ChatGPT

Claude Opus 5のARC-AGI-3 30.2%をどう読むべきか

エグゼクティブサマリー

Anthropicは2026年7月24日にClaude Opus 5を公開し、本文では「ARC-AGI-3で次点モデルの約3倍」と説明しました。30.2%という絶対値は、Anthropic本文そのものではなく、ARC Prizeの公式X投稿で「Claude Opus 5 … 30.2%」として明示され、同時に「従来最高の7.8%はGPT-5.6 Sol (Max)」と位置づけられています。したがって、**30.2%という数値は、現時点ではAnthropic本文の比率表現とARC Prize公式投稿の双方で裏づけられる一方、Anthropicの本文テキスト単体には絶対値が出ていない**、というのがいちばん厳密な言い方です。 [1]

このスコアは大きな前進ですが、ARC-AGI-3の定義上、100%は「未知の環境を**人間並みの行動効率**で全て攻略すること」を意味します。ARC-AGI-3は、静的パズルではなく、未知環境の探索、世界モデルの形成、目標発見、長期計画を測るインタラクティブ・ベンチマークです。よって30.2%は、**未知環境への適応能力が従来のフロンティアモデルより質的に改善した**ことを示す強いシグナルですが、**人間級の汎用エージェント**を意味するにはまだ遠い数字です。 [2]

実務面では、この水準は「新しい社内ツール、変則的なUI、仕様が十分に文章化されていないワークフロー」に対する**初見適応力**の改善を示唆します。研究面では、ARC Prizeが以前にOpus 4.7やGPT-5.5で観察した「局所効果は見えるが世界モデルを作れない」「訓練データ由来の誤ったゲーム類推に引きずられる」「レベルは解けてもゲームを学べない」といった失敗モードを、どこまで減らしたのかを分析する好材料です。 [3]

ただし、**なぜ30.2%まで伸びたか**について、公開情報だけで断定できることは限られます。AnthropicはOpus 5の**1Mトークン文脈、128k出力、適応的 thinking、effort レベル、長時間エージェント作業向け最適化**は公開していますが、**パラメータ数、MoEかどうか、学習データ総量、データ混合比、合成データ比率**は公開資料からは確認できません。したがって、最も妥当な結論は、**基盤モデルの非公開な改良**に、**評価時思考量の拡大、長文脈保持、自己検証の改善**が重なった、というものです。 [4]

前提とエビデンス

ARC-AGI-3は、ARC Prize Foundationが2026年3月に公開した、ARC系列初の**インタラクティブ推論ベンチマーク**です。モデルは自然言語のルール説明を与えられず、未知の環境を自分で探索し、目標を見つけ、状態遷移を学び、行動計画を修正し続ける必要があります。100%は「全ゲームを人間と同等以上の行動効率で攻略した」ことを意味します。 [5]

ARC Prizeの技術報告は、ARC-AGI-3の中核能力を**探索、モデリング、目標設定、計画と実行**の四つに整理し、スコアを単なる正答率ではなく**行動効率**として設計しています。これは、総当たり的な試行錯誤を不利にし、未知環境での素早い仮説形成と修正を高く評価するためです。さらに2026年4月の人間データ公表時には、レベル単位の基準を「第2位人間」から「中央値人間」へ変更するなどのスコア更新が行われましたが、その影響は**人間・AIともに約+0.5ポイント**と説明されており、30.2%という大幅上昇を説明するには小さすぎます。 [6]

重要なのは、ARC Prize自身が**ベンチマーク特化ハーネスによる“かさ上げ”**を強く警戒していることです。公式技術報告は、公開環境に対するタスク特化や、ARC-AGI-3専用戦略・道具選択・人手で埋め込まれた指示がスコアを人為的に押し上げうると明記し、公式スコアでは**同一のシステムプロンプト**を用い、**外部ツールな**

