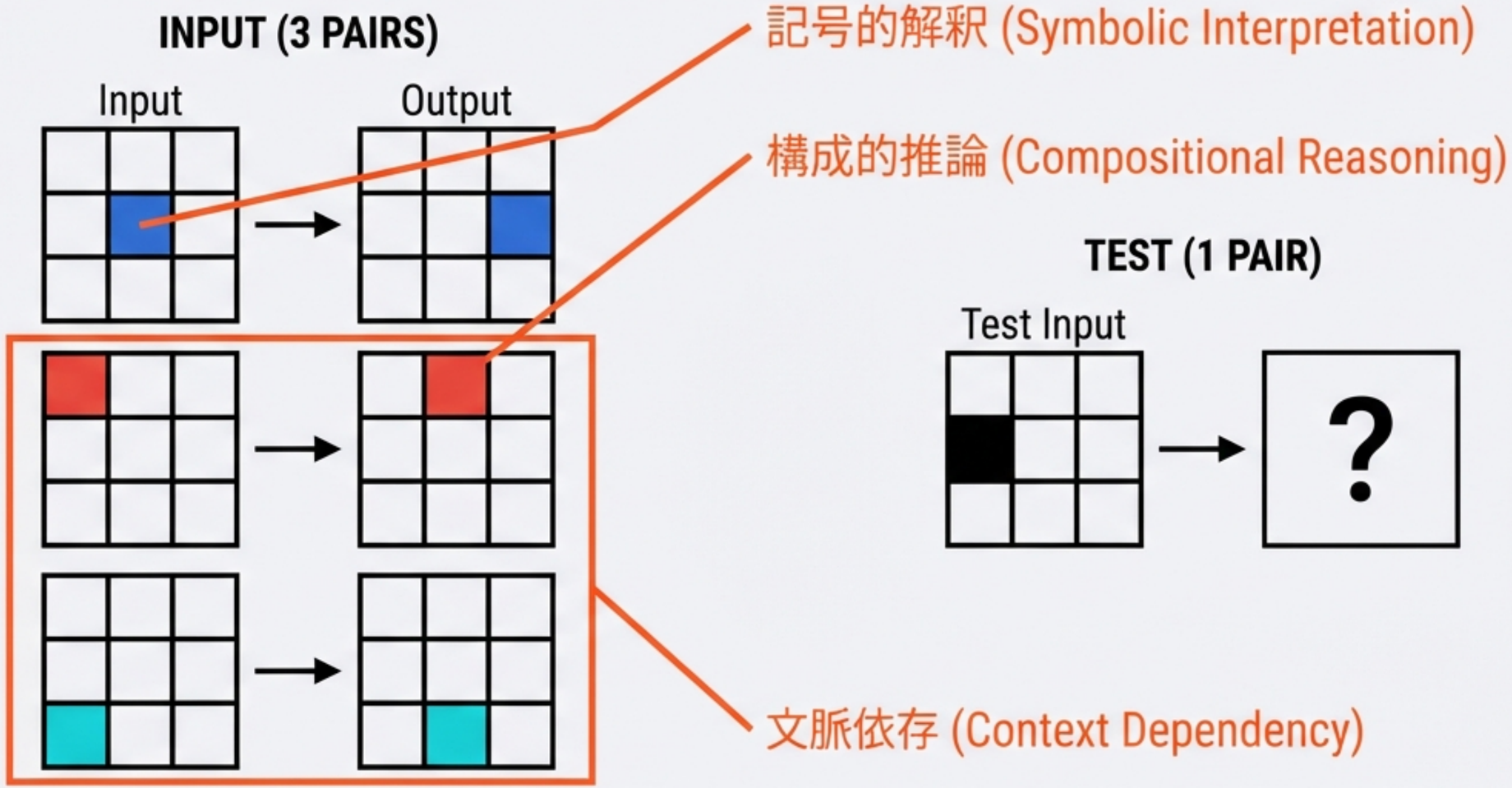
非言語的推論

テキストの統計的確率（次に来る単語）ではなく、記号操作と論理構築を要求する。

新規性

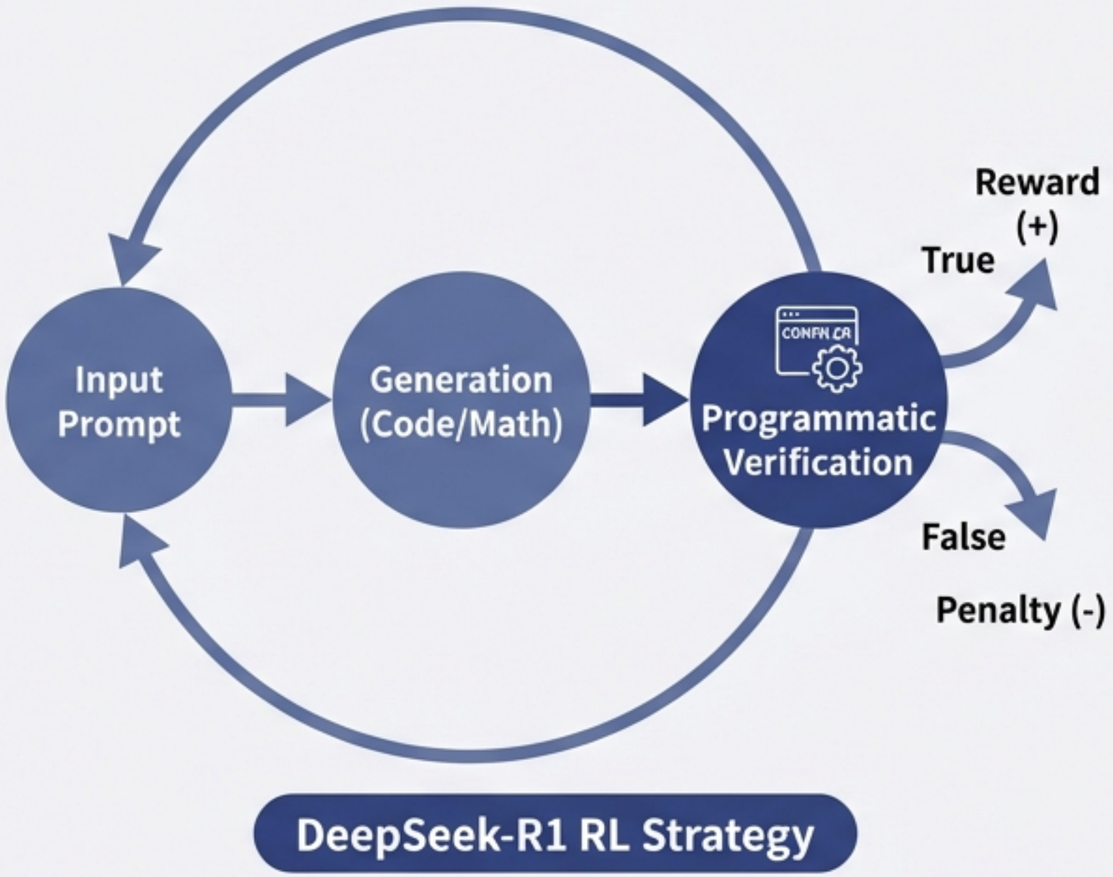
全ての問題が新規作成され、過去のデータの記憶では解けない。

統計的相関では解けないパズル



Takeaway: The model cannot recite the answer; it must construct the logic.

