2026年におけるAI推論の壁

中国製LLMとARC-AGI-2

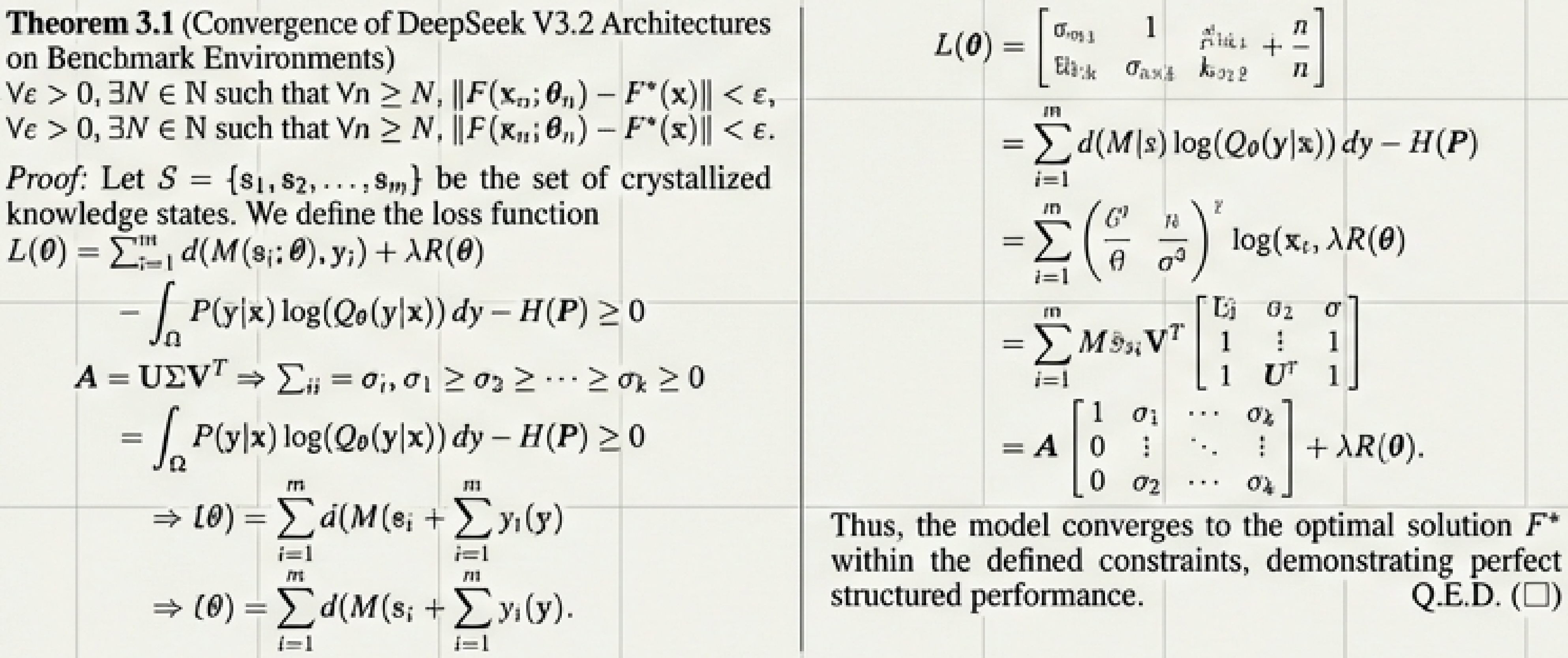


Figure 1.1: Crystallized Intelligence - Solved, Structured, and Perfect Execution.

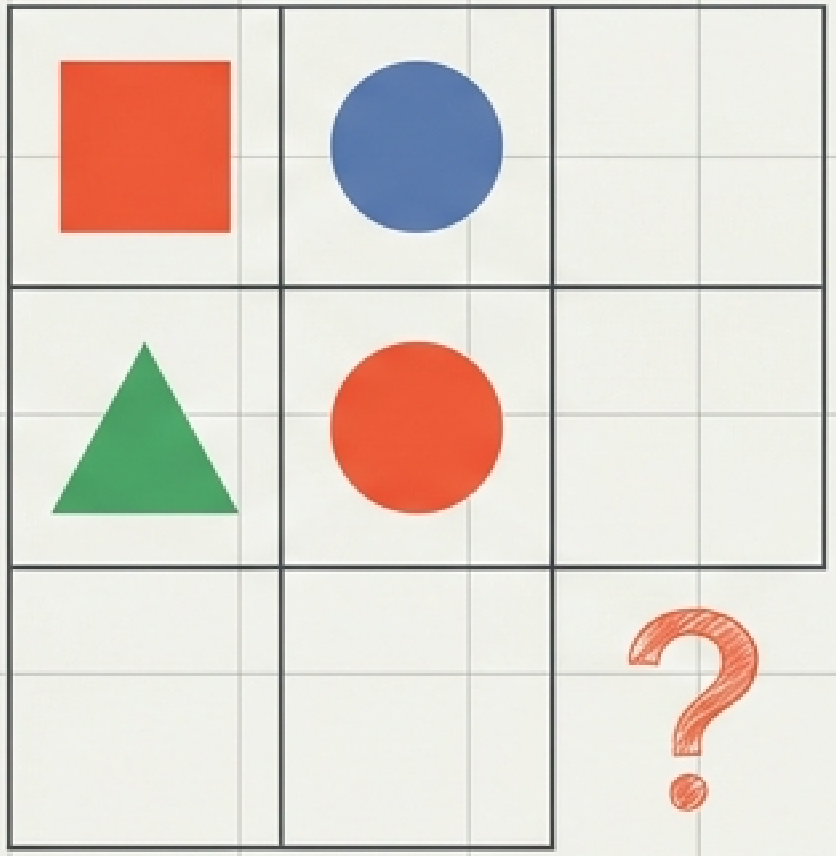


Figure 1.2: Fluid Intelligence/ARC - Unsolved, Adaptable, and Abstract Reasoning Task.