しで、一般目的API越しのモデルを測る方針を示しています。これは、プロンプト工学や専用ハーネスだけで達成された改善を、AGI進歩の証拠と誤読しないためです。 7

今回の30.2%については、**Anthropicの本文テキスト**では「次点の約3倍」としか書かれておらず、**ARC Prizeの公式X投稿**が「30.2%」の絶対値を与えています。前世代のOpus 4.8についても、1.5%という絶対値はARC Prize公式X投稿で「new SOTA」として示されています。よって本稿では、**一次ソースの優先順位**を保ちつつ、絶対値についてはARC Prize公式投稿と、その図表を日本語で読み解いた二次資料を補助的に使っています。 8

実用上と研究上の含意

30.2%というスコアが実務的に意味するのは、モデルが単に“難問に長考できる”だけでなく、**初見の操作体系から有効な仮説を組み立て直す能力**を改善した可能性が高いことです。ARC PrizeはARC-AGI-3を、実世界のエージェントが遭遇する「説明不足のダッシュボード、癖のあるフォーム、未文書化の内部ツール、事前に想定されていない例外ケース」の縮図として位置づけています。したがって、この種のスコア上昇は、一般的な業務エージェントにおける**未知UI適応、曖昧な状態推定、長期タスクでの方針維持**にプラス方向の外挿を許します。 9

ただし、外挿には限界があります。ARC-AGI-3は**ターン制**で、報告でも「リアルタイムの反射運動ではなく、オフライン推論を優先する設計」とされています。つまり、このスコアはロボティクス一般、連続制御、身体性、マルチモーダルな実世界認識、対人相互作用を直接は保証しません。またARC-AGI-3では自然言語指示が与えられないため、企業システムのような言語豊富なタスクとは難しさの分布が異なります。 10

研究上は、スコアそのもの以上に**失敗の質が変わるか**が重要です。ARC PrizeがOpus 4.7とGPT-5.5を分析した際には、主な失敗は「局所的な因果は見えてもグローバルな世界モデルに昇華できない」「見た目が似た既知ゲームへ誤マッピングする」「レベル成功をゲーム理解へ一般化できない」の三類型でした。もしOpus 5の30.2%が本物なら、研究者にとっての重要点は「より多くの環境を解けた」ことよりも、**どの失敗類型が減ったのか**、そして**どの類型が残ったのか**です。 11

安全面では、能力向上は両義的です。AnthropicはOpus 5を、Fable 5に近い能力をより安価に提供するモデルと位置づけつつ、サイバー分類器はFable 5より**約85%少ない頻度**で発火し、**ソースコードの脆弱性発見は許可**する一方で、**バイナリベースの脆弱性探索、侵入テスト、エクスプロイト生成**は遮断すると説明しています。これは防御用途には実務上の利点ですが、同時に**境界面の設計を誤ると攻撃面も広がる**ことを意味します。またAnthropicはOpus 5が対人協力型のミスアライン行動や自律的有害行動の指標で低い値を示したと主張していますが、強いエージェント能力そのものが持つリスクを消すものではありません。 12

スコア急伸の要因分析

公開情報から最も確からしい説明は、**基盤モデル側の大幅改良**です。Anthropicはアーキテクチャの詳細を開示していませんが、Opus 5を「複雑なエージェント的コーディングとエンタープライズ作業向け」、Fable 5を「最も高能力」と整理し、Opus 5に**1Mトークン文脈、128k出力、adaptive thinking、high既定、full effort ladder**を持たせています。さらにプロンプティング文書では、Opus 5が**長時間のエージェント作業、長文脈、サブエージェント協調、自己修正**に強いとされています。ARC-AGI-3が要求するのはまさに、観測履歴を圧縮し、仮説を持続し、必要に応じて捨てて作り直す能力であるため、この適性の一致は偶然ではないはずです。 13

