ARC-AGI-3 における Claude Opus 5 の 30.2%

——その意味と、なぜここまで上がったのか

調査日：2026 年 7 月 25 日

Claude Opus 5

---

1. 要旨

2026 年 7 月 24 日に公開された Claude Opus 5 は、ARC-AGI-3 で **30.2%**を記録した。この数値は Anthropic のシステムカード§8.14.2 に由来し[6][11]、ARC Prize 財団の公式リーダーボードにも「ARC PRIZE | VERIFIED」バッジ付きで掲載されている[3]。すなわち第三者検証済みのスコアである。

ARC-AGI-3 は 2026 年 3 月のローンチ時点で全フロンティアモデルが 1%未満[7]、直前の最高値は GPT-5.6 Sol (Max) の 7.8%であった[3]。30.2%は前記録の約 4 倍にあたり、しかも測定コストはより低い。本稿では、この数値が何を意味し、なぜこれほど跳ね上がったのか、そして実用上どのような含意を持つのかを整理する。

2. ARC-AGI-3 とは何を測るベンチマークか

ARC-AGI-3 は、従来の ARC（静止した図形パズル）とは設計思想が異なる。エージェントを**説明書のないミニゲーム**に放り込み、目的もルールも明示しないまま実際にプレイさせる[1]。与えられるのは 5 つのキー操作と Undo、そして 64×64 グリッド上の座標クリックだけである[2]。操作系をあえて単純化することで、難しさが「操作の習熟」ではなく「論理の発見」に集中する構造になっている。

環境は全 135 個（公開 25／セミプライベート 55／プライベート 55）で構成され、各環境は少なくとも 6 レベルを持つ（公式採点には通常 5 レベルを使用）[2]。人間ベースラインは各環境につき 10 名の被験者を対面・統制環境で実測し、「手数における上位中央値の最良人間」と定義されている。収録された環境はいずれも人間が 100%解けることが確認済みである[2]。人間の環境あたり所要時間の中央値は 7.4 分であった。

3. 「30.2%」の正しい読み方

最も誤解されやすい点だが、ARC-AGI-3 のスコアは正解率ではない。採点指標は **RHAE**

（**Relative Human Action Efficiency／人間比の行動効率**）であり、各レベルについて次式で計算される[2]。

$$\text{スコア} = \min\left(1.15,\ \left(\frac{\text{人間の手数}}{\text{AI の手数}}\right)^2\right)$$

環境スコアはレベルに線形の重みを掛けて算出され（レベル 1 は 1/15、レベル 5 は 5/15 の寄与）、最終スコアは全環境の平均となる[2]。人間ベースラインが 100%である。

したがって 30.2%は「3 割の環境をクリアした」という意味ではない。比を二乗する指標なので、単純化して均一な効率を仮定すれば、**おおよそ人間の 1.8 倍の手数を要している状態**に相当する（$1 \div 1.8^2 \fallingdotseq 0.31$）。効率差が二乗で効くため数値は伸びにくく、逆に言えば数ポイントの差でも実際の探索効率の差はかなり大きい。

4. 公式リーダーボードにおける位置づけ

以下は ARC Prize 公式リーダーボード（ARC-AGI-3 タブ／Leaderboard Breakdown）の掲載内容である[3]。網掛け行が Claude Opus 5。

AI システム	開発元	公開日	種別	ARC-AGI-1	ARC-AGI-2	ARC-AGI-3	コスト/タスク	コスト(v3)
Claude Opus 5 (High)	Anthropic	2026-07-24	CoT	97.5%	88.3%	**30.2%**	$1.45	$20.7K
Claude Opus 5 (Max)	Anthropic	2026-07-24	CoT	97.5%	90.4%	N/A	$2.06	N/A
Grok 4.5 (High)	xAI	2026-07-16	CoT	85.7%	52.6%	0.3%	$0.780	$6.9K
Grok 4.5 (Medium)	xAI	2026-07-16	CoT	87.2%	52.6%	0.3%	$0.750	$8.5K
Grok 4.5 (Low)	xAI	2026-07-16	CoT	79.2%	33.1%	0.3%	$0.400	$8.7K
Inkling	Thinking Machines	2026-07-15	CoT	79.5%	36.5%	N/A	$0.642	N/A
GPT-5.6 Sol (Low)	OpenAI	2026-07-09	CoT	74.5%	42.5%	0.3%	$0.320	$12.8K
GPT-5.6 Sol (Medium)	OpenAI	2026-07-09	CoT	92.5%	67.1%	1.1%	$0.470	$13.0K
GPT-5.6 Sol (High)	OpenAI	2026-07-09	CoT	97.0%	85.4%	2.1%	$0.740	$15.2K
GPT-5.6 Sol (xHigh)	OpenAI	2026-07-09	CoT	97.5%	90.0%	7.0%	$1.04	$19.2K
GPT-5.6 Sol (Max)	OpenAI	2026-07-09	CoT	96.5%	92.5%	7.8%	$1.44	$25.1K
GPT-5.6 Terra (Low)	OpenAI	2026-07-09	CoT	60.2%	18.8%	0.0%	$0.170	$5.1K
GPT-5.6 Terra (Medium)	OpenAI	2026-07-09	CoT	77.0%	37.5%	0.1%	$0.270	$5.9K
GPT-5.6 Terra (High)	OpenAI	2026-07-09	CoT	92.0%	67.1%	0.5%	$0.550	$5.9K
GPT-5.6 Terra (xHigh)	OpenAI	2026-07-09	CoT	94.0%	74.2%	0.7%	$0.690	$6.8K