結晶性知能の覇権と流動性知能の欠落に関する構造的・戦略的分析

本資料は、DeepSeek V3.2およびQwen 3.5のベンチマーク性能と、ARC-AGI-2におけるスコア飛躍の技術的・経済的要因を分析したものである。

Executive Summary（要約）

Situation（現状）

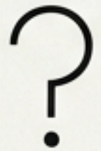
中国製LLMの台頭：2026年現在、DeepSeek V3.2やQwen 3.5は、数学（AIME）や知識（GPQA）において米国のフロンティアモデルと同等、あるいはそれを凌駕するスコアを記録している。

Complication（課題）

1%の衝撃：しかし、未知の抽象推論を測定する ARC-AGI-2 においては、人間（100%）や Gemini 3 Deep Think（84.6%）に対し、　　これらのモデルは1.3%という壊滅的な低スコアに留まっている。

Question（問い）

なぜ「数学の天才」であるAIが、幼児レベルのパズルを解けないのか？

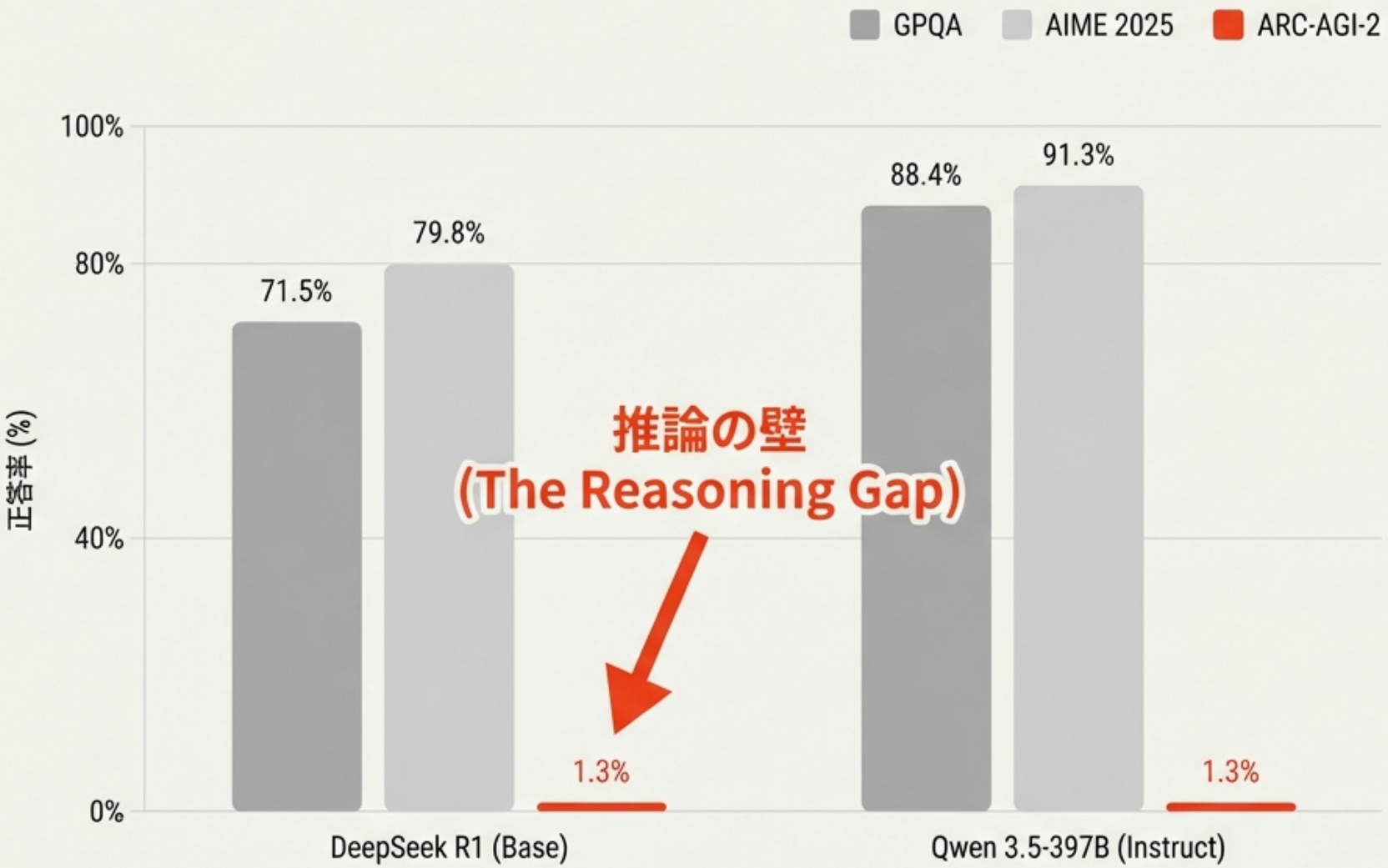


Answer（結論）

構造と戦略の壁：

- 1. アーキテクチャ：自己回帰モデルによる2次元空間情報の1次元化（シンボルグラウンディング問題）。
- 2. 戦略：中国モデルは「推論効率（System 1）」を極限まで最適化しており、ARCが要求する「探索的推論（System 2）」に必要な計算コストを排除する設計思想にある。

