中国製 LLM が ARC-AGI-2 で苦戦する構造的理由

知識の想起と真の推論の断絶

2026 年 2 月 22 日作成（最新リーダーボードデータ反映）

MATH-500 で 97%を叩き出す DeepSeek R1 が、ARC-AGI-2 ではわずか 1.3%しか取れない。この衝撃的なギャップは、現行 LLM が「知識の想起」と「真の推論」の間に抱える根本的な断絶を浮き彫りにしている。ARC-AGI-2 は 2025 年 3 月に François Chollet 率いる ARC Prize Foundation が公開した次世代ベンチマークで、従来のベンチマークとは全く異なる認知能力——「流動性知能（fluid intelligence）」を測定する[1, 2]。2026 年 2 月 22 日時点の最新リーダーボードでは、Google の Gemini 3 Deep Think が 84.6%でトップに立ち、グランプリ条件（85%）に迫る一方、中国製 LLM の最高スコアは Kimi K2.5 の 11.8%にとどまり、その差は 70 ポイント以上に拡大している[8]。

1. ARC-AGI-2 は何が根本的に違うのか

ARC-AGI-2 は、François Chollet と Mike Knoop（Zapier 共同創業者）が共同設立した ARC Prize Foundation によって 2025 年 3 月 24 日にリリースされた[2, 10]。前身の ARC-AGI-1（2019 年発表）が 2024 年末に OpenAI o3 によって 85%超を達成され飽和状態に近づいたことを受け、より困難な後継版として設計された[7]。

従来のベンチマーク（MMLU、HumanEval、MATH 等）が測定するのは**結晶性知能（crystallized intelligence）**——訓練データからの知識想起能力である。これに対 ARC-AGI-2 が測定するのは**流動性知能**——未知の問題に対して最小限の例示から規則を発見し適用する能力だ[1, 5]。各タスクは一意で事前暗記不可能。専門知識や言語能力は不要で、「物体の永続性」「基本的幾何学」「初等的数感覚」といった人間の基本的認知プリミティブのみを要する。

評価セットは各 120 タスク、各タスク 2 回の回答機会（Pass@2）。400 名以上の統制実験で人間パネルスコア 100%、平均個人スコア約 60%、1 タスク平劇 2.7 分であった[1, 2]。学歴や専門分野はスコアと無相関であり、汎用的な問題解決能力を測定していることを裏付けている。

2. 中国製 LLM のスコアが示す衝撃的な落差

以下の表は、2026 年 2 月 22 日時点の ARC Prize 公式リーダーボードにおける中国製 LLM のスコアを示している。

モデル	Author	ARC-AGI-1	ARC-AGI-2	Cost/Task
Kimi K2.5	Moonshot AI	65.3%	**11.8%**	$0.280
Minimax M2.5	Minimax	63.7%	**4.9%**	$0.170
Deepseek R1	Deepseek	15.8%	**1.3%**	$0.080

Deepseek R1 (05/28)	Deepseek	21.2%	**1.1%**	$0.053
Qwen3-235b-a22b (25/07)	Alibaba	11.0%	**1.3%**	$0.004

表 1: 中国製 LLM の ARC-AGI-2 スコア（2026 年 2 月 22 日時点） [8]

中国製 LLM の中で最も高いスコアは Moonshot AI の Kimi K2.5 の 11.8%である。一方、DeepSeek R1 は MATH-500 で 97.3%、MMLU で 90.8%というトップクラスの成績を収めながら、ARC-AGI-2 ではわずか 1.3%にとどまる[3, 8]。Qwen3 についてはチーム自身が ARC-AGI で 41.8%と報告したが、ARC Prize Foundation による独立検証でな ARC-AGI-1 で 11%、ARC-AGI-2 で 1.3%にとどまり、約 4 倍の乖離が指摘されている[9]。

3. 西側モデルとの比較が明らかにする格差の拡大

2026 年 2 月 22 日時点の ARC-AGI-2 リーダーボード上位モデルを見ると、中国勢の停滞と西側モデルの急速な進歩が鮮明になる。

AI System	Author	ARC-AGI-2	Cost/Task	System Type
Human Panel	Human	**100.0%**	$17.00	N/A
Gemini 3 Deep Think (2/26)	Google	**84.6%**	$13.62	CoT
Gemini 3.1 Pro (Preview)	Google	**77.1%**	$0.962	CoT
GPT-5.2 (Refine.)	Johan Land	72.9%	$38.99	Refinement
Claude Opus 4.6 (120K, High)	Anthropic	69.2%	$3.47	CoT
Claude Sonnet 4.6 (High)	Anthropic	60.4%	$2.70	CoT
GPT-5.2 Pro (High)	OpenAI	54.2%	$15.72	CoT
GPT-5.2 (X-High)	OpenAI	52.9%	$1.90	CoT
Opus 4.5 (Thinking, 64K)	Anthropic	37.6%	$2.40	CoT
NVARC	ARC Prize 2025	27.6%	$0.200	Custom
Kimi K2.5	Moonshot AI	**11.8%**	$0.280	CoT
Minimax M2.5	Minimax	**4.9%**	$0.170	CoT
Deepseek R1	Deepseek	**1.3%**	$0.080	CoT
Qwen3-235b (25/07)	Alibaba	**1.3%**	$0.004	Base LLM

