中国製LLMはなぜARC-AGI-2で苦戦するのか：高難度推論ベンチマークが示すAIの現在地

発行日: 2026年2月22日

著者: Manus AI

序論

2026年現在、大規模言語モデル（LLM）の能力は飛躍的に向上し、MMLUやMATHといった主要なベンチマークにおいて、人間を超えるスコアを記録するモデルも珍しくなくなりました。特に、DeepSeekやQwenに代表される中国発のLLMは、これらの知識集約型・スキルベースのテストで驚異的な性能を示し、世界のトップランナーとして認知されています [1]。しかし、その一方で、AIの真の汎用知能（AGI）への到達度を測るために設計された最先端の推論能力ベンチマーク「ARC-AGI-2」においては、これらの高性能モデルが軒並み極めて低いスコアに留まるという、顕著な性能の乖離が見られます。2026年2月時点のリーダーボードでは、多くの中国製LLMのスコアは1%台前半であり、トップ層のモデルが記録する80%超のスコアとは隔絶した結果となっています [2]。

本レポートでは、この「ベンチマーク間のパラドックス」に着目し、なぜ中国製LLMが他のベンチマークで高得点を獲得しながら、ARC-AGI-2では苦戦を強いられているのか、その根本的な原因を多角的に分析します。

ARC-AGI-2：知識ではなく「流動性知能」を測る試金石

ARC-AGI-2が他のベンチマークと一線を画すのは、その設計思想にあります。MMLUやGSM8Kなどが、既存の知識（結晶性知能）をどれだけ正確に記憶・応用できるかを測るのに対し、ARC-AGI（Abstraction and Reasoning Corpus for Artificial General Intelligence）は、未知の課題に直面した際に、少数の例から自律的にルールを抽出し、応用する能力、すなわち「流動性知能」を評価することに特化しています [3]。

> ARC-AGIは、広範な事前経験やドメイン固有のトレーニングに依存せず、効率的に新しい問題を推論し解決する能力、すなわちより一般的で人間らしい流動性知能の形態を評価するために設計されました。
>
> — ARC-AGI-2 Technical Report [4]

ARC-AGI-2は、初代のARC-AGI-1が抱えていた課題（計算力による総当たり攻撃への脆弱性など）を克服し、AIの推論能力の限界をより精密に測定するために、2025年に導入されました。そのタスクは、以下のような人間にとっては直感的でありながら、現在のAIにとっては極めて困難な特性を持っています [4] [5]。

ARC-AGI-2の主要な挑戦	説明
記号的解釈	視覚的なパターンを、それ自体が持つ意味を超えた抽象的な記号として解釈する能力。
構成的推論	複数のルールが相互作用する中で、それらを同時に、あるいは段階的に組み合わせて適用する能力。
文脈に応じたルール適用	周囲の状況（コンテキスト）に応じて、適用すべきルールを動的に変更・選択する能力。
新規性の高さ	全てのタスクが完全に新規であり、訓練データからの記憶やパターンマッチングでは解決できない。

これらの要求は、LLMが膨大なテキストデータから学習した統計的相関に基づく「次に来る単語の予測」という基本原理とは根本的に異なり、真の理解と推論を必要とします。結果として、純粋なLLMはARC-AGI-2でほぼ0%というスコアを記録しており、このベンチマークが現在のAI技術の根源的な課題を浮き彫りにしていることがわかります [6]。

中国製LLMの強みと、その限界

DeepSeekやQwenといった中国製LLMは、特定の領域、特に「検証可能なタスク」において世界最高水準の性能を誇ります。これらは、数学、コーディング、科学といった、明確な正解が存在し、その正しさをプログラム的に検証できる領域です。

この強みの背景には、**強化学習（RL）を中心とした独自の訓練戦略**があります。特にDeepSeek-R1の論文では、人間の専門家によるラベル付け（SFT）を必要とせず、タスクの正解・不正解という報酬信号のみを用いてモデルの推論能力を向上させる手法が詳述されています [7]。このアプローチは、検証可能なドメインにおいて自己対戦や試行錯誤を繰り返すことで、極めて効率的に性能を高めることを可能にしました。その結果、MATHやHumanEvalといったベンチマークでトップクラスのスコアを達成しています。

しかし、この戦略は諸刃の剣でもあります。検証可能なタスクに特化して最適化されたモデルは、正解が一つに定まらず、ルール自体をその場で発見する必要があるARC-AGI-2のような「開放的な推論タスク」への汎化能力を十分に獲得できません。モデルは、明確な報酬が得られない未知の環境で、どのように思考の連鎖（Chain-of-Thought）を構築し、どの戦略が有望かを探求する術を知らないのです。