性能の乖離：ベンチマークスコアにおける「推論の壁」



従来のベンチマークスコアは、もはや「汎用な知能」の指標としては機能してていない。この極端な乖離こそが、現在のLLMが抱える認知能力の偏りを示している。

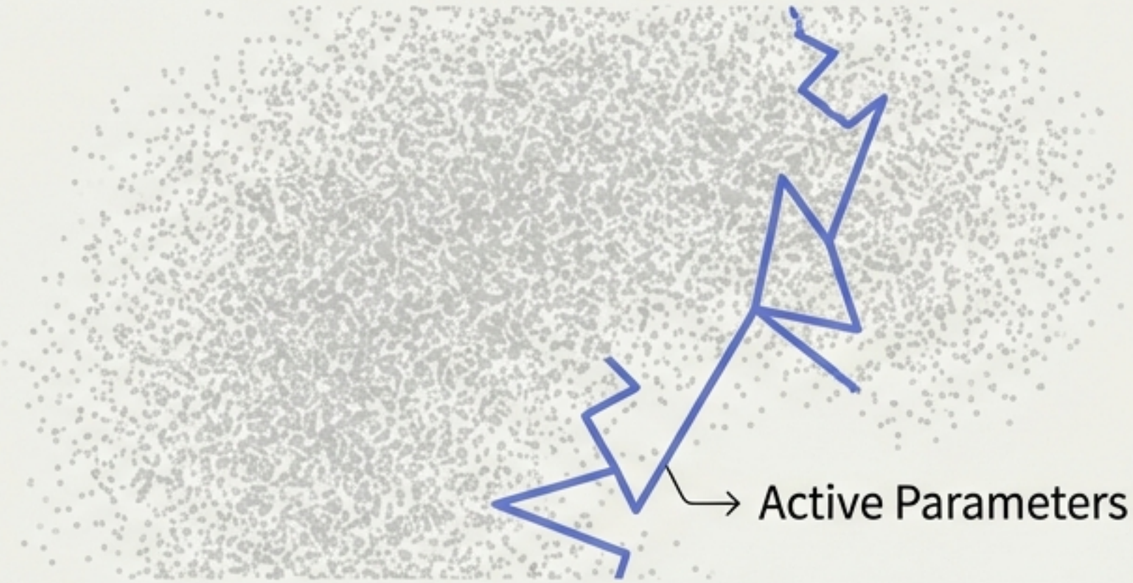
結晶性知能の極致：中国製LLMの真の強み

結晶性知能（Crystallized Intelligence）：過去に学習した知識やアルゴリズムを、適切な文脈で効率的に検索・適用する能力。

DeepSeek Sparse Attention (DSA) と MoE により、人類史上最も効率的な「想起」システムを実現。

Sparse Attention

MoE



DeepSeek V3.2	
AIME 2025: **96.0%**	HMMT: **99.2%**

Qwen 3.5	
GPQA Diamond: **88.4%**	LiveCodeBench: **83.6%**

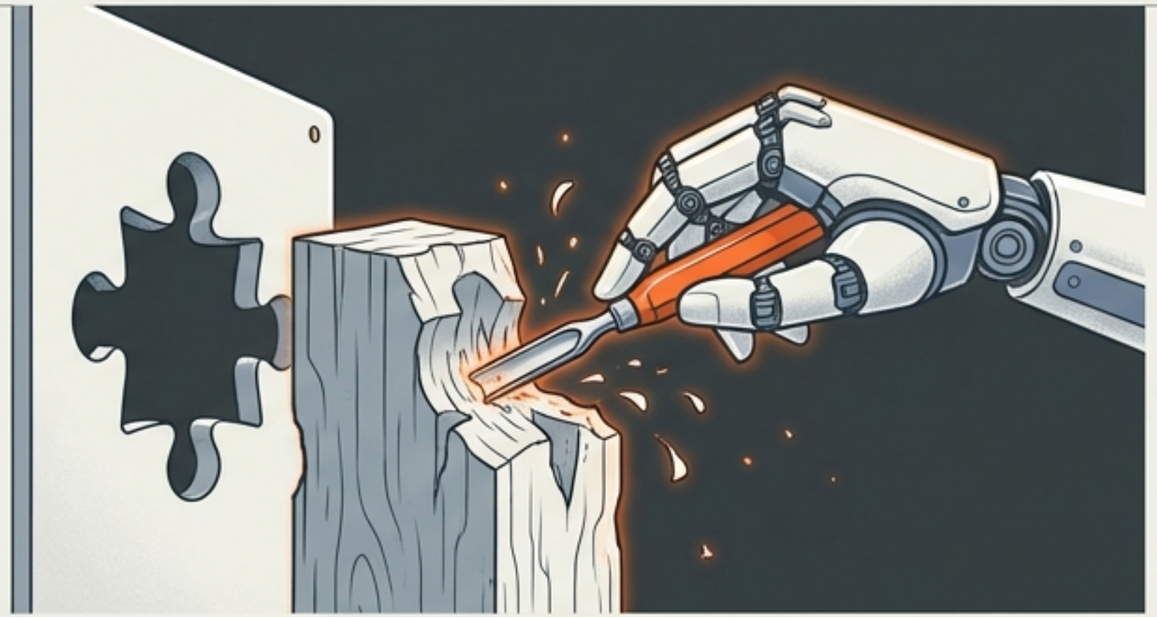
Swiss Scientific Editorial

推論の錯覚：なぜベンチマークはハッキングされるのか

Recall vs. Reasoning

意味的重複 (Semantic Duplicates) / 補間 (Interpolation)

未知への適応 (Adaptation) / 外挿 (Extrapolation)

