中国製 LLM が ARC-AGI-2 で低スコアにとどまる理由：他ベンチマーク高得点との乖離を分析

エグゼクティブサマリー

2025 年 3 月に ARC Prize Foundation が公開した ARC-AGI-2 は、AI の「流動的知能」（未知の状況への適応能力）を測定するベンチマークである。2026 年 2 月現在、DeepSeek や Qwen 等の中国製 LLM は、SWE-Bench、AIME、GPQA-Diamond など従来型ベンチマークで欧米フロンティアモデルに匹敵するスコアを叩き出している。にもかかわらず、ARC-AGI-2 においては大幅に低いスコアにとどまっている。この乖離は中国製モデルに限った問題ではないが、ARC-AGI-2 で高得点を達成するために必要な「テスト時計算（test-time compute）」の大規模投資において、欧米大手に対して遅れをとっていることが主因である。[1][2][3]

ARC-AGI-2 の設計思想：なぜ従来ベンチマークと違うのか

ARC-AGI-2 は、François Chollet が提唱した「流動的知能」の概念に基づき、AI が訓練データに含まれない未知のパターンに対して効率的に適応できるかを評価する。従来のベンチマークとの根本的な違いは以下の通りである。[4][1]

- **記憶ではなく適応を測定**：MMLU、GPQA-Diamond、HumanEval などは学習済み知識の正確な再現・適用を問うが、ARC-AGI-2 は学習データに存在しないパターンの即座の理解を要求する[5][6]
- **ブルートフォースの排除**：ARC-AGI-1 では大量の計算資源投入で高得点が可能だったが、ARC-AGI-2 はそれを防ぐ設計に改良された[7][1]
- **効率性の評価**：スコアだけでなく 1 タスクあたりのコストも評価指標に含まれる。人間は 1 タスク約 17 ドルで解決可能だが、AI システムには数百ドルかかる場合がある[8]

AI が苦手とする 3 つの認知課題

ARC-AGI-2 のテクニカルレポートは、現行の AI 推論システムが特に苦手とする認知課題を 3 つ特定している。[2]

1. **シンボルの意味的解釈**：記号を視覚パターン以上の「意味」として解釈する能力が欠如している
2. **複数ルールの同時適用**：相互作用する複数のルールを同時に適用するタスクで失敗する。単一のグローバルルールは適用可能だが、複合ルールは処理できない
3. **文脈依存的ルール適用**：文脈に応じてルールを異なる方法で適用する必要がある場合、表面的なパターンに固執し、根底にある選択原理を理解できない

これらの課題は、単にモデルサイズを拡大したり、学習データを増やしたりするだけでは解決できない「質的な壁」を意味する。[9][8]

ARC-AGI-2 リーダーボードの現状（2026 年 2 月）

主要モデルのスコア推移

2026 年 2 月 20 日時点の ARC-AGI-2 スコアを時系列で整理すると、欧米フロンティアモデルの急速な改善と、中国製モデルの停滞が鮮明になる。

モデル	開発元	ARC-AGI-2 スコア	時期	備考
Gemini 3 Deep Think（アップグレード版）	Google	84.6%	2026 年 2 月	現行最高スコア（公式モデル）[10][11]
Gemini 3.1 Pro	Google	77.1%	2026 年 2 月	[12]
Johan Land 氏システム（GPT-5.2 改造）	個人	72.9%	2026 年 2 月	GPT-5.2 ベースのカスタムシステム[13]
Claude Opus 4.6	Anthropic	68.8%	2026 年 2 月	[14]
GPT-5.2 Thinking	OpenAI	52.9%	2025 年末	[15]
GPT-5.1	OpenAI	17.6%	2025 年	[15]
Gemini 2.5 Pro	Google	4.9%	2025 年	[16]

o3-low	OpenAI	4.0%	2025年3月	[17]
DeepSeek R1	DeepSeek	1.3%	2025年3月	[18][1]
o1-pro	OpenAI	1.2%	2025年3月	[8]
GPT-4.5 / Claude 3.7 Sonnet	各社	~1%	2025年3月	[1]

