ChatGPT

中国製LLMが多くのベンチマークで高得点でもARC-AGI-2で低迷しやすい理由に関する調査報告書

エグゼクティブサマリー

本報告書は、**2026年2月20日**時点までに公開された一次情報（主に ARC Prize [1] の公式ドキュメント／技術報告、主要ベンチマークの原著論文、主要な中国発LLMの技術報告・モデルカードなど）を中心に、「中国製LLMが MMLU・C-Eval・HumanEval 等で高得点を取りやすい一方、ARC-AGI-2では相対的に低得点（＝人間平均・最上位モデルに比べ大きなギャップ）になりやすい」背景を、テスト設計・評価方法・モデル能力/訓練・評価実務・リーク/過学習・計測アーティファクトの観点から分析する。 [2]

結論を要約すると、ARC-AGI-2は（i）**知識依存を極小化**した視覚的・抽象的な「規則学習→一般化」課題で、（ii）**出力が厳密一致（ピクセル完全一致）**で部分点がなく、（iii）**pass@2（2案提出）**という限定的な探索余地しか与えず、（iv）さらに（Kaggle等の設定では）**計算資源・時間・外部ツール**に強い制約がかかるため、一般的な言語・知識ベンチマークで強いモデルでも、その強み（膨大なコーパス/試験問題形式への最適化、長文説明、暗黙知、言語推論のパターン学習）が点数に直結しにくい。 [3]

一方で、ARC-AGI-2でも2025年末〜2026年初頭にかけて「**テスト時学習（test-time training）**」「プログラム合成」「検証・自己修正ループ」等を強化したシステムが伸び、2026年2月20日時点では（報道ベースではあるが）ARC-AGI-2で**77.1%**の検証済みスコアが言及されている（中国製モデルの同等の公式公開スコアは、一次情報では相対的に乏しい/未確認が多い）。 [4]

したがって「中国製LLMがARC-AGI-2で伸びにくい」要因は、国籍というよりも、**（A）ARC-AGI-2自体が要求する帰納バイアス（2D操作・プログラム誘導・短い例からの規則抽出）と、（B）一般的LLMが得意な“言語知識ベンチ最適化”のズレ、（C）評価ハーネス/プロンプト/温度/ツール可否など実務設定の差**が主因である可能性が高い。改善には、表現（encoding）・探索（search）・検証（verification）・適応（adaptation）を明示的に組み込んだ実験が必要になる。 [5]

現象の定義と、2026年2月20日までに確認できた実証的状況

本稿でいう「中国製LLM」は、主に中国に拠点を置く企業・研究組織が開発・公開した大規模言語モデル群（例：アリババ（Alibaba） [6] 系のQwen、DeepSeek [7]、智譜AI（Zhipu AI） [8] 系のGLM、Moonshot AI [9] 系のKimi、MiniMax [10] 等）を指す。 [11]

「多くのベンチマークで高得点」の根拠として、中国発モデルの技術報告には、MMLU・HumanEval・GSM8K・BBH等での高スコアが明示されている。たとえばQwen2技術報告は、Qwen2-72B（base）で **MMLU 84.2 / HumanEval 64.6 / GSM8K 89.5 / BBH 82.4** 等を報告している。 [12] またQwen2.5の公式ブログは、Qwen2比で **MMLU 85+・HumanEval 85+・MATH 80+** といった改善を述べている（ただし“+”表記のため厳密値は未指定）。 [13] さらにKimi K2論文は、LiveCodeBench v6、AIME 2025、GPQA-Diamond、SWE-Bench Verified等で強い成績を示すと述べる。 [14]

一方ARC-AGI-2は、2025年3月24日に公開・アナウンスされ、公開時点では「純粋なLLM（単発推論）は0%」「公開済み推論システムでも一桁%」と説明されている。 [15] また公開直後の分析記事では「主要推論モデルをテストしたが、（当時の）最良でもごく低い」旨が述べられている。 [16] その後、2025年末のARC