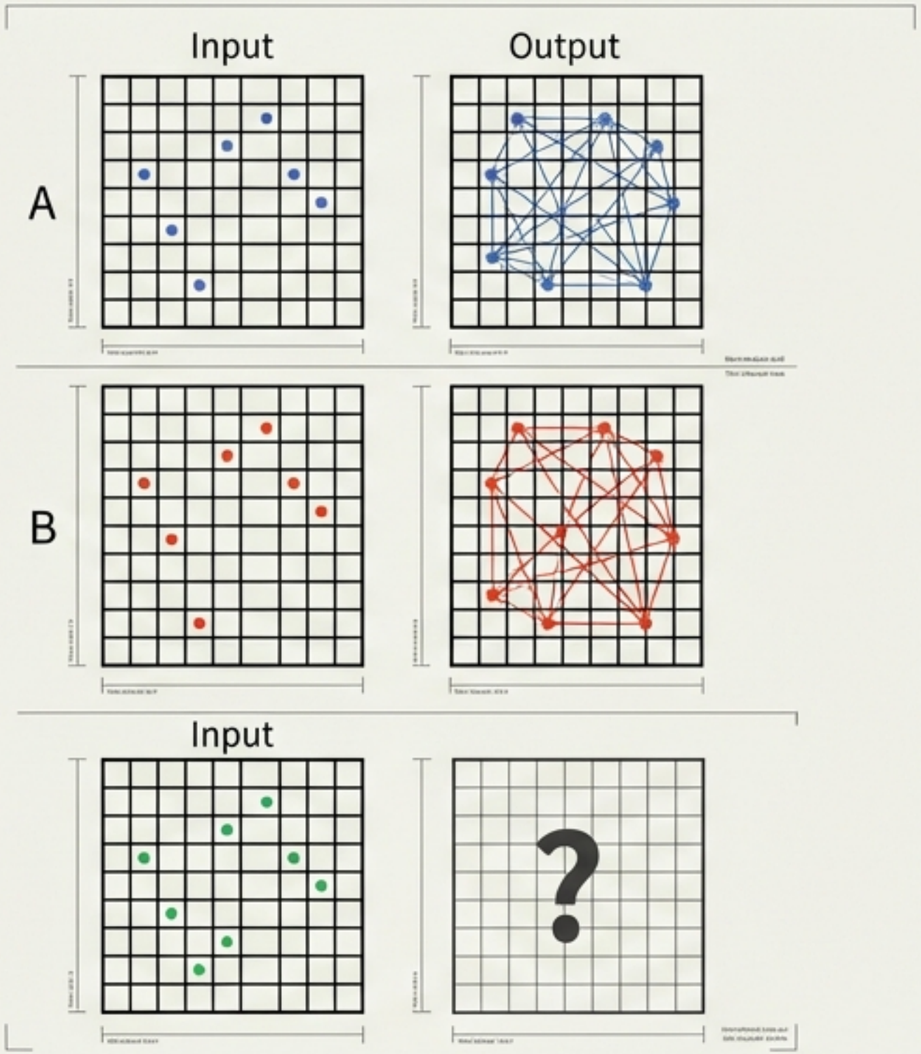


ソフトな汚染：問題の変数を変えただけのデータが訓練セットに含まれている場合、モデルは「思考」しているのではなく、解法テンプレートを「検索」しているに過ぎない。

「これは未知の規則への適応ではなく、過去のデータの高度な補間である。」

真の試金石：ARC-AGI-2とは何か

Train



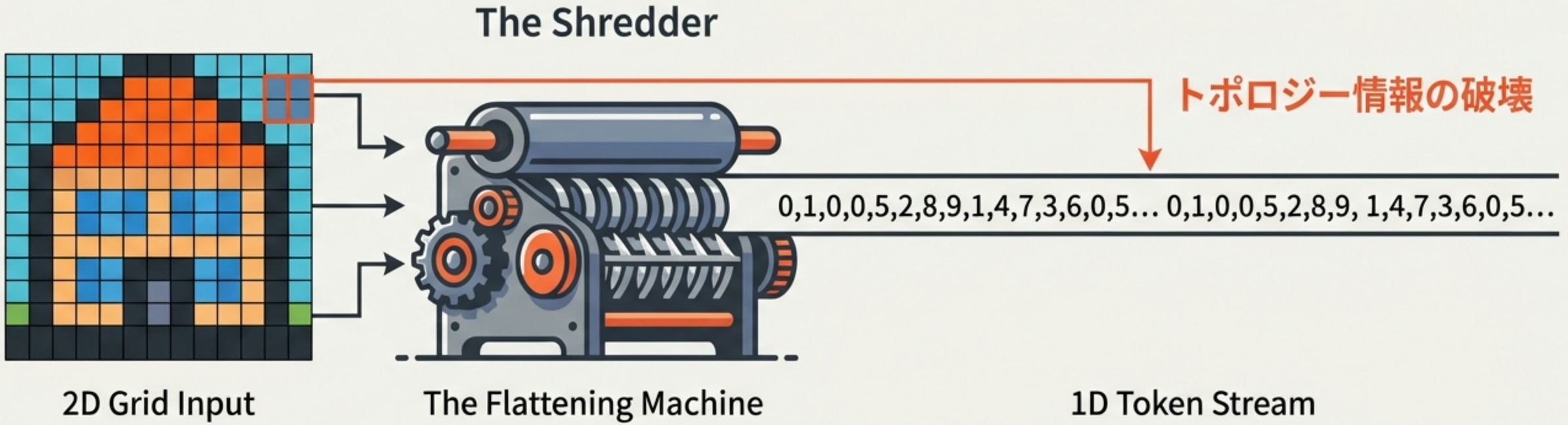
ARC-AGI-2: 3つの基準

1. **完全な初見** (Novelty): インターネット上に存在しないユニークなルール。
2. **コア知識のみ** (Core Priors): オブジェクトの永続性、対称性など、幼児レベルの物理的直感のみを使用。
3. **効率性** (Efficiency): ブルートフォースを排除し、限られたリソースでの解決を要求。