注目すべきは、2025年3月の初期段階では中国製モデル（DeepSeek R1：1.3%）も欧米モデル（o1-pro：1.2%、GPT-4.5：~1%）も同様に低スコアだった点である。しかし、その後の約1年間でGPT-5.2が52.9%、Gemini 3 Deep Thinkが84.6%と飛躍的に向上したのに対し、中国製モデルの公表されたARC-AGI-2スコアの大幅な更新は確認されていない。[18][1]

2026年春節の中国モデルラッシュとベンチマーク

2026年2月の春節前後に、Qwen 3.5、GLM-5、MiniMax M2.5、Kimi K2.5が一斉にリリースされた。これらは従来型ベンチマークで欧米フロンティアモデルに匹敵する性能を示している。[3]

ベンチマーク	Qwen 3.5	GLM-5	Kimi K2.5	Claude Opus 4.6	GPT-5.2
SWE-Bench Verified	76.4%	77.8%	76.8%	80.9%	80.0%
AIME 2025	91.3	92.7	96.1	92.8	100
GPQA-Diamond	—	86.0	87.6	87.0	92.4
BrowseComp	69.0	76%	78.4	37.0-59.2	57.8-65.8

出典：各社自社報告ベース[3]

しかし、これらの春節モデル群のARC-AGI-2スコアは公式リーダーボードに登場していない。各社が公表したベンチマーク一覧にもARC-AGI-2は含まれておらず、テスト自体を実施していないか、低スコアのため公表を避けている可能性がある。

中国製LLMがARC-AGI-2で低スコアにとどまる構造的要因

要因 1：テスト時計算（Test-Time Compute）への投資格差

ARC-AGI-2 で高スコアを達成している欧米モデルに共通するのは、「テスト時計算」の高度化である。GPT-5.2 や Gemini 3 Deep Think は、推論時にモデルが「より長く考える」能力を持ち、複雑な問題に対してチェーン・オブ・ソート（思考の連鎖）を深く展開できる。[15][11]

Gemini 3 Deep Think の 84.6%達成は、まさにこの「テスト時計算」アプローチの成果であり、Google は「応答生成前にモデルがより長く『考える』能力に焦点を当てた」と明言している。GPT-5.2 も、セルフベリフィケーション（自己検証）やプロンプティング技術の成熟の恩恵を受け、ARC-AGI-1 の 72.8%から ARC-AGI-2 の 52.9%へと、前世代比で大幅な改善を実現した。[11][15]

一方、中国製モデルの多くは MoE（Mixture of Experts）アーキテクチャによるコスト効率を追求する方向に注力しており、推論時に大量の計算資源を投入する「テスト時計算」への注力は相対的に少ない。中国製モデルの設計哲学は「最高性能かつ最安クラス」のポジションを目指すものであり、1 タスクあたり数十ドルの計算コストをかける ARC-AGI-2 対策とは方向性が異なる。[3]

要因 2：ARC-AGI-2 が測定する能力の本質的な違い

従来型ベンチマーク（SWE-Bench、AIME、GPQA-Diamond 等）は、訓練データに含まれる知識や既知のパターンの応用を問う「結晶性知能」に近い。大量の高品質データで訓練され、推論チェーンを最適化した中国製モデルは、このタイプのタスクに極めて強い。[6][5]

対照的に、ARC-AGI-2 は「流動的知能」――未知の問題への即時適応――を測定する。これは以下の点で根本的に異なる。[2]

- **訓練データの量やモデルサイズでは解決しにくい**：ARC-AGI-2 はタスクごとに新規のパターンを要求するため、スケーリングの恩恵を受けにくい[5][7]
- **構成的推論能力が必要**：複数のルールを同時に把握・適用する能力、文脈に応じた柔軟なルール切り替えが求められる[9][2]
- **表面的ショートカットが通用しない**：研究では、高精度のモデルでさえ「意図された抽象化」ではなく「表面的なショートカット」に基づいてタスクを解いていることが示されている[19]

