2026年における中国製LLMの性能乖離：従来型ベンチマークの飽和とARC-AGI-2における流動性知能の欠如

Gemini 3.1 pro

2026年2月20日現在、人工知能の推論能力を評価するパラダイムは根本的な転換点を迎えている。Qwen 3.5（Alibaba）やDeepSeek V3.2およびR1（DeepSeek）に代表される中国製大規模言語モデル（LLM）は、数学、プログラミング、大学院レベルの科学知識を問う従来型のベンチマークにおいて、OpenAIのGPT-5.2やGoogleのGemini 3.1 Proに匹敵、あるいは部分的に凌駕するスコアを記録している[1]。しかし、抽象的推論と未知の問題解決能力を純粋に測定する「ARC-AGI-2（Abstraction and Reasoning Corpus for Artificial General Intelligence v2）」においては、これらのトップティアモデルであっても依然として一桁台前半という極めて低い得点に低迷しているという顕著な性能乖離が観察されている[6]。

本報告書は、この性能乖離の根底にある技術的、アーキテクチャ的、そして評価論的な要因を網羅的に分析する。中国製LLMが「結晶性知能（Crystallized Intelligence）」の模倣において極めて高い効率と圧倒的なコストパフォーマンスを誇る一方で、なぜ「流動性知能（Fluid Intelligence）」の獲得において構造的な障壁に直面しているのかを解き明かす。最新のMixture-of-Experts（MoE）アーキテクチャの構造的限界、テスト時計算量（Test-Time Compute）の経済的制約、そして空間的・位相幾何学的推論における自己回帰モデルの根本的な欠陥という観点から、現在のAI開発が直面している「推論の壁（The Reasoning Gap）」の正体を詳細に論じる。

従来型ベンチマークにおける中国製LLMの覇権と結晶性知能の極致

2025年後半から2026年初頭にかけて、中国のAI研究所はアーキテクチャの最適化と大規模な強化学習（RL）を通じて、驚異的なコストパフォーマンスとベンチマークスコアを達成し、世界のAIエコシステムに多大な衝撃を与えた[7]。これらのモデルは、特定のドメイン知識やアルゴリズム的思考を問うテストにおいて、クローズドな欧米製フロンティアモデルと同等以上の能力を発揮している。

知識集約型タスクと数学的推論での圧倒的パフォーマンス

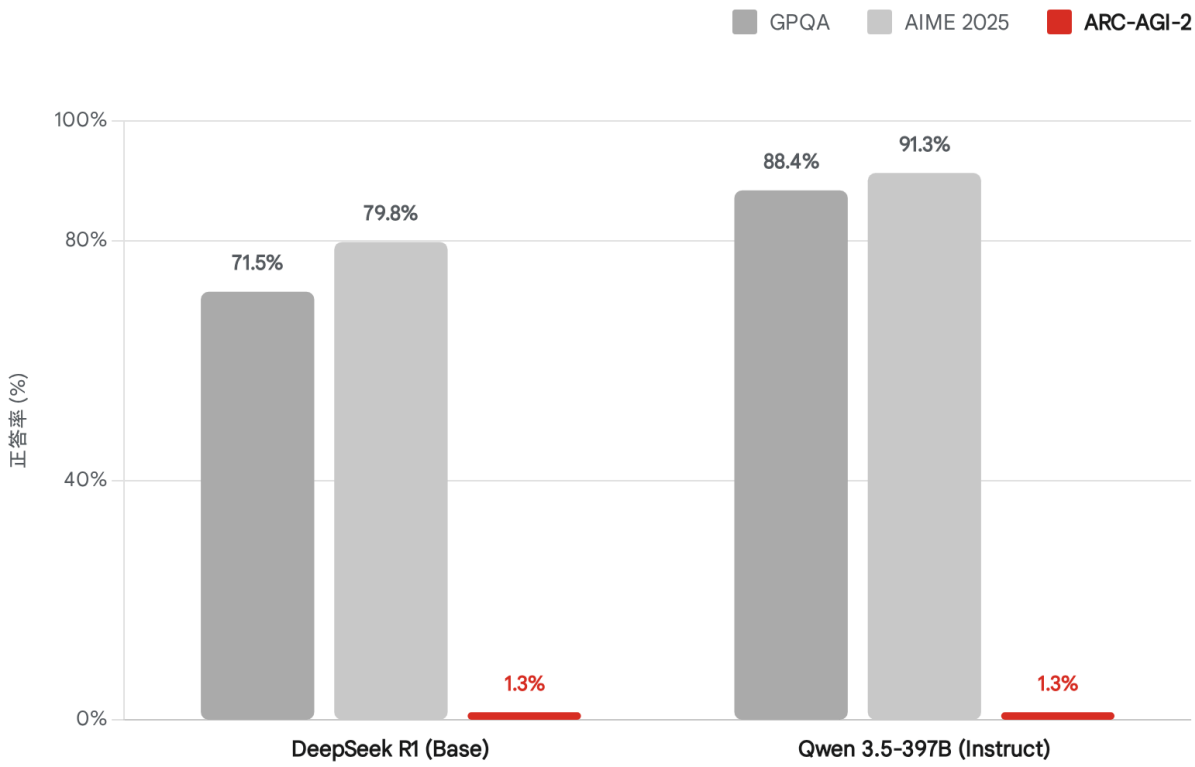
DeepSeekが2025年末にリリースしたDeepSeek V3.2は、128Kのコンテキストウィンドウと独自開発のDeepSeek Sparse Attention（DSA）を採用し、6710億（671B）の総パラメータのうち、トークンごとにわずか370億（37B）のパラメータのみをアクティブにする極めて高効率な構造を実現している[8]。このアーキテクチャの洗練により、数学オリンピックレベルの複雑な推論を要求するAIME 2025において、GPT-5 Highの94.6%を上回る96.0%という驚異的なスコアを記録した[8]。さらに、HMMT 2025（Harvard-MIT Mathematics Tournament）においては、Gemini 3 Proの97.5%を凌ぐ99.2%を達成し、