人間とAIの断絶：直感と計算のギャップ

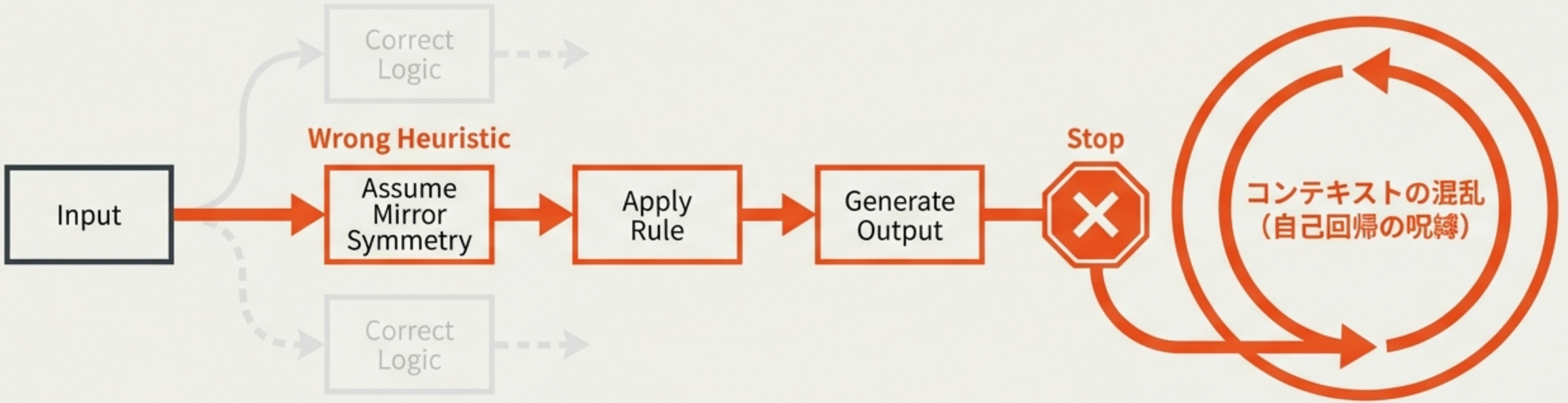
人間にとってARCは「視覚的な直感」であり、特別な訓練なしに解ける。一方、LLMにとっては「理解不能な数字の羅列」である。このギャップは、現在のAIが記号の意味を理解していない（シンボルグラウンディング問題）ことを示唆している。

構造的欠陥 I：空間コンテキストの喪失 (The Flattening)



シンボルグラウンディング問題：トランスフォーマーは2次元画像を「1次元のトークン列」として処理する。この過程で、上下左右の「隣接性」や「内側/外側」といった空間的コンテキストが失われる。モデルはピクセルを空間的実体ではなく、単なる数値の並びとして計算している。

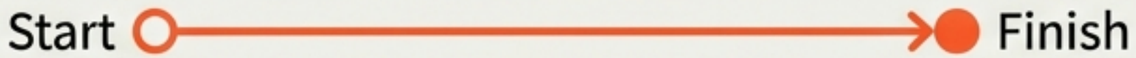
構造的欠陥 II：構成的推論の失敗



中国製モデルは単純なグローバルルールは扱えるが、「AならX、BならY」といった条件分岐や構成的推論において、一度誤ったヒューリスティクス (安易なパターン) に固執すると、自己回帰的な性質により軌道修正ができなくなる。

思考システムの欠如：System 1 vs System 2

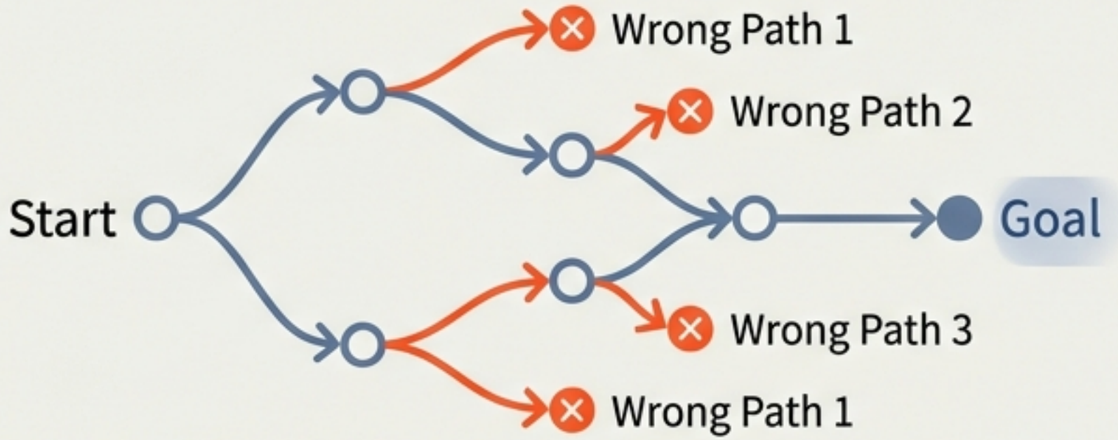
System 1 (Chinese Models)



直感・高速 (DeepSeek V3.2, Qwen 3.5)