さらに、既存のベンチマークには**「データ汚染（Data Contamination）」**の問題が常につきまといます。これは、ベンチマークのテスト問題が、意図的か否かにかかわらずモデルの訓練データに含まれてしまい、モデルが推論ではなく「記憶」によって解答してしまう問題です [8]。

中国製LLMが高いスコアを出す背景には、このデータ汚染の可能性も指摘されており、真の汎化能力を測る上での課題となっています。ARC-AGI-2は、全てのタスクが新規に作成され、非公開であるため、このデータ汚染のリスクを極限まで低減しており、モデルの「素の」推論能力を厳格に評価します。

ARC-AGI-2で高得点を出すモデルとの比較

2026年2月、GoogleのGemini 3 Deep ThinkがARC-AGI-2で84.6%という驚異的なスコアを記録し、大きな注目を集めました [9]。この成功は、単一の巨大なモデルによるものではなく、**テスト時計算（Test-Time Compute）**、**プログラム合成（Program Synthesis）**、そして複数の思考プロセスを組み合わせる**洗練された推論ハーネス（Reasoning Harness）**といった、システムレベルのアプローチの賜物です [10]。

これらのシステムは、問題に直面した際に、単一の答えを出すのではなく、

1. 複数の異なる推論経路（プログラムや思考の連鎖）を生成する。
2. それぞれの経路の妥当性を自己評価・検証する。
3. 最も有望な解を統合・洗練させて最終的な答えを導き出す。

といった、より人間らしい動的な問題解決プロセスを実行します。これは、中国製LLMが得意とする特定のドメインに最適化された単一の推論戦略とは対照的です。ARC-AGI-2での成功は、モデルの巨大さ以上に、いかにして柔軟で適応的な推論アーキテクチャを構築するかにかかっていることを示唆しています。

結論

中国製LLMが他のベンチマークで高い評価を得ながらARC-AGI-2で苦戦する理由は、単一の要因に起因するものではなく、複数の構造的な問題が絡み合った結果です。

1. **ベンチマークの性質の違い**：中国製LLMは、知識の記憶と応用を測る「結晶性知能」のテスト（MMLU等）に強い一方、ARC-AGI-2が測る未知の課題解決能力「流動性知能」への適応ができていない。
2. **訓練戦略の特化**：強化学習を用いて数学やコーディングといった「検証可能なタスク」に最適化された訓練戦略が、ルールの発見自体を要求される開放的な推論タスクへの汎化を妨げている。
3. **アーキテクチャと推論方式の限界**：単一モデルによる静的な推論に依存しており、テスト時に動的な探索やプログラム合成を行う高度な推論ハーネスを十分に組み込めていない。

この現象は、現在のLLM開発における「スコアのための最適化」と「真の汎用知能の探求」との間のギャップを象徴しています。中国製LLMのARC-AGI-2での苦戦は、そのモデルの劣後を示すものではなく、むしろ現在のAI技術が共通して直面している、より根源的な課題——すなわち、**統計的パターン認識から、記号的操作と構成的一般化を伴う真の推論能力へといかにして**

飛躍するか——を浮き彫りにしていると言えるでしょう。ARC-AGI-2は、その困難さゆえに、次世代のAIが目指すべき方向を指し示す、重要な道標であり続けています。

参考文献

[1] Sallam, M., et al. (2025). Chinese generative AI models (DeepSeek and Qwen) rival flagship models for clinical utility and safety. PMC.

[2] ARC Prize. (2026). Leaderboard. Retrieved from

[3] ARC Prize. (n.d.). What is ARC-AGI? Retrieved from

[4] Chollet, F., et al. (2026). ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems. arXiv:2505.11831v2.

[5] Morales Aguilera, F. (2025). The Evolving Landscape of AI Intelligence: A Comparative Analysis of ARC-AGI-1 and ARC-AGI-2. Medium.

[6] Jha, A. (2025). LLMs Hit 0% on ARC-AGI-2 benchmark: Exposing the Limits of AI Generalization. LinkedIn.

[7] Guo, D., et al. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

[8] The Grigorian. (2025). When Benchmarks Lie: Why Contamination Breaks LLM Evaluation. Medium.

[9] Google. (2026). Gemini 3 Deep Think: AI model update designed for science. Google Blog.

[10] Barla, N. (2026). ARC-AGI In 2026: Why Frontier Models Still Don’ t Generalize. Adaline Labs.