Prize 2025結果分析では、Kaggle制約下の上位解法が **~24%**（private eval）に到達したことが報告される。[17]

2026年2月20日までの範囲で、ARC-AGI-2の最上位スコアは（少なくとも報道上）大きく伸びている。たとえば2026年2月20日付の報道は、ARC-AGI-2で**77.1%**の検証済みスコアに言及している。[18] ただし、中国製モデルが同一の「ARC Prize Verified」枠組み・同一セット（semi-private / private）・同一条件で得たスコアが、一次情報として体系的に公開されている例は、少なくとも本調査で確認した一次資料からは限定的であり、「中国製が低迷」と断言するには、**公開スコアの欠落（未公開/未確認）**という観測上の制約が残る。[19]

この“欠落”自体が重要で、ARC-AGI-2は（後述のとおり）評価ハーネス・計算資源・ツール有無・リーク対策等がスコアに強く影響し、ベンチマーク数値の横比較が難しい。そのため本報告書は、「中国製LLMが弱い」という単純な能力差仮説ではなく、**ベンチマーク設計と評価実務の相互作用**として説明・検証可能な仮説群を提示する。[20]

ARC-AGI-2のテスト設計、タスク種、スコアリング、限界

ARC-AGIは、フランソワ・ショレ（François Chollet）[21] の「On the Measure of Intelligence」で提示された、抽象推論・一般化能力（いわゆる流動性知能に近い概念）を測ることを狙ったコーパス（ARC）を土台にする。[22] ARC-AGI-2は2025年に導入された新バージョンで、形式自体はARC-AGI-1と同一（少数の入出力例→テスト入力に対する出力を生成）だが、AIにとって難しく、人間には比較的容易であるよう調整されている。[23]

ARC-AGIタスクは、数個（典型的に3〜5組）のtrain入出力ペアと、1〜3個のtest入力を含み、各タスクは整数（色）からなる2Dグリッドとして表現される。出力は「色と位置が完全一致」する必要がある。[24]

ARC-AGI-2のデータ分割は、少なくとも「Public Train / Public Eval / Semi-Private Eval / Private Eval」が明示され、Semi-Privateは限定的に第三者へ露出した可能性がある一方、Privateは第三者非露出を想定すると説明される。[25] またARC-AGI-2では難易度校正が導入され、Public Eval/ Semi-Private/ Privateが人間・AIの成績で概ね同程度（<1pp）になるよう調整するとされる。[26]（この校正が成功しているか、どの程度頑健かは、タスクの有限標本性や評価条件差に依存するため、後述の「計測アーティファクト」として注意が必要である。）

スコアリング（重要）は、ARC Prizeのガイドにより、各test出力に対して**2つの候補出力（attempt_1/attempt_2）**を提出し、どちらかが完全一致すればそのtest出力は1点、両方外れれば0点、最終スコアは全test出力にわたる平均（タスクに複数test入力がある場合はそれらも含め平均）であると定義される。[24] ここでは部分点がなく、わずかな形状ミスや色ミスで0点になるため、生成モデルの「だいたい合う」能力は点数に反映されにくい。[27]

ARC-AGI-2の人間側の性質として、ARC Prizeは「ARC-AGI-2の全タスクが少なくとも2人により2試行以内（pass@2）で解ける」旨を述べ、平均的参加者の平均を60%とする説明もある。[28] この“人間容易”は、ARC-AGI-2が知識よりも抽象規則の発見・一般化を狙っていることと整合するが、裏返すと「モデルが人間のような帰納バイアスを持たない限り点が伸びない」構造になりやすい。[29]

既知の限界として、ARC Prize自身が「ベンチマークは不完全」であり、また理想的なAGIの“リトマス試験”ではないことを繰り返し示唆している（例：グランドプライズを85%に設定しつつ、ベンチマークの不完全性も明記）。[30] さらに、Semi-Privateが限定的に露出し得る点は、「リーク/汚染の完全排除が構造的に難しい」ことを示す。[31]