一発書きの生成 (Single-shot)。パターン認識。
ARCスコア: 1.3% (失敗)

System 2 (Reasoning Models)



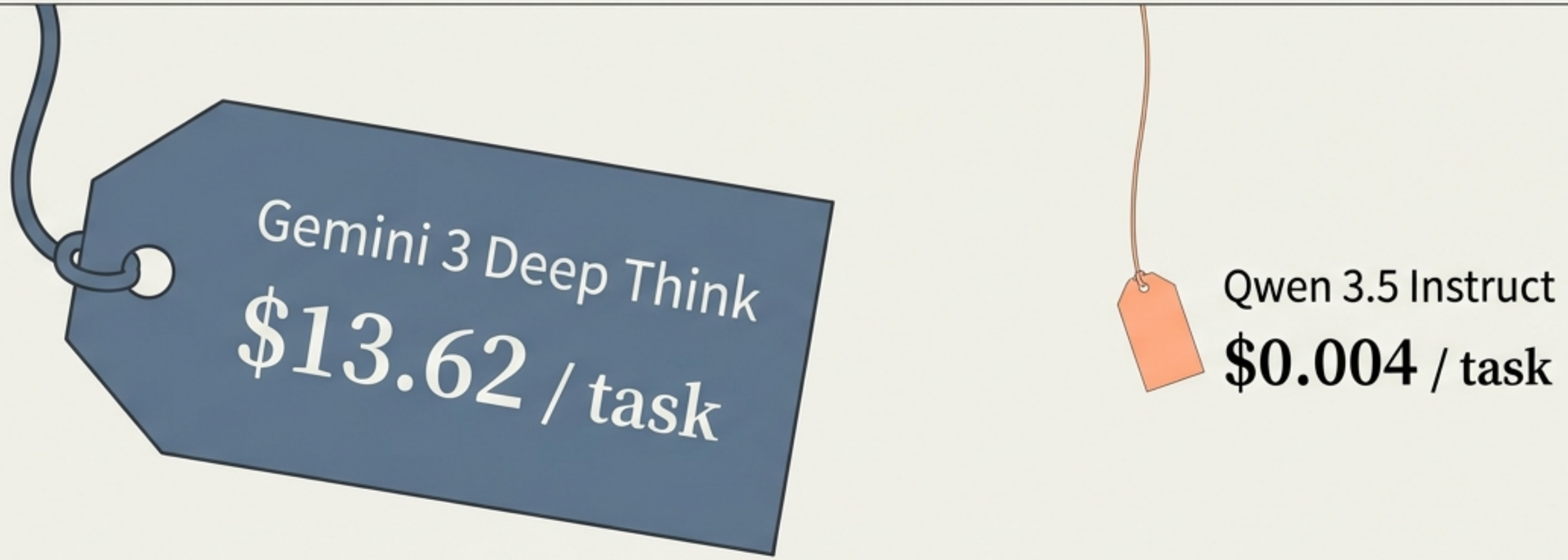
熟慮・低速 (Gemini 3 Deep Think)

探索・自己検証 (Tree Search) 。論理推論。
ARCスコア: 84.6% (成功)