数学的推論の領域において世界最高峰の地位を確立している[8]。

同様に、AlibabaのQwen 3.5-397B-A17Bは、Gated Delta Networks（線形アテンション）と疎なMixture-of-Expertsを融合させたハイブリッドアーキテクチャを活用し、視覚と言語の早期融合（Early fusion）トレーニングを行っている[11]。このモデルは、大学院レベルの高度な科学知識を問うGPQA Diamondにおいて88.4%を記録し、LiveCodeBench v6では83.6%というトップクラスの成績を収めている[12]。さらに、プログラミングのベンチマークであるSWE-bench Verifiedにおいても、Claude Opus 4.6の80.6%に迫る76.4%を達成している[1]。

これらの卓越したスコアは、モデルが膨大な事前学習データから獲得したパターン認識能力と知識の蓄積、すなわち心理学における「結晶性知能（Crystallized Intelligence）」の極致を示している[14]。MMLU（Massive Multitask Language Understanding）やGPQA、あるいは高度なコーディングテストといったベンチマークは、表面的には高度な論理的推論を要求しているように見えるが、本質的には訓練データ内に存在する膨大な概念の補間や、テキストベースの論理パターンの再現、確立されたアルゴリズムの適用によって解決可能である[15]。中国製LLMは、大規模な強化学習を用いたポストトレーニングを通じて、これらの既存パターンの抽出と適用プロセスを極限まで最適化することに成功したのである[18]。

中国製LLMにおけるベンチマークスコアの極端な乖離



QwenおよびDeepSeekのベースモデルは、大学院レベルの知識（GPQA）や高度な数学（AIME 2025）で極めて高いスコアを記録するが、未知の抽象パターンの解決を要求するARC-AGI-2においては1%台に低迷している。

Data sources: Introl, Digital Applied, ARC Prize, Hugging Face

評価指標の飽和とデータ汚染による「推論の錯覚」

これらの高スコアが「真の未知の推論能力（True Out-of-Distribution Reasoning）」を反映しているかについては、学術界およびAI評価機関から強い疑義が呈されている[20]。LLMの性能評価において、訓練データにテストデータが混入する「データ汚染（Data Contamination）」や、特定のベンチマークに過剰適合させる「ベンチマーク・ゲーミング（Benchmark Gaming）」が深刻な構造的欠陥として浮上している[21]。

自然言語処理の研究分野では、n-gramの一致といった単純な文字列比較による従来の汚染検出フィルターでは検知できない「意味的重複（Semantic Duplicates）」によるソフトな汚染が広範に発生していることが指摘されている[26]。例えば、CodeForcesのプログラミング課題やAIMEの数学問題において、問題の表面的な文脈や変数を変更しただけの意味的重複データが訓練コーパスに大量に含まれている場合、モデルは未知の論理問題をその場で推論して解いているのではなく、事前学習や強化学習の過程で記憶した解法手順（アルゴリズムのテンプレート）を確率的に呼び出しているに

過ぎない[20]。

医学や法学の専門試験、あるいは大学院レベルの物理学を問うMMLUやGPQAなどの標準化されたテストは、事実の蓄積と既存のフレームワークの適用を評価するように設計されている[15]。中国製LLMが高得点を叩き出している分野は、まさにこの「パターン暗記と統計的な類似性検索」が極めて有効に機能する領域である[20]。人間の専門家が時間をかけて学ぶアルゴリズムを、モデルは膨大なパラメータ空間内に圧縮された知識表現として保持している。しかし、これは過去のデータに存在するパターンの高度な補間であり、人間のように新しい環境や全く未知の規則に対して適応する「推論」とは根本的に異なる認知プロセスである[20]。