「検証可能なタスク」への過剰適応

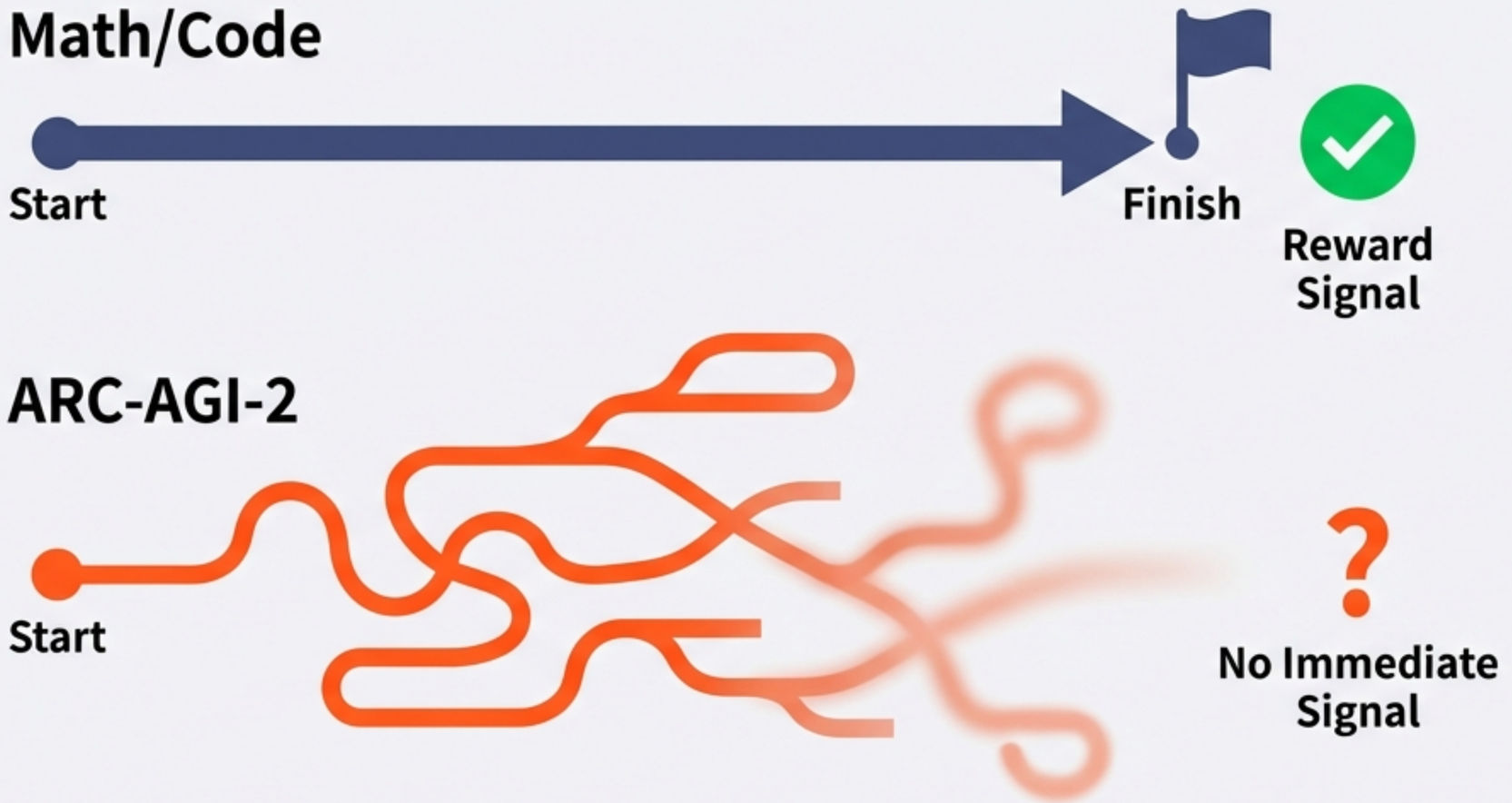


中国製LLMの強みは、DeepSeek-R1論文にあるような「正解が明確な領域」での強化学習にある。

数学やコード生成では、答えが合っているかプログラム的に判定できるため、効率的に学習が進む。

結果：正解ルートが確立されたタスクには無類の強さを発揮する。

報酬なき荒野での迷走

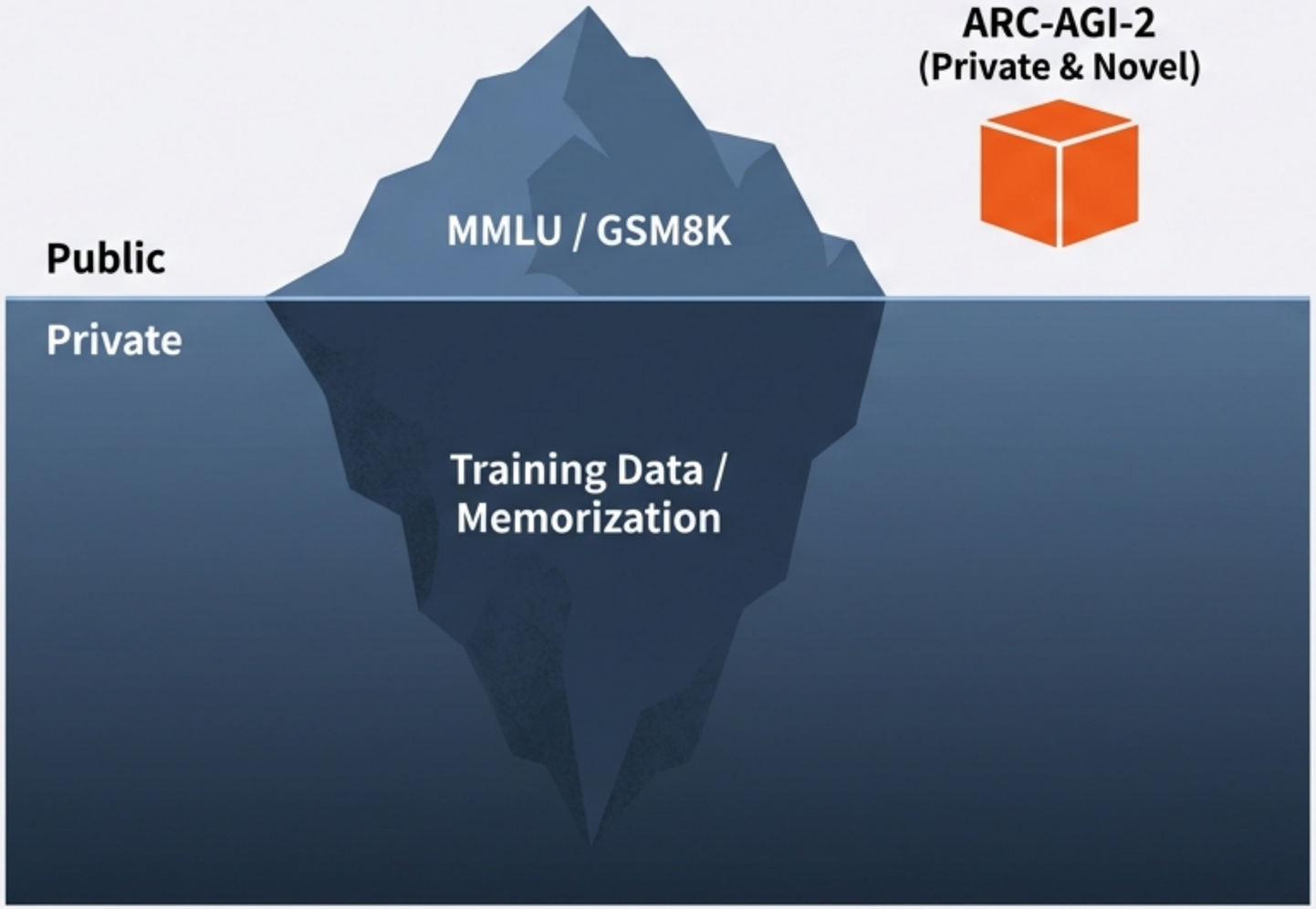


ARC-AGI-2のような「開放的な推論タスク」には、即時の報酬信号が存在しない。

検証可能なタスクに特化したモデルは、報酬がない環境での「思考の連鎖（Chain-of-Thought）」の探索方法を知らない。

ルールそのものを発見しなければならない状況下で、モデルは立ち尽くしてしまう。

データ汚染という幻影

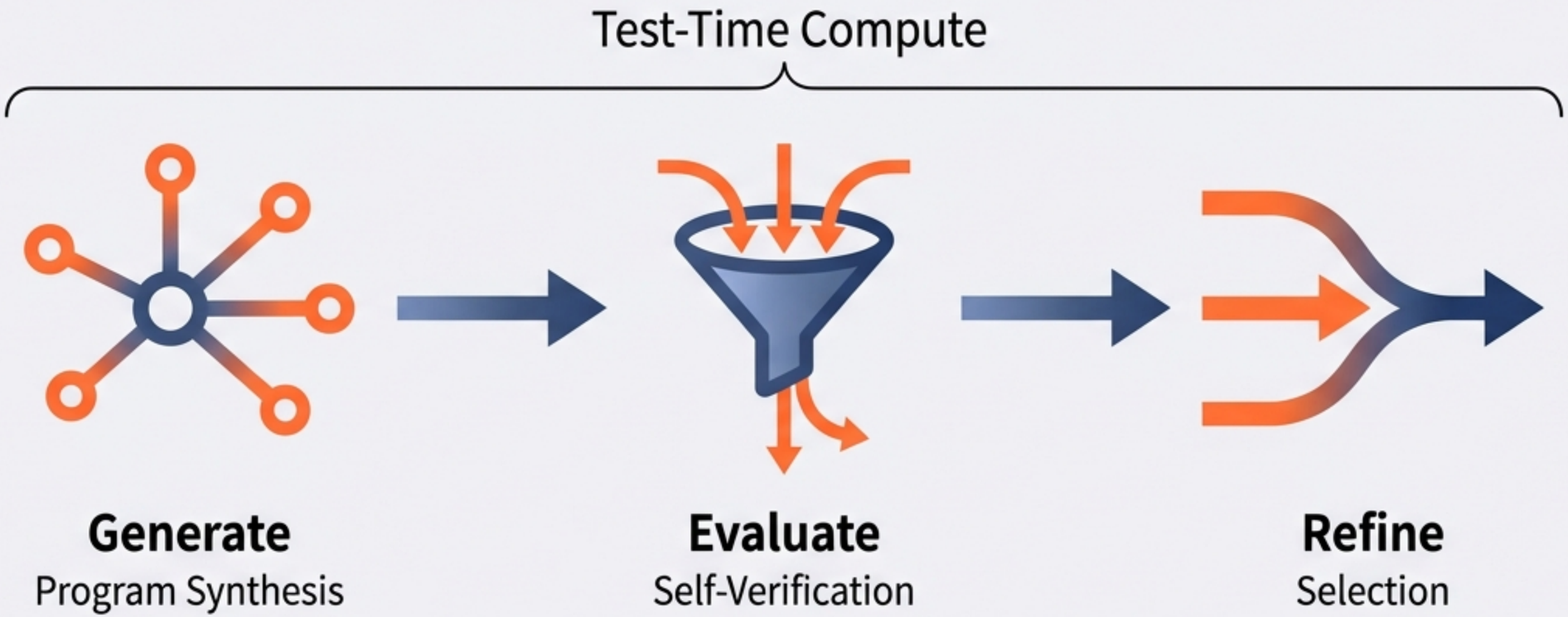