ARC-AGI-2と、上位を取りやすい他ベンチマークの構造差

ARC-AGI-2で苦戦しても、MMLU・C-Eval・HumanEval等で高得点を取り得るのは、測っているもの（能力因子）と、問題形式・採点関数が大きく違うためである。以下では依頼の観点（「ARC-AGI-2の差分」）を明示するため、代表ベンチを構造比較する。 32

ベンチマーク設計の比較表

ベンチマーク	入力と出力の形式	主に測る要素	採点の性質	リーク/最適化が起きやすい要因
ARC-AGI-2	少数の2Dグリッド入出力例→**2Dグリッドを生成**	抽象規則の発見、2D操作の合成、少数例からの一般化	**完全一致（0/1）**、pass@2、部分点なし	Public部分は学習・チューニング可能。Semi-Privateは限定露出の可能性が明示されるため、完全な非汚染は難しい。 33
MMLU	多分野の**選択式（多肢選択）**問題	幅広い知識＋言語的推論	1問1点（選択肢一致）で比較的安定	公開試験問題由来で、学習混入や“形式慣れ”が起きやすい。 34
C-Eval	中国語中心の**多肢選択**（難易度層あり）	中国語文脈での知識＋推論	選択肢一致。CoT設定などで差が出る	中国語試験データに近く、学習・最適化・暗記の影響を受けやすい。 35
HumanEval	関数仕様（docstring等）→**コード生成**	プログラム合成/アルゴリズム的推論	**pass@k**が一般的（複数サンプルで成功確率が増える）	有名問題の学習混入、ユニットテストへの“当て方”最適化が起きやすい。 36

この比較から、ARC-AGI-2が他ベンチと最も異なるのは、（i）言語知識の寄与が小さく、（ii）2Dスキル誘導（回転・反転・抽出・塗り替え等の合成）が中心で、（iii）生成物が厳格フォーマットかつ完全一致採点、（iv）pass@2でも探索余地が小さい、という点である。 37

この差は、中国製に限らず「**知識・言語推論で強いモデルほど有利なベンチ**」と「**帰納・プログラム誘導で強いシステムほど有利なARC系**」のミスマッチとして説明できる。 38

能力・訓練・アラインメントがARC-AGI-2に効く点と効かない点

依頼の観点（モデル能力、訓練データ/アラインメント）に沿って、ARC-AGI-2に効きやすい要素を「効く理由」「効きにくい理由（落とし穴）」として整理する。 39

推論（長い思考）とチェーン・オブ・ソート

ARC Prizeのガイドは、単純な「Direct LLM Prompting」は<5%程度と述べ、さらに合成データで大量学習しても~10%程度という言及を置く（少なくとも当時のKaggle文脈）。 30　これは、**“思考の長さ”だけでなく、2D規則の探索・検証を含むアルゴリズム構造**が必要であること（＝出力がピクセル完全一致のため、言語的な“それっぽい説明”が点に結びつかない）を示唆する。 33

一方、2025年末のARC Prize 2025上位解法の記述は、テスト時学習や自己改良・反復精緻化（refinement loop）等、**探索と適応**を強く含む設計が伸びたことを示している。 40　したがって「推論が強い」中国製

LLMが他ベンチで高得点でも、ARC-AGI-2では“推論能力”の定義が違い、**探索・検証・手続き化**が不足すると伸びない、という説明が成立する。 41

ツール使用とプログラム合成

ARC-AGI系の高スコア解法は、離散的プログラム探索、DSL合成、アクティブ推論（test-time　fine-tuning）などの系譜が整理されており、LLM単体より「外部の探索器・検証器・実行器」を組み込む方向が重要視されてきた。 42