ARC-AGI-2の設計思想と流動性知能の真の測定

この「推論の錯覚」を打破し、AIの真の汎用能力を測定するために、2025年3月にARC Prize FoundationとFrançois Cholletによってリリースされたのが「ARC-AGI-2」である[16]。ARC-AGIは、システムの「スキル獲得効率（Skill-acquisition efficiency）」、すなわち開発者が予期しなかった未知の問題にいかに適応し、事前知識なしで新しい法則を見つけ出せるかという「流動性知能（Fluid Intelligence）」を測定することを唯一の目的として設計されている[14]。

ARC-AGI-2の技術的特性とブルートフォースの排除

ARC-AGI-2のフォーマットは、1x1から最大30x30のグリッド上に最大10色のピクセルが配置された抽象的な視覚パズルである[16]。AIシステムは、数個（通常2〜3個）の入力・出力の例示ペアから、背後にある隠された変換ルール（Abstraction）を論理的に推論し、それを新しい入力グリッドに対して正確に適用しなければならない[16]。

このベンチマークがMMLUや数学の試験と決定的に異なるのは、以下の厳密な設計原則に基づく点である。

第一に、完全な事前学習の無効化である。ARC-AGI-2のすべてのタスクは完全にユニークであり、インターネット上のいかなる訓練データにも存在しない新規な論理パターンで構成されている[30]。したがって、意味的重複やパターンの暗記、あるいは検索拡張生成（RAG）によるスコアの底上げは完全に不可能に設定されている[16]。

第二に、コア知識プライア（Core Knowledge Priors）への限定である。タスクを解くために必要な知識は、Elizabeth Spelkeらの発達心理学で提唱されるような、オブジェクトの永続性、目標指向性、基本的な計数能力、位相幾何学的な概念（接続性、対称性、包含関係など）といった、人間が先天的に持つ、あるいは幼少期に言語を介さず獲得する認知のプリミティブのみに厳格に制限されている[14]。言語的知識や歴史的事実、プログラミング言語の構文といった専門知識は一切不要であるため、従来型ベンチマークにおける「結晶性知能」のアドバンテージは完全に無効化される[14]。

第三に、計算資源によるブルートフォース（総当たり攻撃）の排除と「知能の効率性」の測定である。ARC-AGI-2は、単純なプログラム探索（Program Search）や計算集約的な推測を防ぐように意図的に設計されている[16]。さらに、バージョン1からの重要な変更点として、タスクごとの「コスト（推論時の

計算量）」を追跡し、正解するだけでなく、いかにリソース効率よく解答を導き出せるかを評価対象としている[6]。知能とは単なる処理能力ではなく、限られた資源を用いて未知の法則をいかに効率的に獲得するかという適応力の指標であるという哲学が貫かれている[14]。

驚異的な人間とのスコア・ギャップと現在の到達点

ARC-AGI-2において最も特筆すべき事実は、人間と最新のAIモデルとの間に横たわる絶望的とも言えるスコアの乖離である。ARC Prize Foundationが厳密な管理下で実施したテストによれば、400人の非専門家の人間（特別な訓練を受けていない一般人）が1,417のユニークなタスクに挑み、すべてのタスクが「2回の試行以内で少なくとも2人の人間によって」解決された[16]。全体としての人間ベースラインは100%の解決率を誇り、1タスクあたりの平均所要時間はわずか2.3分であった[29]。これは、ARC-AGI-2が人間の流動性知能にとっては極めて直感的で容易な問題であることを証明している。

一方で、2026年2月時点における中国製オープンウェイトモデルのパフォーマンスは壊滅的である。Qwenチームのフラッグシップ推論モデルである Qwen3-235b-a22b InstructのARC-AGI-2における公式スコアはわずか1.3%にとどまっている[6]。同様に、DeepSeek R1（ベースモデル）のスコアも1.3%に過ぎない[6]。これは、ランダムな推測と大差ないレベルであり、未知の問題に対する適応力が根本的に欠如していることを示している。

システム / モデル	**ARC-AGI-2 スコア**	**GPQA Diamond**	**1タスクあたりの推論コスト**	**備考 / アプローチ**
人間（非専門家平均）	**100.0%**	N/A	N/A	平均2.3分で解決、直感的な空間・位相推論[29]
Google Gemini 3 Deep Think	84.6%	N/A	$13.62	大規模な並列推論、検証機関による確認済み[17]
Google Gemini 3.1 Pro	77.1%	94.3%	N/A	ネイティブ・マルチモーダル、System 2 統合[1]
Anthropic Claude Opus 4.6	68.8%	91.3%	N/A	高度な適応的推論、外部ツール活用[1]
OpenAI GPT-5.2	52.9%	92.4%	N/A	論理チェーンによる逐次思考プ