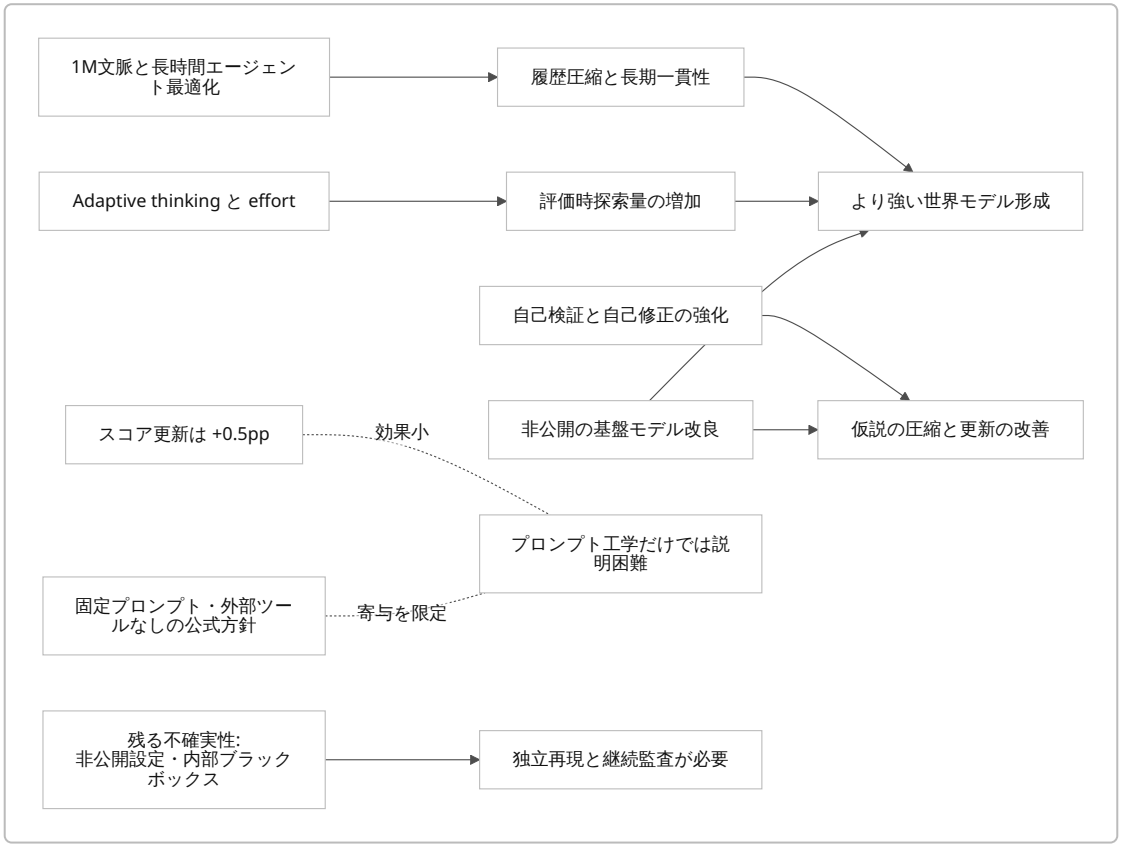
評価時推論の寄与も大きいとみられます。Opus 5は`low`から`max`までのeffortを持ち、`xhigh`や`max`では大きな`max_tokens`を与えて「subagents and tool calls」の余地を確保するよう文書化されています。これは、**テスト時計算量を増やすことで性能を押し上げる設計**が製品仕様の中核に入っていることを示します。もっとも、ARC Prizeの公式ARC-AGI-3スコアは外部ツールを与えない方針なので、少なくともフロントの評

価ハーネス由来の道具拡張が主因ではありません。ここで効いているのは、外部ツールではなく、**モデル内部またはAPI背後の思考量・探索量**である可能性が高いです。[14]

もうひとつ大事なのは、**自己検証の内在化**です。Anthropicのプロンプティング文書は、Opus 5は「言われなくても自分で検証する」「再確認やダブルチェックを明示指示すると過剰検証になる」と述べます。これは推測ですが、ARC PrizeがOpus 4.7に観察した「誤った理論への早すぎる圧縮」「成功理由の誤学習」を減らすには、**正しい自己検証のコスト効率**が鍵です。つまり、単純に“たくさん考えた”のではなく、**誤理論をより早く棄却できるようになった**ことが、30.2%のかなりの部分を説明している可能性があります。[15]