要因 3：米中間の半導体規制とハードウェア制約

中国 AI 企業は米国の対中半導体輸出規制により、最先端の NVIDIA GPU へのアクセスが制限されている。GLM-5 が Huawei Ascend チップのみで学習を完了したことはその象徴的事例である。[3]

ARC-AGI-2 で高スコアを達成するには、推論時に大量の計算資源を投入する必要がある。Johan Land 氏のシステムは 1 タスクあたり約 39 ドルの計算コストを要し、o3-low でさえ 1

タスク約 200 ドルかかると推定されている。中国企業がこの規模のテスト時計算を効率的に実行するには、計算インフラの制約が障壁となりうる。[20][18]

要因 4：戦略的優先順位の相違

中国 AI 企業の競争戦略は、実用的なベンチマーク（コーディング、数学、知識問答、ブラウジング）での高得点と圧倒的な低価格を両立させることに重点を置いている。Spring 節モデルラッシュで各社が競って公表したベンチマークは、SWE-Bench、AIME、GPQA-Diamond、BrowseComp など、開発者コミュニティに直接訴求する実用的指標である。[3]

ARC-AGI-2 は研究コミュニティでは重視されているものの、現時点では顧客獲得や商業的差別化に直結するベンチマークとは見なされていない可能性が高い。中国 AI 企業にとっては、API 料金が Claude Opus 4.6 の 1/5〜1/40 という価格競争力を訴求することの方が、ARC-AGI-2 で高スコアを追求するよりも事業上の優先度が高いと考えられる。[3]

欧米モデルの急速な改善が示す技術的転換点

ARC-AGI-2 における欧米モデルの急速なスコア改善は、単なる漸進的改良ではなく、技術的なパラダイムシフトを示唆している。

GPT-5.1 から GPT-5.2 への移行で、ARC-AGI-2 スコアは 17.6%から 52.9%へと+35 ポイント跳躍した。さらに、Gemini 3 Deep Think のアップグレードにより 84.6%を達成し、人間の平均スコア（約 60%）を大幅に超えた。[10][15][11]

この急速な改善の鍵となっているのは以下の技術的進歩である。

- **深層テスト時推論**：モデルが応答前に長時間の推論チェーンを展開し、自己検証を行う能力[15][11]
- **マルチモデル反射的推論**：Johan Land 氏のシステムに見られるように、GPT-5.2 をベースにマルチモデル構成で精度を向上させるアプローチ[13]
- **視覚パターンのネイティブ処理**：Gemini のマルチモーダルアーキテクチャは、テキスト媒介の推論ではなく視覚パターンをネイティブに処理する優位性を持つ[12]

今後の見通し

中国製 LLM が ARC-AGI-2 で低スコアにとどまっている状況は、固定的なものではない。DeepSeek V4 は未発表ながらリーク情報では「別格」の性能が示唆されており、テスト時計算能力の強化が含まれている可能性がある。1M トークンのコンテキストウィンドウ拡張や Engram 技術の導入は、推論能力の質的向上を示唆する。[3]

ただし、ARC-AGI-2 で欧米モデルに追いつくには、中国 AI 企業がコスト効率最優先の戦略から、推論時の計算資源投入を伴う「深い思考」モデルの開発にもリソースを振り向ける必要がある。ARC-AGI-2 のベンチマークとしての重要性が業界全体で認知されるにつれ、この方向への投資が加速する可能性はある。

ARC Prize 2025 の目標は、ARC-AGI-2 で 85%の正確率をタスクあたり 0.42 ドルで達成することである。この「高精度かつ低コスト」という目標は、まさに中国 AI 企業が得意とする効率性追求の方向性と一致しており、中国勢が本格参入した場合には競争のダイナミクスが大きく変わる可能性を秘めている。[18]