Thinking				ロセス[2]
DeepSeek R1 (Base CoT)	**1.3%**	71.5%	**$0.080**	トークン生成制約、テスト時適応機構の欠如[6]
Qwen3-235b-a22b Instruct	**1.3%**	N/A	**$0.004**	線形アテンション、空間把握能力の限界[6]

GoogleのGemini 3.1 Pro（77.1%）や、専用の推論モードであるGemini 3 Deep Think（84.6%）といった一部の最先端クローズドモデルは、人間には及ばないものの顕著な高得点を達成している[2]。しかし、これらのモデルは後述する莫大な「テスト時計算量（Test-Time Compute）」と高度に統合された推論パイプラインに依存しており、QwenやDeepSeekのような中国のオープンモデルが追求する標準的な推論プロセスとは根本的に異なるアプローチをとっているのである[17]。

中国製LLMがARC-AGI-2で破綻する構造的理由

中国製LLMが各種の学術ベンチマークで圧倒的な成績を収めながら、ARC-AGI-2で1%台に沈む理由は、単なる訓練データの不足やパラメータの規模の問題ではない。これは、現在のLLMが依拠する自己回帰型（Autoregressive）トランスフォーマーモデルが抱える根本的なアーキテクチャの限界と、テキスト生成のパラダイムそのものに起因する深い認知論的欠陥である。

1. シンボルグラウンディング問題と2Dグリッドの1次元化による空間コンテキストの喪失

ARC-AGI-2のタスクは、本質的に2次元の空間的、位相幾何学的なパズルである。タスクを解くためには、色のついたピクセルを「単なる数値や記号」としてではなく、「連続したオブジェクト」「境界線」「内側と外側」「対称性」といった物理的・空間的な意味を持つ実体として解釈する必要がある[28]。これは認知科学において「シンボルグラウンディング問題（Symbol Grounding Problem）」と呼ばれる課題であり、言語モデルは現実世界の物理的法則を真に理解しているわけではなく、記号を記号として操作しているに過ぎない[28]。

人間は2次元のグリッドを見た瞬間、視覚野を通じて図形全体のゲシュタルト（全体性）を把握し、ピクセル間の位相幾何学的なつながりを瞬時に理解する。しかし、Qwen 3.5やDeepSeek V3.2のような自己回帰モデルにとって、2Dグリッドは一次元のトークン列（文字列）に平坦化（Flattening）されて入力される[34]。トランスフォーマーの自己注意機構（Self-Attention）は、シーケンス内の各トークン間の数学的関係性を計算することはできるが、2次元マトリックスを1次元にシリアライズする過程で、上下左右の物理的な「隣接性」や「閉じた空間」といった位相幾何学的なコンテキストは破壊されてしまう。

歴史的なデータから明らかになっているように、人間が直感的に「閉じた図形の中を別の色で塗りつぶす」と理解できるタスクであっても、LLMはそれを「行と列のインデックスに基づく複雑な数値的関

係性の計算」として処理しなければならない[34]。研究報告によれば、最先端の推論モデルであっても、シンボルに視覚的パターンを超えた意味を割り当てる「シンボリック解釈（Symbolic Interpretation）」において著しく失敗する[31]。モデルは、オブジェクト間の接続性や対称性を数値的にチェックしようと試みるが、空間的な相対配置や位相幾何学的な構造（トポロジー）を真に理解していないため、些細な位置のズレや文脈の変化で決定的なエラーを犯し、無効なパスを生成し続ける[31]。

具体的には、LLMは「心的回転（Mental folding）」や「抽象的なマップのナビゲーション」といった空間推論能力を自発的に獲得（創発）しておらず、テキストベースの統計的なピクセル予測に終始している[34]。中国製モデルは、論理展開が一次元のテキストストリームに沿って進行する数学的方程式（AIMEなど）のトークン生成においては極めて優秀であるが、多次元的な状況把握が必要なARCのような視覚的抽象化タスクでは、そのアーキテクチャの前提自体が足枷となり完全に破綻するのである[28]。

2. コンテキスト依存のルール適用と構成的推論の致命的限界

ARC-AGI-2のテクニカルレポートが指摘するもう一つの重大な欠陥は、AIシステムにおける「構成的推論（Compositional Reasoning）」と「コンテキストに依存したルール適用（Contextual Rule Application）」の失敗である[31]。