- 既存ベンチマークの高得点は、テスト問題が訓練データに含まれてしまっている「記憶」の結果である可能性が否定できない。
- ARC-AGI-2は非公開（Private）かつ新規（Novel）であるため、この「カンニング」が通用しない。
- 1%というスコアこそが、現在のLLMの「素の推論能力」を冷徹に示している。

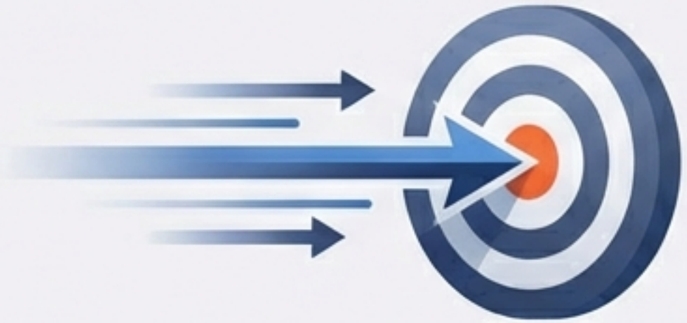
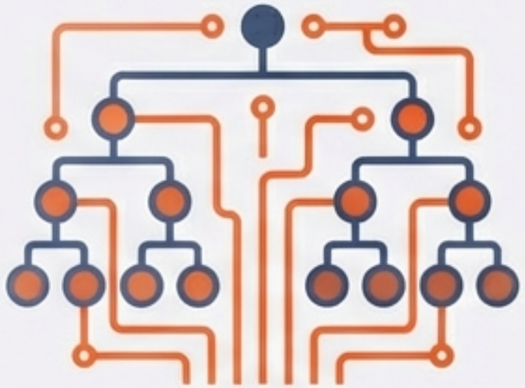
84.6%の解：Gemini 3 Deep Thinkのアプローチ

ARCを攻略したのは「巨大なモデル」ではなく「洗練されたシステム」である。

- **テスト時計算 (Test-Time Compute)**：回答前に時間をかけ、複数の推論経路を探索・検証する。
- **プログラム合成**：答えを直接出すのではなく、答えを導く「プログラム」を生成する。