ただしKaggle本戦では「インターネット不可」等の制約が明記されるため、一般的な“Web検索ツールで補う”型のエージェント能力は直接には使えない。 43　ここで効くツールは、**ローカルで完結する探索（例：DSL探索、Python実行、候補解生成→検証）**である。中国製LLMがツール利用に強いとしても、それがWebエージェント的な強み（検索・外部知識）に偏っている場合、ARC-AGI-2の得点には直結しにくい。 44

長文コンテキストとマルチモーダル

ARC-AGI-2の各タスクは2Dグリッドが中心で、長文文脈を必要としない場合が多い。したがって、一般の長文ベンチで効く「超長文保持」や「大量文書からの検索」能力は、ARC-AGI-2では限界的になりやすい。 45

マルチモーダルについても、ARC-AGI-2は本質的には**2D離散記号操作**であり、画像として見るかJSONとして見るかは表現の問題である。ただし、モデル側に「2D空間の帰納バイアス（平行移動・回転・対称性・領域抽出）」が入っているかどうかはスコアに効き得る。 46　ここで、中国製LLMが“マルチモーダル＝自然画像理解”方向に最適化されていても、ARCで必要な「低次元・離散・構造」への適応が弱いと、効果は限定的になり得る（要検証）。 29

学習データとアラインメント

MMLUやC-Evalのような試験形式のデータは、公開済みであるがゆえに、訓練コーパス混入や“形式チューニング”が起きやすい。 47 対してARC-AGI-2は、Semi-Private/Privateセットの存在を明確にし、公開データとの重なりやショートカットを避ける意図が強い。 48

この構造上、「ベンチマーク上の“高得点”＝能力」になりにくい領域では、（中国製に限らず）ベンチ固有の過学習の恩恵が減るため、相対順位が入れ替わり得る。 49

評価セットアップ、リーク/過学習、計測アーティファクト

ここでは依頼の観点（評価設定、ローカリゼーション、リーク、計測アーティファクト）を、ARC-AGI-2の実務仕様に即して整理する。 50

評価セットアップ要因

ARC-AGI系の評価は、ハーネスの違い（promptテンプレ、出力フォーマット、リトライ、attempt数、温度、max　tokens等）で結果が変わる。ARC　Prizeのベンチマーク用リポジトリは、CLI引数として `--num_attempts`、温度、最大出力トークン、プロバイダ（API/ローカル）など多数の設定項目を明示している。 51　つまり「同じモデル名」でも、**プロンプト・温度・デコード戦略・attempt数**が揃っていないスコア比較は危険である。 52

さらに、ARC　Prizeのリーダーボードは「推論レベル（思考時間）」の違いで同一モデルの複数点が結ばれる、と説明されている。 53 これは、ARC-AGI-2が“推論時間を使うほど伸びるが逓減する”性質を想定していることを意味する。中国製LLMが他ベンチで強くても、ARC-AGI-2で「十分な推論時間」「適切な探索ルー

プ」を許容する設定になっていなければ、スコアが伸びない（あるいは評価されていない）可能性がある。[54]

ローカリゼーション（中国語プロンプト vs 英語プロンプト）については、ARCタスク自体が非言語的である一方、**指示文・フォーマット指定・例示**は言語で与えられるため、理論的には影響し得る。ARC Prizeのログ例では英語の指示文が用いられている。[55] ただし、影響の大きさは未確定であるため、後述の実験で検証すべき対象になる。

リーク、過学習、ベンチ特化チューニング

ARC-AGI-2はSemi-Private/Privateによりリークを抑える設計だが、Semi-Privateは限定露出の可能性が明記される（＝完全な非汚染ではない）。[56] またARC Prizeの年次分析は、汚染（knowledge-dependent overfitting）や“ベンチに合わせた過学習”への警戒を示す文脈を含む。[57]

逆に、MMLU/C-Eval/HumanEvalは公開であり、モデル開発者・利用者が評価セットを把握しやすい。HumanEvalはpass@kが有効であること自体が論文で議論されており、サンプリング回数を増やすと成功率が上がる性質がある。[58] この「繰り返し生成で上げる」戦略は、ARC-AGI-2でもpass@2という形では許されるが、kが2しかないため伸びしろが小さい。[24]