中国製LLMを含む多くのモデルは、単一のグローバルなルール（例えば「すべての青い四角を一番右に移動させる」）を抽出して適用することには長けている[31]。しかし、ARC-AGI-2のタスクの多くは、複数のルールが相互に作用する複雑な構成を持っている。例えば、「形Aが存在する場合はルールXを適用し、形Bが存在する場合はルールYを適用する」といった文脈依存の条件分岐や、ルールの適用順序が結果を左右する問題である。

自己回帰モデルは、表面的なパターンに固執する傾向があり、基礎となる選択原則（Selection principles）を深く理解せずに、訓練データでよく見た統計的にありふれたヒューリスティクス（例えば、単純な鏡面反射、色の置換、単純なカウント）に飛びついてしまう[31]。誤った仮説に一度アンカリングされると、後続のトークン生成はその誤ったコンテキストに引きずられ、軌道修正が極めて困難になる。これは「コンテキストの混乱（Context Confusion）」と呼ばれ、LLMが以前に生成した誤った推論チェーンのトークンに強い影響を受け、行き詰まる現象である[34]。

3. System 1 vs System 2 推論：自己回帰生成における探索の欠如

現在、AI業界のコンセンサスは、LLMの推論能力を根本的に向上させるためには、事前学習のスケールアップだけでなく、推論時（テスト時）により多くの計算リソースを割り当てる「テスト時計算量（Test-Time Compute: TTC）」の拡大と、テスト時適応（Test-time adaptation）が必要不可欠であるという点にある[17]。

ノーベル経済学賞受賞者のダニエル・カーネマンが提唱した人間の認知モデルにおける「System 1（直感的で高速な思考）」と「System 2（論理的で遅い、熟慮的な思考）」になぞらえれば、標準的なLLMの自己回帰的なトークン生成はSystem 1に該当する[37]。ARC-AGI-2の未知のパズルを解くためには、一発の直感で答えを出すのではなく、仮説を立て、状態空間を探索し、誤りを自己認識して

バックトラックするSystem 2の推論が必須である[37]。

中国製LLMのベースモデルや標準的な指示チューニングモデル（Qwen 3.5やDeepSeek R1）は、Chain-of-Thought（CoT）プロンプティングを通じて推論ステップをテキストとして出力する機能は持っているものの、内部的に複数の推論パスを並行して評価し、動的に探索空間を剪定するような本格的なテスト時適応機構を備えていない[6]。これらのモデルは、与えられた入力に対して決定論的、あるいは単純な確率的サンプリングに基づく一本道のトークン生成（Single-shot inference）を行っているに過ぎず、ARC-AGI-2の広大な探索空間の中で容易に迷子になってしまうのである[6]。

テスト時計算量と知能効率のジレンマ：中国モデルの戦略的制約

中国製LLMがARC-AGI-2で高得点を獲得できない最大の要因の一つは、彼らのモデルアーキテクチャと開発戦略が、ARC-AGI-2を解くために必要な「膨大な推論コスト」と真っ向から対立している点にある。

莫大なコストを要求するSystem 2推論

GoogleのGemini 3 Deep ThinkがARC-AGI-2で84.6%という前人未到のスコアを叩き出した理由は、単なるモデルの賢さではなく、並列推論（Parallel reasoning）システムを活用し、複数の仮説を同時に評価、修正、マージする膨大なテスト時計算量を投入しているからである[17]。このプロセスの結果、Gemini 3 Deep Thinkの1タスクあたりの推論コストは$13.62という法外な金額に達している[17]。これは、ARC-AGIの1つのパズルを解くために、一般的なAPIコールの数万倍の計算資源を消費していることを意味する。

中国製LLMが追求する「極限のコストパフォーマンス」

対照的に、中国のAI戦略の中核は、オープンウェイトモデルの提供と、推論コストの劇的な削減によるAIの民主化・実用化にある。DeepSeek V3.2は、100万入力トークンあたりわずか0.28という破格の安さを実現しており、$GPT-5.2$の数十分の一のコストで運用可能である[2, 8, 42]。$Alibaba$の$Qwen3.5$も同様に、100万トークンあたり約0.18という極めて高いコスト効率を誇り、8倍の処理スループットを実現している[13]。

この極限の効率性を達成するために、中国製モデルはスパースなMoEアーキテクチャや、DeepSeek Sparse Attention（DSA）のような革新的なアテンション機構を採用し、アクティブなパラメータ数と計算の複雑さを大幅に削減している[8]。しかし、この「推論時の計算量を最小化する」というアーキテクチャの方向性は、ARC-AGI-2の複雑なパズルを解くために必要な「推論時の計算量を最大化して探索空間を広げる」というアプローチと完全に逆行しているのである。