ARCを解くには「立ち止まって考える」機能が不可欠だが、標準的なオープンモデルは「話し続ける」ことしかできない。

Swiss Scientific Editorial

知能効率のジレンマ：コストと推論のトレードオフ

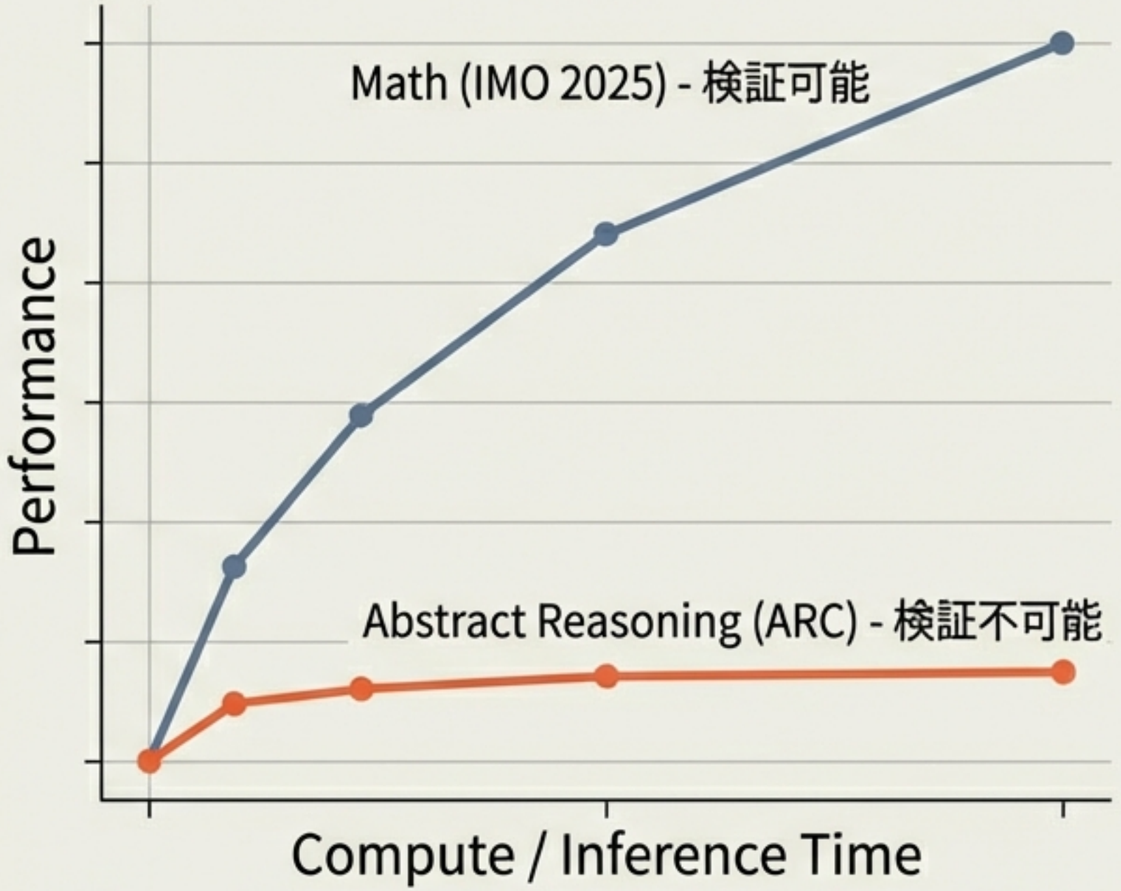


Googleの戦略 (Test-Time Compute):
莫大な計算資源を投入して並列推論を行う「ブルートフォース推論」。スコアは高いがコストは法外。

中国の戦略 (Inference Efficiency):
APIの実用性と民主化を優先し、推論コストを極限まで下げる。スコアは低いが実用的。

中国製モデルの低スコアは技術的敗北ではなく、「効率性」を選んだ戦略的選択の結果である。

例外と限界：DeepSeek-V3.2-Speciale



DeepSeek-V3.2-Specialeは推論時間を増やした強化モデル。数学ではGeminiに匹敵するが、ARCでは依然として苦戦する。

ARCのような「未知の抽象ドメイン」では、探索空間が無限に近く、モデル自身が正解かどうかを判定する「検証関数」を構築できないため、単に計算量を増やしても正解に到達できない。

総括：「推論の壁」の正体

Strategy（戦略）

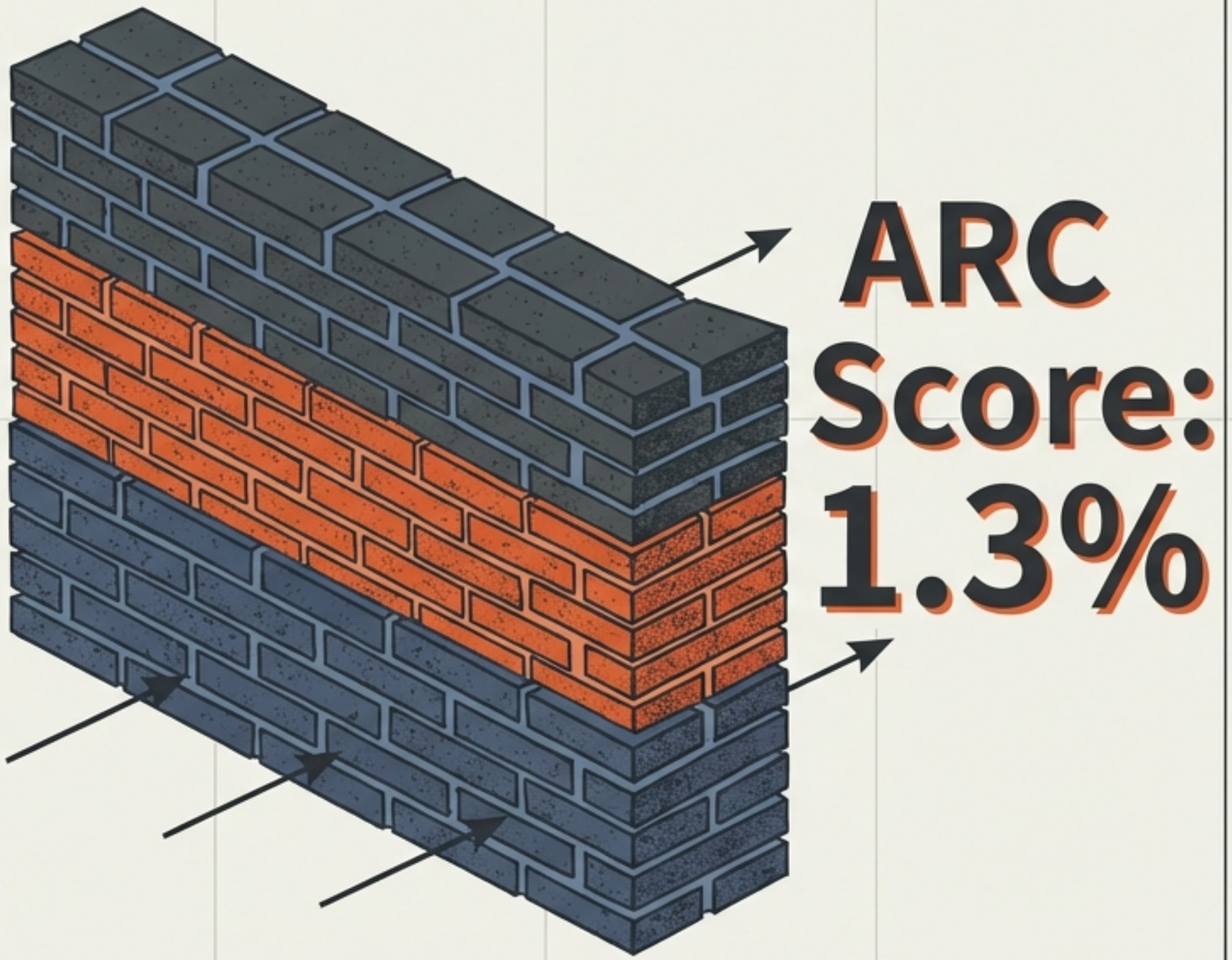
実用性を重視する中国モデルは、ARCに必要な「非効率な探索（System 2）」を意図的に排除している。