公開データのスケールや組成については慎重であるべきです。モデル概要ではOpus 5の**訓練データカットオフが2026年5月**、信頼知識カットオフも2026年5月と示されていますが、**データ量、合成データ比率、RLの詳細、アーキテクチャ様式**は確認できません。したがって、「インタラクティブ環境の軌跡データを大量に学習したからだ」といった説明は現時点では根拠不足です。最も厳密な表現は、**学習側の大規模改良があったことは確実だが、その中身は非公開**です。[16]

評価アーティファクトや“ゲーム化”の可能性も残ります。ARC Prizeは、ARC-AGI-3専用の人間指示やハーネス設定すら性能を不自然に押し上げると技術報告で警告しています。他方で、今回の30.2%はARC Prize公式がSOTAとして発信しているので、少なくとも**公然たる公開セット過学習だけではない**とみるのが自然です。それでも、Anthropic本文が絶対値や詳細設定を詳述していない以上、**どのeffortレベルで、何試行平均で、内部的にどれだけ思考量を使ったのか**は、なおブラックボックスです。[17]

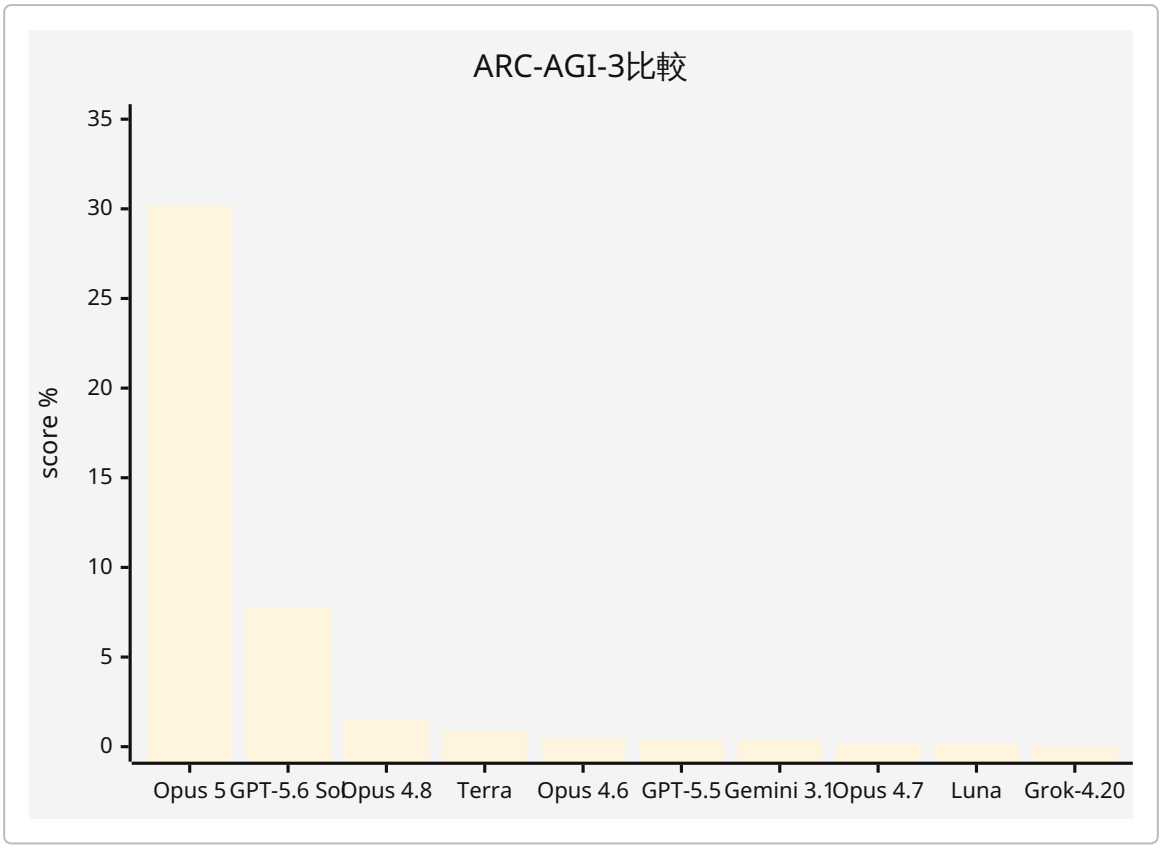


上図のうち、**固定プロンプト・外部ツールなし、スコア更新は+0.5ポイント程度、Opus 5のthinking/effort/長文脈**という三点は公開ソースで確認できます。そこから「基盤能力の改善が主因」という結論は、公開事実に基づく推論としては最も自然です。[18]

モデル比較と時系列

まずARC-AGI-3だけを見ると、2026年3月のリリース時点ではフロンティアAIは1%未満でした。その後、5月のGPT-5.5/Opus 4.7分析では0.43%と0.18%、6月のOpus 4.8で1.5%、7月のGPT-5.6 Solで7.78%、そして7月24日のOpus 5で30.2%へと、短期間で伸びています。これは**継続的改善**ではなく、少なくとも見かけ上は**段差的な飛躍**です。 [19]

モデル	ARC-AGI-3 スコア	データ種別	日付の目安	出典
Claude Opus 5	30.2%	ARC-AGI-3	2026-07-24	ARC Prize公式X投稿、Anthropic公開日 [20]
GPT-5.6 Sol Max	7.78%	Semi-Private	2026-07-09	ARC Prize verified results [21]
Claude Opus 4.8	1.5%	ARC-AGI-3	2026-06ごろ	ARC Prize公式X投稿経由 [22]
GPT-5.6 Terra Max	0.80%	Semi-Private	2026-07-09	ARC Prize verified results [23]