表に示されたARC-AGI-2のデータがこの事実を如実に物語っている。Qwen3-235b-a22b Instructの1タスクあたりのコストはわずか0.004、$DeepseekR1$は0.080である[6]。中国製モデルは、推論コストをGemini Deep Thinkの1000分の1から3000分の1に抑えた結果、ARC-AGI-2のスコアも1.3%にとどまっている。彼らのモデルは、APIとしての実用性やローカル環境での展開を前提として

いるため、一つの問題に対して数分から数十分の計算リソースを占有するような探索的推論（Tree of Thoughtsや自己検証の反復）をデフォルトでは実行できない設計になっているのである[19]。

François Chollet が定義する「知能」は、単なるタスクの達成度ではなく、費やしたリソースに対する「効率（Intelligence Efficiency）」を含む[6]。Gemini 3 Deep Thinkがスコアで圧倒している一方で、そのリソース消費量は膨大であり、真の意味で効率的な知能を獲得しているとは言い難い。逆に、中国製モデルはリソース効率では究極の域に達しているが、流動性知能そのものが不足しているというジレンマが存在する。

DeepSeek V3.2-Specialeの例外と示唆

中国のAI研究所がこの問題に無自覚なわけではない。DeepSeekは、この制約を一時的に取り払い、オープンモデルの推論能力の限界を押し広げるために「DeepSeek-V3.2-Speciale」という高計算量バリアントをリリースした[44]。

このSpecialeバリアントは、出力トークン長や推論時間の制約を緩和し、大規模な強化学習（RL）プロトコルによって最適化された深い推論を可能にしている[44]。その結果、数学オリンピック（IMO 2025）や情報オリンピック（IOI 2025）において金メダルレベルのパフォーマンスを達成し、Gemini 3.0 Proに匹敵する推論能力を示したと報告されている[43]。

しかし、DeepSeek-V3.2-Specialeが強みを発揮するのは、数学や競技プログラミングのような「厳密なルールが存在し、検証関数が明確なドメイン」である。ARC-AGI-2のような「ルール自体を数個の例から帰納的に発見しなければならない未知の抽象ドメイン」においては、探索空間が無限に近く、検証関数をモデル自身が構築しなければならないため、単純に推論チェーン（CoT）を長くするだけでは正解に辿り着くことは困難である[34]。実際、Specialeバリアントは高コストを許容するAPI専用の実験的モデルとして位置づけられており、中国製モデルが日常的に提供する「低コスト・高効率」という基本価値からは逸脱している[45]。

結論：流動性知能の壁と汎用人工知能（AGI）への道筋

2026年2月の時点で、中国製LLMは驚異的な技術的洗練を達成しており、MMLUやGPQA、プログラミングベンチマークにおいて欧米のフロンティアモデルと互角以上の戦いを繰り広げている。疎なMoEアーキテクチャ、線形アテンション、そして革新的なスパースアテンション（DSA）の導入により、限られたパラメータ数と計算リソースで最大限の「結晶性知能」を引き出し、驚異的な低価格でAPIを提供する技術において、中国のAIエコシステムは間違いなく世界をリードしている。

しかし、ARC-AGI-2のスコアが容赦なく浮き彫りにしたように、現在のLLMのパラダイムは、統計的なパターンマッチングと学習データの高度な補間の域を出ていない。彼らは、過去の人類の知識を圧縮した巨大なデータベースとしては完璧に機能するが、物理世界に根ざした「流動性知能」や、未知のルールを帰納的に推論する空間的・位相幾何学的な理解力を根本的に欠いているのである。

中国製モデルがARC-AGI-2で高得点を獲得できない理由は、決して技術力の劣後ではない。それは、自己回帰トランスフォーマーというアーキテクチャが1次元のシーケンス予測に最適化されていること、そして彼らのビジネス戦略が「テスト時計算量を最小化することによる極限の効率化」を志向し

ていることの必然的な結果である。

今後、AI業界が真の汎用人工知能（AGI）に向けてこの「推論の壁」を突破するためには、単にパラメータを増やしたり、訓練データに合成データを追加したりする既存のスケール則（Scaling Laws）の延長だけでは不十分である。2次元・3次元の空間的関係性やグラフ構造をネイティブに保持・処理できるニューロシンボリック（Neuro-symbolic）なアーキテクチャへの移行や、莫大なテスト時計算量を消費することなく、モデル内部で効率的に仮説の生成・検証・バックトラックを実行できる新たなSystem 2推論機構の確立が求められる。