Learning (学習)

データ汚染のない「完全な未知」に対して、パターン補間は無力である。

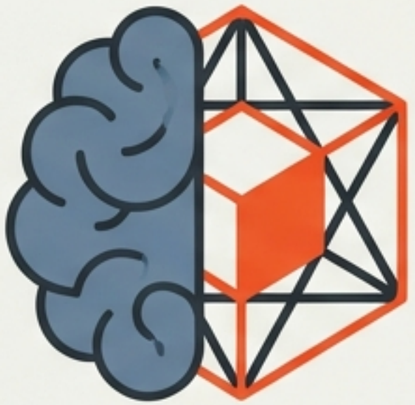
Architecture (アーキテクチャ)

1次元トランスフォーマーは2次元の位相幾何学をネイティブに理解できない。



AGIへの道筋：スケーリング則のその先へ

Neuro-symbolic AI（ニューロシンボリック）



空間的関係性やグラフ構造をネイティブに処理できるアーキテクチャ。

Test-Time Training（テスト時学習）

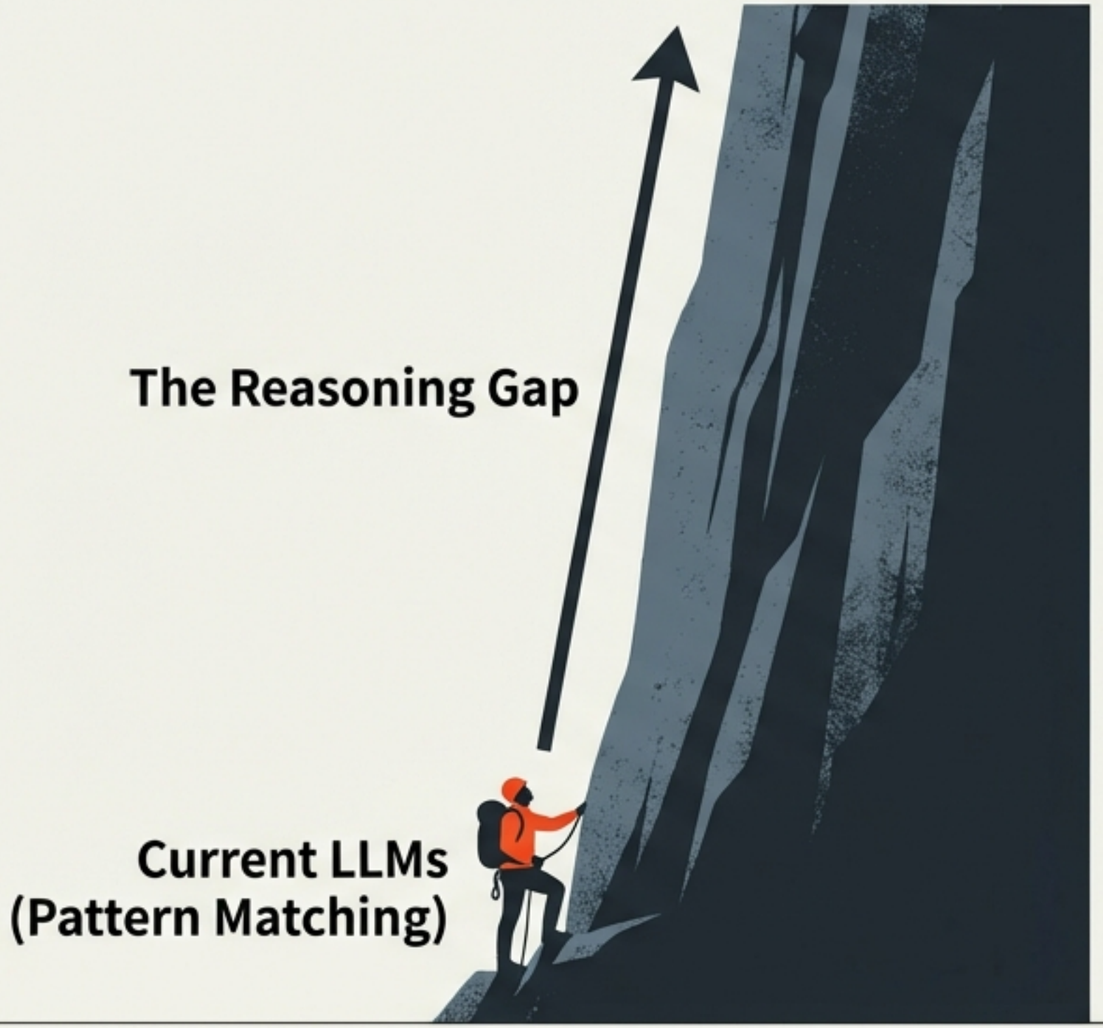
モデルが問題に合わせてリアルタイムに重みを更新し適応する。

System 2 Integration（探索的推論）

生成ではなく、探索と自己検証をループさせる推論エンジンの統合。

「補間（Interpolation）」から「外挿（Extrapolation）」へ。既知のパターンの適用から、未知のルールの発見へ。

結論：1.3%からの挑戦



中国製LLMのARCにおける低スコアは、敗北ではなく「現在地の正確な地図」である。

人間が2分で解けるパズルにスーパーコンピュータが屈する現状。このギャップを埋めるプロセスこそが、真の汎用人工知能(AGI)への最短ルートとなる。