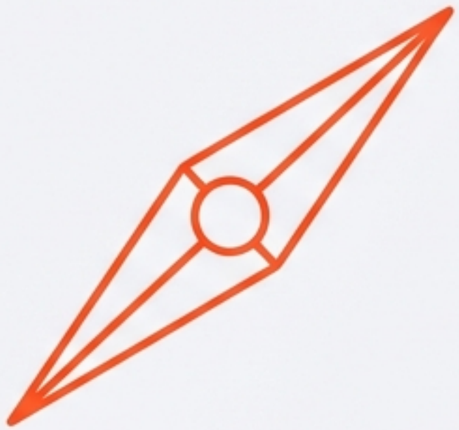
直感（System 1）から、熟慮（System 2）へ

System 1 (DeepSeek)	System 2 (Gemini 3)
	
Fast Intuitive Pattern Matching	Slow Deliberate Symbolic Manipulation

中国製LLM: 「次に来る単語」を予測する直感的アプローチ。既知のパターンの適用に強い。

次世代システム: 複数の思考プロセスを組み合わせ、動的に解を探索する。これがARC攻略の鍵となった。

AGIへの道標



- 中国製LLMの苦戦は、モデルの劣等性ではなく、AI技術全体が直面する「統計から推論へ」の過渡期を示している。
- 「スコアの最適化」を超え、真の汎用知能へ到達するためには、未知のルールを解き明かす構成的推論能力が不可欠である。
- ARC-AGI-2は、その進化を測るための唯一無二の羅針盤であり続ける。