表 2: ARC-AGI-2 リーダーボード（上位モデルと中国製モデル、赤字は中国製） [8]

最も注目すべきは、Google の Gemini 3 Deep Think が 84.6%でグランプリ条件（85%）にあとわずか 0.4 ポイントに迫っている点だ[8]。また Gemini 3.1 Pro (Preview)は 77.1%をコスト $0.962/タスクで達成しており、コスト効率の面でも突出している。Claude Opus 4.6 は 69.2%を$3.47 で達成し、人間個人平均（~60%）を超えた。一方、中国製 LLM の最高スコアは Kimi K2.5 の 11.8%であり、**トップとの差は 72.8 ポイント**に達している。

さらに重要なのは、主要 AI ラボ（Google、OpenAI、Anthropic、xAI）がすべてモデルカードに ARC-AGI スコアを記載するようになった一方、中国のラボは ARC-AGI スコアの公式報告を

行っていないという事実だ[6]。

4. なぜ中国製 LLM は ARC-AGI-2 で失敗するのか——技術的分析

ARC-AGI-2 Technical Report は、フロンティア AI 推論システムの失敗を 3 つのカテゴリに分類している[1, 5]。

4.1 記号解釈（Symbolic Interpretation）の失敗

AI システムは対称性チェックやミラーリングなどの操作は試みるが、記号そのものに文脈依存的な意味を付与することができない。ARC-AGI-2 では入力例中で記号の意味が暗黙的に定義され、それを理解して適用する必要がある[5, 12]。これは LLM のパターンマッチングアプローチでは根本的に対処困難である。

4.2 構成的推論（Compositional Reasoning）の失敗

単一のグローバルルールは適用できても、複数のルールが相互作用する場合に破綻する。ARC-AGI-2 は複数規則の同時発見・適用を要求する[1, 5]。

4.3 文脈依存的ルール適用の失敗

モデルは表層的パターンに固着し、文脈に応じたルール変化の「背後の選択原理」を理解できない[5]。Chollet はこれを「LLM はベクトルプログラムのリポジトリとして機能するが、その場で新しいプログラムを合成する能力が欠けている」と表現した[15]。

5. 強化学習アプローチの限界とテストタイム計算の壁

DeepSeek R1 が採用する GRPO ベースの強化学習は、検証可能な報酬信号がある領域（数学、コーディング）では極めて有効だが、ARC-AGI-2 のタスクには RL 訓練に組み込める自然な検証器が存在しない。DeepSeek R1 の論文自体が「純粋な RL の成功は信頼性の高い報酬信号に依存する」と認めている[3]。

R1-Zero の純粋 RL 訓練は自己検証や反省といった推論行動を自発的に発現させたが、ARC Prize Foundation の分析は「SFT は CoT 推論のドメイン汎用性を高めるために必要」と結論づけた[4]。RL だけではドメイン間の汎化が困難なのだ。

テストタイム計算の大量投入も、ARC-AGI-2 では壁にぶつかる。ARC Prize Foundation は「対数線形スケーリングでは ARC-AGI-2 を突破できない」と述べている[2, 7]。ただし、2026 年 2 月時点で Gemini 3 Deep Think が 84.6%に到達したことは、Refinement ループや新たな推論アーキテクチャの有効性を示唆している[8]。

また、NVIDIA の NVARC チームはカスタムアプローチで 27.6%を$0.20/タスクで達成しており、規模よりもアーキテクチャの革新が重要であることを示している[8, 14]。

6. 中国勢が出遅れる背景——ベンチマーク最適化文化と研究方向性

第一に、研究の優先順位の違い。DeepSeek、Alibaba（Qwen）、Moonshot の公開研究は、数学・コーディング・知識 QA といった検証可能な報酬信号に基づく RL 最適化に集中している。ARC-AGI-2 で進歩を示した西側のアプローチ——Refinement ループ、テストタイム学習、合成データによる事前適応——は、中国のラボから公には報告されていない[3, 6]。

第二に、ベンチマーク報告の姿勢の違い。主要 AI ラボがモデルカードに ARC-AGI スコアを記載する一方、中国のラボは ARC-AGI-2 スコアの積極的な報告・分析を行っていない。Qwen3 のスコア乖離問題（自己報告 41.8% vs 検証値 11%）は透明性の懸念も提起している[6, 9]。