ARC-AGI-2における低スコアは、中国製LLMの敗北を意味するものではない。むしろ、これまでのベンチマーク・ゲーミングによって隠蔽されていた「言語モデルが本質的に理解できない領域」を明確に可視化し、次世代のAI研究が向かうべき新たなフロンティアを正確に指し示しているのである。人間の非専門家がわずか2分で解ける直感的なパズルに、最先端のスーパーコンピュータが屈する現状こそが、我々がAGIに到達するまでに越えなければならない最大の試練であることを証明している。

引用文献

1. Benchmark Wars 2026: ARC-AGI-2, GPQA Diamond, and the HLE ..., 2月 22, 2026にアクセス、https://exzilcalanza.info/ai-benchmark-wars-2026-arc-agi-gpqa-hle-controversy/
2. Best AI Models 2026: Gemini vs Claude vs GPT Compared ..., 2月 22, 2026にアクセス、https://serenitiesai.com/articles/best-ai-models-comparison-2026
3. OpenAI vs DeepSeek vs Qwen: The Ultimate AI Battle of 2025, 2月 22, 2026にアクセス、https://medium.com/@shivashanker7337/openai-vs-deepseek-vs-qwen-the-ultimate-ai-battle-of-2025-a6e7c1c9c008
4. Qwen 2.5-Max outperforms DeepSeek V3 in some benchmarks, 2月 22, 2026にアクセス、https://www.artificialintelligence-news.com/news/qwen-2-5-max-outperforms-deepseek-v3-some-benchmarks/
5. DeepSeek V3 vs Qwen3 Max Benchmarks: Coding, Math ..., 2月 22, 2026にアクセス、https://spectrumailab.com/blog/deepseek-v4-vs-qwen3-max-thinking-chinese-ai-models-beating-gpt5
6. Leaderboard - ARC Prize, 2月 22, 2026にアクセス、https://arcprize.org/leaderboard
7. A 2025 Comparison of DeepSeek R1, Qwen 2.5 and Claude 3.7, 2月 22, 2026にアクセス、https://www.preprints.org/frontend/manuscript/f194c9e503cef47e350e0d2aa2305423/download_pub
8. DeepSeek-V3.2 Matches GPT-5 at 10x Lower Cost | Introl Blog, 2月 22, 2026にアクセス、https://introl.com/blog/deepseek-v3-2-open-source-ai-cost-advantage
9. DeepSeek V3.2 Vs ChatGPT 5.1 Vs Gemini 3 Pro - AceCloud, 2月 22, 2026にアクセス、https://acecloud.ai/blog/deepseek-v3-2-vs-chatgpt-5-1-vs-gemini-3-pro/
10. deepseek-ai/DeepSeek-V3 - GitHub, 2月 22, 2026にアクセス、

https://github.com/deepseek-ai/DeepSeek-V3
11. qwen3.5-397b-a17b Model by Qwen | NVIDIA NIM, 2月 22, 2026にアクセス、https://build.nvidia.com/qwen/qwen3.5-397b-a17b/modelcard
12. Qwen/Qwen3.5-397B-A17B · Hugging Face, 2月 22, 2026にアクセス、https://huggingface.co/Qwen/Qwen3.5-397B-A17B
13. Qwen 3.5: 397B MoE Benchmarks, Pricing & Complete Guide, 2月 22, 2026にアクセス、https://www.digitalapplied.com/blog/qwen-3-5-agentic-ai-benchmarks-guide
14. What is ARC-AGI? - ARC Prize, 2月 22, 2026にアクセス、https://arcprize.org/arc-agi
15. What is MMLU? LLM Benchmark Explained and Why It Matters, 2月 22, 2026にアクセス、https://www.datacamp.com/blog/what-is-mmlu
16. Is Your AI Smart Enough? Test It with ARC AGI v2! - Labellerr, 2月 22, 2026にアクセス、https://www.labellerr.com/blog/arc-agi-v2/
17. Google Gemini 3 Deep Think Hits 84.6% on ARC-AGI-2, Beating ..., 2月 22, 2026にアクセス、https://www.implicator.ai/google-gemini-3-deep-think-hits-84-6-on-arc-agi-2-beating-gpt-5-and-claude-2/
18. DeepSeek-V3.2 Technical Report Is Pure Gold - The AI Timeline, 2月 22, 2026にアクセス、https://mail.bycloud.ai/p/deepseek-v3-2-technical-report-is-pure-gold
19. OpenAI o3 vs DeepSeek r1: An Analysis of Reasoning Models, 2月 22, 2026にアクセス、https://blog.promptlayer.com/openai-o3-vs-deepseek-r1-an-analysis-of-reasoning-models/
20. Why do LLM's perform so well on academic tests but so poorly on ..., 2月 22, 2026にアクセス、https://www.reddit.com/r/ArtificialInteligence/comments/1p87ado/why_do_llms_perform_so_well_on_academic_tests_but/
21. In Benchmarks We Trust ... Or Not?, 2月 22, 2026にアクセス、https://aclanthology.org/2025.emnlp-main.1208.pdf
22. The state of open-source LLMs as of summer 2024, or - The Motte, 2月 22, 2026にアクセス、https://www.themotte.org/post/1075/the-state-of-opensource-llms-as
23. Benchmarking Large Language Models Under Data Contamination, 2月 22, 2026にアクセス、https://aclanthology.org/2025.emnlp-main.511/
24. GitHub - SeekingDream/Static-to-Dynamic-LLMEval: The official, 2月 22, 2026にアクセス、https://github.com/SeekingDream/Static-to-Dynamic-LLMEval
25. Foundation Models in Agriculture: A Comprehensive Review, 2月 22, 2026にアクセス、https://pdfs.semanticscholar.org/7421/75cedc929ba9e5ecaf86dadf9a746c33808e.pdf
26. Soft Contamination Means Benchmarks Test Shallow Generalization, 2月 22, 2026にアクセス、https://arxiv.org/html/2602.12413v1
27. LLM benchmarks in 2025: What they prove and what your business, 2月 22, 2026にアクセス、https://www.lxt.ai/blog/llm-benchmarks/