AI システム	開発元	公開日	種別	ARC-AGI-1	ARC-AGI-2	ARC-AGI-3	コスト/タスク	コスト(v3)
GPT-5.6 Terra (Max)	OpenAI	2026-07-09	CoT	96.5%	83.9%	0.8%	$1.09	$7.9K
GPT-5.6 Luna (Low)	OpenAI	2026-07-09	CoT	34.2%	5.1%	0.2%	$0.070	$2.3K

表 1：ARC Prize 公式リーダーボード ARC-AGI-3 タブ（2026 年 7 月 25 日時点、画面上部から 17 行分）。「コスト(v3)」は ARC-AGI-3 評価一式の実行費用。

5. なぜこれほど上がったのか

ARC Prize 財団は前世代（GPT-5.5／Opus 4.7）の挙動を分析し、フロンティアモデルに共通する 3 つの失敗モードを特定している[4]。これが裏返しの答えになる。

第一に、**局所的観察を世界モデルに統合できない**こと。「ACTION3 を押すとこの物体が回転する」までは把握できるのに、「回転は位置決めの手段だから、動かす前に向きを合わせるべきだ」という一段上の理解に繋がらない。

第二に、**訓練データからの誤った抽象化**。見た目が少し似ているだけで、テトリスや Sokoban、Frogger といった既知のゲームだと決めつけてしまう。財団の表現を借りれば「局所的な視覚的類似が、そのままゲーム全体の理論になってしまう」。誤った理論はその後の行動選択を丸ごと乗っ取る。

第三に、**理解しないままの勝利**。誤った仮説を抱えたままレベル 1 を 37 手でクリアしてしまい、その誤解が「正しかった」ものとして固定化し、レベル 2 以降で立て直せなくなる。

要するに欠けていたのは、観察を再利用可能な原理に**圧縮**する力、仮説を能動的に**反証・更新**する力、そして「なぜ今うまくいったのか」を確認する**メタ認知**である。Opus 5 の飛躍は、この「自分の理論を疑い続けるループ」が実用水準に達したことの現れと読むのが自然である。Anthropic が同モデルの特徴として挙げる「もっともらしい答えで止まらず、成功するまで作業を続ける」持続性[5][8]は、まさに ARC-AGI-3 が測る能力と同じ方向を向いている。

6. 表から読み取れる三つの論点

6.1 コスト効率――スコアが高く、かつ安い

GPT-5.6 Sol (Max) は 7.8%を得るのに$25.1K を要したのに対し、Opus 5 (High) は 30.2%を$20.7K で達成している[3]。スコアが約 4 倍でありながら費用は約 2 割低い。RHAE が手数の少なさを評価する指標である以上、これは整合的である。無駄な試行が減れば、スコアが上がる

と同時に実行コストも下がる。前節の「仮説更新ループ」という説明を、コスト面からも裏づける結果と言える。

6.2 effort 設定との非単調性——30.2%は上限ではない可能性

OpenAI 系は effort を上げるほど ARC-AGI-3 スコアが単調に伸びている（Sol: 0.3 → 1.1 → 2.1 → 7.0 → 7.8%）。一方 Opus 5 は Max 設定での ARC-AGI-3 測定が「N/A」となっており[3]、最高効率設定での値が存在しない。ARC-AGI-2 では High 88.3%に対し Max 90.4%と伸びていることを踏まえると、30.2%は Opus 5 の到達上限ではない可能性がある。なお報道では、ベンチマークによっては max 設定でわずかに悪化する例も指摘されており[9]、effort とスコアの関係は単純ではない。

6.3 v1・v2 との対比——強みの所在

ARC-AGI-1 では Opus 5 (97.5%) と GPT-5.6 Sol (xHigh) が同値であり、ARC-AGI-2 では Opus 5 (High) 88.3%に対し Sol (Max) が 92.5%と、むしろ OpenAI 側が上回る[3]。にもかかわらず v3 だけが 30.2%対 7.8%と桁で開く。これは、v1／v2 が「静的パズルの正解率」を測るのに対し、v3 が「未知環境での探索効率」という別種の能力を測っているためである。この 3 列の並びは、**Opus 5 の優位が静的推論そのものではなく、対話的な仮説検証ループにある**ことを明確に示している。