計測アーティファクト（閾値、部分点なし、時間/コスト、確率性）

ARC-AGI-2はピクセル完全一致で、部分点が標準スコアに入らない。ARC Prize自身も、研究用途として「pixel correctness」を参照する場合があるが、競技スコアは0/1であると明記する。[26] この仕様は、モデルが「規則の一部は捉えた」場合でも点に反映されず、改善が観測しづらい（＝学習シグナルが粗い）。[30]

さらに、ARC Prize 2025では評価コスト・計算制約が競技要件として明示され、Kaggleではrun-time上限やハードウェア条件（例：GPUノートブック12時間等）が規定される。[43] そのため、同じモデルでも「推論を長く回して当てる」設定は競技では不利になり得る。[59]

確率性（温度やサンプリング）も重要で、ARC Prizeのベンチハーネスは温度等を設定可能としている。[51] したがって、**温度1.0/0.0、top-p、思考トークン上限、attempt間の相関制御**などが、pass@2の実効性能を左右し得る（後述の実験で定量化可能）。

技術的説明仮説、検証実験、改善提案

ここでは依頼の観点（技術的説明と検証実験、実務的提案）を、反証可能な形で提示する。ARC-AGI-2は「LLM単体」より「探索＋検証＋適応」を含むシステムで伸びてきたという公式分析と整合するため、改善策もシステム化が中心になる。[44]

仮説セット

仮説A：表現（encoding）のミスマッチが主因
2Dグリッドをテキスト列として与えると、空間的帰納バイアスが弱いモデルほど規則抽出に失敗しやすい。視覚エンコーダや2D専用トークナイザ、あるいは2D操作を前提にした内部表現が不足していると、言語ベンチで強くてもARCで伸びない。[29]

仮説B：探索不足（pass@2に最適化されていない）
HumanEval等ではpass@kがスコアを押し上げるが、ARCはpass@2で探索幅が小さいため、候補生成の多様性と検証がより重要になる。複数候補を“独立”に出せていない（attempt_1とattempt_2が実質同じ）場合、スコアが伸びない。[60]

仮説C：検証器（verifier）欠如による“微小ミス→0点”問題
完全一致採点では、境界条件の1ピクセルミスが致命的。生成物を自動検査し、規則に反する箇所を修正するループがないと伸びにくい。 30

仮説D：評価実務（プロンプト・温度・フォーマット）が不利
ARCはフォーマット違反や出力不能が即0扱いになり得る。モデルが英語指示やJSON出力に得意でない／安全・アラインメントにより出力が冗長化する等で、正解規則を掴んでも点にならない。 61

仮説E：ベンチマーク特化で稼いだ“知識・試験最適化”がARCに転移しない
MMLU/C-Evalは多肢選択で知識寄与が大きい。ARCは知識をほぼ使わないため、スケーリングや試験最適化の利得が出にくい。 62

検証実験の設計

以下は「中国製LLMがARC-AGI-2で相対的に低い理由」を検証するための最小実験群である（データはPublic Evalを用い、semi-private/privateはアクセス可能な範囲に限定）。ARC Prizeが公開するスコア定義（pass@2、完全一致）に合わせる。 63