References

1. A new, challenging AGI test stumps most AI models - The Arc Prize Foundation has a new test for AGI that leading AI models from Anthropic, Google, and D...
2. ARC-AGI-2 A New Challenge for Frontier AI Reasoning Systems - Technical context and description of the ARC-AGI-2 Benchmark
3. DeepSeek V4 も間近？中国春節 AI ラッシュ・5 モデル全解説 - あの「DeepSeek ショック」以降、中国 AI 各社は熾烈な競争を続けており、各社が DeepSeek V4 の発表前に見出しを飾ろうと競い合った結果、2026 年の春節が「 ...
4. 汎用 AI 向け新テスト「ARC-AGI-2」で人間が AI に圧勝 AI の新た ... - 米時間 2025 年 3 月 24 日、非営利法人 Arc Pr...
5. ARC-AGI In 2026: Why Frontier Models Still Don't Generalize - ARC-AGI-2 exposes the real gap: generalization efficiency under budget constraints, where refinement...
6. Best LLM for reasoning in 2026: ARC-AGI-2 benchmark results - ARC-AGI-2 is the hardest reasoning benchmark in AI. See which LLM scores highest in 2026 and what th...
7. ARC-AGI-2: A New Challenge for Frontier AI Reasoning ...
8. ARC-AGI-2：人間には簡単だが AI には難しい抽象的推論能力を ... - ARC-AGI-2 ベンチマークの誕生背景と特徴 2025 年 3 月 24 日、ARC Prize Foundation は人工知能の抽象的推論能力を評価するための新しいベンチマーク「ARC-AGI-2」をリリー...
9. AI の「本当の賢さ」を測る：新テスト「ARC-AGI-2」が暴く推論 ... - AGI の進捗を測る新ベンチマーク ARC-AGI-2 が登場。OpenAI o3 ですら苦戦する難易度。AI が苦手な「流動性知能」や「効率性」を重視する意義を解説。AGI 開発の最前線と AI の現在地を理解する...
10. Gemini 3 Deep Think Upgrade Released, Breaks ... - The upgraded version's ARC-AGI-2 score, as shown in the graph on the right of the image below, has r...
11. Is This AGI? Google's Gemini 3 Deep Think Shatters ... - Redefining AGI with 84.6% on ARC-AGI-2. The ARC-AGI benchmark is an ultimate test of intelligence. U...
12. Benchmark Wars 2026: ARC-AGI-2, GPQA Diamond, and ... - Benchmark Wars 2026: ARC-AGI-2, GPQA Diamond, and the HLE Scoring Controversy. Gemini 3.1 Pro leads ...

13. Johan Land 氏 ARC-AGI-2 で 72.9%達成 - 2026 年 2 月 3 日、ARC-AGI-2 Leaderboard に、 Johan Land 氏が GPT-5.2 に手を加えたシステムが、ARC-AGI-2（Abstraction and Reasoning...

14. Claude Opus 4.6 が ARC-AGI-2 で 68.8% - この ARC-AGI-2 で、2026 年 2 月 3 日 Johan Land 氏が GPT-5.2 に手を加えたシステムが 72.9%という驚異的な正答率を記録しましたが、商用版である Claude Opus 4.6 が ....

15. GPT-5.2 & ARC-AGI-2: A Benchmark Analysis of AI ... - An in-depth analysis of OpenAI's GPT-5.2 achieving a 54% score on the ARC-AGI-2 benchmark for abstra...

16. ARC-AGI v2 Benchmark: Complete Leaderboard & Performance Analysis (2025) - Comprehensive ARC-AGI v2 benchmark results comparing 3+ AI models from 3 organizations. Top performe...

17. 主要 AI モデルはどれも“歯が立たない”、新しい「人間には簡単 ... - 今回は、AGI（汎用人工知能）の進歩を測定するために設計された新しいベンチマークテスト「ARC-AGI-2」が登場し、最先端の AI モデルが挑戦した、その結果報告を取り上げます。

18. 新测试挑战 AI 智能水平：ARC-AGI-2 让顶尖模型碰壁 - 根据 Arc Prize 排行榜，诸如 OpenAI 的 o1-pro 和 DeepSeek 的 R1 等 “推理型” AI 模型在 ARC-AGI-2 测试中的得分仅在 1% 到 1.3% 之间，而更为强大...

19. Do AI Models Perform Human-like Abstract Reasoning ... - arXiv

20. ARC-AGI-2 と AI 汎用性評価：Johan Land 氏の 72.9%達成と多 ...