7. 実用上、何が良いのか

ARC-AGI-3 は事前学習の知識で押し切れないよう設計されている。環境は非公開かつ新規で、ルールは実行時にしか判明しないため、訓練データとのパターンマッチが機能しない[1][7]。ここでのスコアは「知っていること」の量ではなく、**未知の状況で仮説を立て、検証し、修正する速さ**——いわゆる流動的知能／スキル獲得効率——を反映する。

実務への含意は直接的である。ドキュメントの存在しない社内ツール、初見の UI、仕様書のないレガシーコード、想定外のレスポンスを返す API。こうした「誰も教えてくれない環境」でエージェントを動かす場面が、まさにこの能力に対応する。従来はプロンプトや足場（scaffolding）を人間が丁寧に設計して補っていた部分が、モデル側の地力で埋まるということであり、エージェント運用の初期設計コストを押し下げる効果が期待できる。

8. 留意点

第一に、**人間との差は依然として大きい**。人間ベースラインは全環境 100%であり、30.2%との間には 3 倍以上の隔たりがある[2]。

第二に、**指標の解釈に注意**が要る。RHAE は効率比の二乗であるため、30%を「人間の 3 割の能力」と読むのは誤りである（前掲§3）。

第三に、**第三者集計サイトの更新には遅れがある**。2026 年 7 月 25 日時点で llm-stats や BenchLM の ARC-AGI-3 リーダーボード集計には GPT-5.6 系のみが並び、Opus 5 は反映されていない[12]。ただしこれは集計サイト側のラグであって、ARC Prize 公式リーダーボードの状態とは別問題である。公式には検証済みとして掲載されている[3]。この二つを混同しないこと。

第四に、**ベンチマーク 1 本の飛躍が日常業務の性能向上に比例するとは限らない**。とはいえ、「訓練データで殴れない」よう設計されたテストで約 4 倍の差をつけたという事実は、他ベンチマークにおける数ポイント差とは質的に異なる指標だと評価してよい。

参考文献

[1] ARC Prize Foundation, “ARC-AGI-3”. https://arcprize.org/arc-agi/3

[2] ARC Prize Foundation, “ARC-AGI-3: A New Challenge for Frontier Agentic Intelligence”（技術報告書, PDF）. https://arcprize.org/media/ARC_AGI_3_Technical_Report.pdf

[3] ARC Prize Foundation, “Leaderboard”（ARC-AGI-3 タブ／Leaderboard Breakdown）. https://arcprize.org/leaderboard ／ 2026 年 7 月 25 日、ブラウザ表示のスクリーンショットにより内容を確認。

[4] ARC Prize Foundation, “Analyzing GPT-5.5 & Opus 4.7 with ARC-AGI-3”. https://arcprize.org/blog/arc-agi-3-gpt-5-5-opus-4-7-analysis

[5] Anthropic, “Introducing Claude Opus 5”（2026 年 7 月 24 日）. https://www.anthropic.com/news/claude-opus-5

[6] Anthropic, “System Card: Claude Opus 5”（2026 年 7 月 24 日, PDF），§8.14.2. https://www-cdn.anthropic.com/c5fbac3f0b1280a933ebd26d3cb8bb9f5bdeaf48/Claude%20Opus%205%20System%20Card.pdf

[7] DataCamp, “ARC-AGI-3: The New Interactive Reasoning Benchmark”. https://www.datacamp.com/blog/arc-agi-3

[8] DataCamp, “Claude Opus 5: Anthropic’s New Flagship, Explained”. https://www.datacamp.com/blog/claude-opus-5

[9] The Decoder, “Anthropic claims its new Claude Opus 5 delivers near-Fable 5 performance at half

the token price”. https://the-decoder.com/anthropic-claims-its-new-claude-opus-5-delivers-near-fable-5-performance-at-half-the-token-price/

[10] OfficeChai, “Claude Opus Scores 30% On ARC-AGI 3, Triples Previous Best Score By Any Model”. https://officechai.com/ai/claude-opus-5-arc-agi-3/

[11] BenchLM, “Claude Opus 5 Benchmarks, Pricing & Speed (July 2026)”. https://benchlm.ai/models/claude-opus-5

[12] llm-stats, “ARC-AGI-3 Leaderboard”. https://llm-stats.com/benchmarks/arc-agi-3

注：本文中の上付き番号は文末の参考文献番号に対応する。表 1 の数値は、閲覧者が取得した ARC Prize 公式リーダーボード画面のスクリーンショットから転記したものであり、画面上部から 17 行分を収録している（それ以降の行は画面外のため未収録）。