実験	変数	必要データ/環境	指標	期待される観測（仮説が正しければ）
表現アブレーション	グリッド表現（JSON直列 vs 画像化 vs 2Dトークン）	ARC-AGI-2 Public Eval、同一モデル	pass@2、フォーマット率、誤りタイプ分布	2D寄り表現で顕著改善（仮説A） 64
多様な2候補生成	attempt_2の生成法（温度差、探索差、自己反対仮説など）	同上	pass@2、attempt間の相関（重複率）	attempt間独立性↑でpass@2↑（仮説B） 65
検証・修正ループ	verifier有無（規則チェック、差分修正）	DSL/Python実行環境（Kaggle制約想定ならローカルのみ）	pass@2、1ピクセル誤り率	“ほぼ正解→完全正解”への変換が増える（仮説C） 66
プロンプト言語/指示の比較	英語指示 vs 中国語指示、最小指示 vs 詳細指示	同上	pass@2、フォーマット違反率	言語差が出るなら差分が統計的に有意（仮説D） 67
競技制約感度	推論時間/計算量/探索幅を段階的に制限	Kaggle制約相当の時間・計算上限	pass@2 と $/task（推定）	制約下で急落する方式は実用/競技で不利（仮説B/Cと整合） 68

評価フロー図（mermaid）

ARC Prizeのスコア定義・データ分割・ハーネス設定を抽象化すると、評価は次の流れで理解できる。 69

```
flowchart TD
  A[ARC-AGI-2 データセット選択<br/>public / semi-private / private] --> B[タスク整形<br/>グリッド表現・指示文・出力フォーマット]
```

```
  B --> C[モデル推論<br/>温度・max tokens・思考設定]
  C --> D[候補出力 2案<br/>attempt_1 / attempt_2]
  D --> E{完全一致判定<br/>pixel-perfect}
  E -->|一致| F[そのtest出力=1]
  E -->|不一致| G[そのtest出力=0]
  F --> H[全test出力で平均=最終スコア]
  G --> H
  C --> I[コスト/時間計測<br/>token課金・wall time]
  I --> J[スコア×コストで可視化<br/>リーダーボード]
  H --> J
```

重要イベントのタイムライン（mermaid）

ARC-AGI-2を巡る設計・バージョン・公開結果は、少なくとも以下の節目がスコア解釈に影響する。 70

```
timeline
  2019 : 「On the Measure of Intelligence」でARC提案
  2025-03-24 : ARC-AGI-2 公開・ARC Prize 2025 アナウンス
  2025-05-20 : ARC-AGI-2 概要報告（当時は低スコアが中心）
  2025-12-05 : ARC Prize 2025 結果分析（Kaggle制約下で~24%等）
  2026-02-20 : 報道上、ARC-AGI-2で77.1%の検証済みスコア言及
```

実務的な改善提案（研究者/エンジニア向け）

最後に、「中国製LLM（あるいは同様に一般ベンチで強いLLM）をARC-AGI-2で伸ばす」ための、優先度が高い実装上の提案をまとめる。これらは国籍非依存に有効だが、“既存の言語ベンチ最適化”とは別物の投資が必要になる。 71

第一に、**2D専用の探索・検証ループを外付け**する。具体的には、（a）候補規則をDSLプログラムとして表現し、（b）trainペアに一致するまで探索し、（c）test入力へ適用して出力を生成し、（d）2候補として多様な探索経路を残す。ARC Prizeのガイドが挙げるDSL合成やアクティブ推論の系譜に沿う。 30

第二に、**pass@2に最適化した多様化**を入れる。HumanEvalでpass@kが効くことは原著で明確であり、ARCでもpass@2の2回に“独立性”を確保する設計が鍵になる。温度差だけでなく、仮説空間（例：対称性優先/連結成分優先/色置換優先）を変えるなど、意図的な探索分岐が必要である。 65

第三に、**フォーマット堅牢化**を徹底する。ARCは出力不能や形式逸脱が致命的になり得るため、（a）出力を厳格JSONに限定させる、（b）生成後にパーサで必ず正規化し、（c）サイズ不整合や不正色があればルール推定段階に戻す、という“失敗前提”の実装が要る。ARC Prizeのベンチハーネスがリトライやスコアリングを実装していることからも、この層の重要性が示唆される。 72

第四に、**評価条件を公開・固定**し、再現可能な比較を行う。ARCは設定差が大きく、同一モデルでも温度・思考トークン上限・attempt生成法で結果が変わるため、（a）プロンプト、（b）デコード設定、（c）試行回数、（d）計算制約、（e）ツール可否、（f）中国語/英語指示、を実験ログとして残す必要がある。 73