第三に、Chollet が指摘する「スキルと知能の混同」。MMLU や MATH での高得点は知識想起能力の高さを示すが、ARC-AGI-2 はまさにその能力が通用しない領域を標的にしている。Chollet は「スキルと知能は別物だ」と述べ、Meta の Yann LeCun も「LLM は人間レベルの知能への行き止まり」と評している[15]。

7. 結論——ARC-AGI-2 が突きつける問いの本質

2026 年 2 月 22 日時点で、ARC-AGI-2 のリーダーボードは現行 AI 研究の重要な真実を浮き彫りにしている。Gemini 3 Deep Think が 84.6%でグランプリ条件（85%）に迫り、Claude Opus 4.6 が 69.2%で人間個人平均を超えるなか、中国製 LLM の最高スコアは Kimi K2.5 の 11.8%にとどまる[8]。DeepSeek R1 の MATH-500 で 97.3%と ARC-AGI-2 で 1.3%といけ 96 ポイントの落差は、「知識の蓄積と検索」と「未知の問題への適応」が根本的に異なる認知能力であることの定量的証拠だ[3, 4]。

中国製 LLM が低スコアにとどまる理由は、（1）トランスフォーマーの次トークン予測が統計的パターン補完を最適化し真のプログラム合成を行わないこと、（2）RL 訓練が狭いドメインに最適化されること、（3）テストタイム計算の大量投入でも対数線形的にしかスケールしないこと、（4）Refinement ループ等の新パラダイムへの取り組みが公にはなされていないこと——の 4 層構造で説明できる[1, 2, 3, 5, 6]。

Gemini 3 Deep Think が 84.6%に到達しグランプリ条件（85%、$0.42/タスク以下）に迫る一方、そのコストは$13.62 と基準を大幅に上回る。2026 年 3 月にはさらに困難な ARC-AGI-3（インタラクティブ推論を要求）のリリースが予定されている[2, 6]。ARC-AGI-2 が示しているのは、現在の LLM パラダイムの延長線上には汎用知能が存在しない可能性であり、中国製 LLM の低スコアはその最も鮮明な表出にほかならない。

参考文献

[1] Chollet, F. et al., "ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems," arXiv:2505.11831v2, 2025. https://arxiv.org/html/2505.11831v2

[2] ARC Prize Foundation, "Announcing ARC-AGI-2 and ARC Prize 2025," arcprize.org, March 2025. https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025

[3] DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," arXiv:2501.12948v1, 2025. https://arxiv.org/html/2501.12948v1

[4] ARC Prize Foundation, "R1-Zero and R1 Results and Analysis," arcprize.org, 2025. https://arcprize.org/blog/r1-zero-r1-results-analysis

[5] ARC Prize Foundation, "ARC-AGI-2 Technical Report," arcprize.org, 2025. https://arcprize.org/blog/arc-agi-2-technical-report

[6] ARC Prize Foundation, "ARC Prize 2025 Results and Analysis," arcprize.org, 2025. https://arcprize.org/blog/arc-prize-2025-results-analysis

[7] ARC Prize Foundation, "OpenAI o3 Breakthrough High Score on ARC-AGI-Pub," arcprize.org, 2024. https://arcprize.org/blog/oai-o3-pub-breakthrough

[8] ARC Prize Foundation, "ARC-AGI-2 Leaderboard," arcprize.org, accessed February 22, 2026. https://arcprize.org/arc-agi/2/

[9] BigGo News, "Qwen3-235B-A22B-Thinking-2507 Faces Benchmark Accuracy Questions," biggo.com, July 2025. https://biggo.com/news/202507251922_Qwen3_Benchmark_Accuracy_Questioned

[10] TechCrunch, "A new, challenging AGI test stumps most AI models," March 2025. https://techcrunch.com/2025/03/24/a-new-challenging-agi-test-stumps-most-ai-models/

[11] IntuitionLabs, "GPT-5.2 & ARC-AGI-2: A Benchmark Analysis of AI Reasoning," 2025. https://intuitionlabs.ai/articles/gpt-5-2-arc-agi-2-benchmark

[12] Labellerr, "Is Your AI Smart Enough? Test It with ARC AGI v2!" 2025. https://www.labellerr.com/blog/arc-agi-v2/

[13] eWEEK, "New AI Benchmark ARC-AGI-2 'Significantly Raises the Bar for AI'," 2025. https://www.eweek.com/news/ai-benchmark-arc-agi-2/

[14] NVIDIA Developer, "NVIDIA Kaggle Grandmasters Win Artificial General Intelligence Competition," 2025. https://developer.nvidia.com/blog/nvidia-kaggle-grandmasters-win-artificial-general-intelligence-competition/

[15] Freethink, "LLMs are a dead end to AGI, says François Chollet," 2025. https://www.freethink.com/robots-ai/arc-prize-agi

[16] LessWrong, "ARC-AGI-2 human baseline surpassed (updated)," 2025. https://www.lesswrong.com/posts/DX3EmhmwZjTYp9PBf/ai-performance-has-surpassed-a-human-baseline-on-arc-agi-2