28. Why Might The LLM Market Not Achieve AGI? - Aire AI App-Builder, 2月 22, 2026にアクセス、https://aireapps.com/articles/why-might-the-llm-market-not-achieve-agi/
29. GPT-5.2 & ARC-AGI-2: A Benchmark Analysis of AI Reasoning, 2月 22, 2026にアクセス、https://intuitionlabs.ai/articles/gpt-5-2-arc-agi-2-benchmark
30. ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems, 2月 22, 2026にアクセス、https://arxiv.org/html/2505.11831v2
31. ARC-AGI-2 A New Challenge for Frontier AI Reasoning Systems, 2月 22, 2026にアクセス、https://arcprize.org/blog/arc-agi-2-technical-report
32. Google's new AI model with double the reasoning power - Xpert.Digital, 2月 22, 2026にアクセス、https://xpert.digital/en/google-gemini-3.1-pro/
33. SPaRC: A Spatial Pathfinding Reasoning Challenge - ACL Anthology, 2月 22, 2026にアクセス、https://aclanthology.org/2025.emnlp-main.526.pdf
34. Building an ARC-2 Solver: My Multi-Agent Socratic Reasoning, 2月 22, 2026にアクセス、https://pub.towardsai.net/building-an-arc-2-solver-my-multi-agent-socratic-reasoning-journey-305c081611c6
35. Language Models and Spatial Reasoning: What's Good, What Is Still, 2月 22, 2026にアクセス、https://towardsdatascience.com/language-models-and-spatial-reasoning-whats-good-what-is-still-terrible-and-what-is-improving-175d2099eb4c/
36. François Chollet: How We Get To AGI (Transcript) - The Singju Post, 2月 22, 2026にアクセス、https://singjupost.com/francois-chollet-how-we-get-to-agi-transcript/
37. Evolutionary System 2 Reasoning: An Empirical Proof - arXiv.org, 2月 22, 2026にアクセス、https://arxiv.org/pdf/2512.05760
38. We tested every major AI reasoning system. There is no clear winner., 2月 22, 2026にアクセス、https://arcprize.org/blog/which-ai-reasoning-model-is-best
39. Notable AI Models - Epoch AI, 2月 22, 2026にアクセス、https://epoch.ai/data/notable_ai_models.csv
40. DeepSeek R1 vs OpenAI o3: Ultimate 2026 Reasoning Model, 2月 22, 2026にアクセス、https://www.humai.blog/deepseek-r1-vs-openai-o3-ultimate-2026-reasoning-model-comparison/
41. Gemini 3 Deep Think: A Comprehensive Analysis of G... - UniFuncs, 2月 22, 2026にアクセス、https://unifuncs.com/s/0UFRPUuR
42. DeepSeek V3.2: Open-Source Reasoning at Gold Medal Level, 2月 22, 2026にアクセス、https://www.e2enetworks.com/blog/deepseek-v3-2-open-source-reasoning
43. DeepSeek-V3.2: Pushing the Frontier of Open Large Language, 2月 22, 2026にアクセス、https://arxiv.org/html/2512.02556v1
44. DeepSeek-V3.2-Speciale: Open-Weights Reasoning Close To The, 2月 22, 2026にアクセス、https://medium.com/@leucopsis/deepseek-v3-2-speciale-open-weights-reasoning-close-to-the-frontier-models-d43cd5da22d9

45. DeepSeek-V3.2 Release, 2月 22, 2026にアクセス、
https://api-docs.deepseek.com/news/news251201