第五に、**リークと“ベンチ特化”を分離して測る**。ARC-AGI-2はSemi-Private/Privateの概念を明示するため、Publicでの改善が、未知タスクへの一般化か、公開セット最適化かを分けて検証すべきである。特にSemi-Privateは露出可能性があると明示されるので、「最終的にはPrivate相当（非露出）で再検証」する設計が重要になる。 74

優先度付き参考情報源

一次情報（最優先）として、ARC-AGI-2の評価解釈はARC　Prize公式（データ分割・スコア定義・競技制約・年次分析）を軸にすべきである。[75]　MMLU/C-Eval/HumanEval/ARCの定義は原著論文（arXiv）を優先した。[76]　中国製LLMの“他ベンチで高得点”は、各モデルの技術報告・公式ブログ・論文（可能な限り一次）を採用した。[77]　2026年2月20日時点のARC-AGI-2高スコアの言及は報道に依存しているため、再現条件の一次資料が入手でき次第、検証が望まれる。[78]

---

[1] [2] [3] [5] [7] [9] [23] [24] [26] [27] [30] [33] [37] [41] [42] [43] [45] [61] [63] [68] [69] [71] [75] https://arcprize.org/guide
https://arcprize.org/guide

[4] [18] [78] https://venturebeat.com/technology/google-launches-gemini-3-1-pro-retaking-ai-crown-with-2x-reasoning
https://venturebeat.com/technology/google-launches-gemini-3-1-pro-retaking-ai-crown-with-2x-reasoning

[6] [20] [50] [51] [52] [67] [72] [73] https://github.com/arcprize/arc-agi-benchmarking
https://github.com/arcprize/arc-agi-benchmarking

[8] [21] [32] [34] [47] [62] [76] https://arxiv.org/abs/2009.03300
https://arxiv.org/abs/2009.03300

[10] [25] [28] [31] [56] [74] https://arcprize.org/arc-agi/2/
https://arcprize.org/arc-agi/2/

[11] [12] [77] https://arxiv.org/abs/2407.10671
https://arxiv.org/abs/2407.10671

[13] https://qwenlm.github.io/blog/qwen2.5/
https://qwenlm.github.io/blog/qwen2.5/

[14] https://arxiv.org/abs/2507.20534
https://arxiv.org/abs/2507.20534

[15] [38] [48] https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025
https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025

[16] https://arcprize.org/blog/analyzing-o3-with-arc-agi
https://arcprize.org/blog/analyzing-o3-with-arc-agi

[17] [39] [40] [44] [49] [57] [66] https://arcprize.org/blog/arc-prize-2025-results-analysis
https://arcprize.org/blog/arc-prize-2025-results-analysis

[19] https://arcprize.org/policy
https://arcprize.org/policy

[22] [29] [46] [64] https://arxiv.org/abs/1911.01547
https://arxiv.org/abs/1911.01547

[35] https://arxiv.org/abs/2305.08322
https://arxiv.org/abs/2305.08322

[36] [58] [60] https://arxiv.org/abs/2107.03374
https://arxiv.org/abs/2107.03374

53 54 59 https://arcprize.org/leaderboard
https://arcprize.org/leaderboard

55 https://huggingface.co/datasets/arcprize/arc_agi_v2_public_eval/raw/main/qwen3-235b-a22b-instruct-2507/2b83f449.json
https://huggingface.co/datasets/arcprize/arc_agi_v2_public_eval/raw/main/qwen3-235b-a22b-instruct-2507/2b83f449.json

65 https://arxiv.org/pdf/2107.03374
https://arxiv.org/pdf/2107.03374

70 https://raw.githubusercontent.com/arcprize/ARC-AGI-2/main/changelog.md
https://raw.githubusercontent.com/arcprize/ARC-AGI-2/main/changelog.md