Claude Opus 4.6 Max	0.50%	Semi-Private at launch	2026-03	ARC-AGI-3技術報告 [24]
GPT-5.5	0.43%	Semi-Private	2026-05-01	ARC Prize分析記事 [11]
Gemini 3.1 Pro Preview	0.40%	Semi-Private at launch	2026-03	ARC-AGI-3技術報告 [24]
GPT-5.4 High	0.20%	Semi-Private at launch	2026-03	ARC-AGI-3技術報告 [24]
Claude Opus 4.7	0.18%	Semi-Private	2026-05-01	ARC Prize分析記事 [11]
GPT-5.6 Luna Max	0.18%	Semi-Private	2026-07-09	ARC Prize verified results [25]
Grok-4.20 Beta 0309 Reasoning	0.10%	Semi-Private at launch	2026-03	ARC-AGI-3技術報告 [24]



この棒グラフで重要なのは、**Opus 5だけが別の分布に見える**ことです。7.78%から30.2%への増加は、単なる“少し強い次世代”というより、ARC-AGI-3が測る能力のボトルネックの一部を越え始めた可能性を示します。もっとも、30.2%でも人間100%からは大きい差があり、依然として**未飽和 benchmark**です。 [26]

関連ベンチマークでもOpus 5は強いですが、ここはデータの出どころを分けて読むべきです。**Frontier-Bench v0.1**は公式サイトがGPT-5.6 Sol 34.4%、Fable 5 33.8%、Opus 4.8 21.1%を明示しており、74タスク・7領域のエージェント型端末作業を測ります。一方、**Opus 5の43.3%、AutomationBench 26.0%、OSWorld 2.0 70.6**といった絶対値は、Anthropic本文では主として図表・相対表現で示され、ここではその数値を日本語の二次資料で補いました。したがって、これらは**妥当だが再掲値であり、独立にverified pageで確認済みという意味ではない**、と読むのが厳密です。 [27]

ベンチマーク	何を測るか	Opus 5	比較対象	ソースの強さ
ARC-AGI-3	未知の対話環境での探索・世界モデル形成・目標発見・計画	30.2	GPT-5.6 Sol 7.78、Opus 4.8 1.5	公式投稿 + verified page + ARC資料 [28]
Frontier-Bench v0.1	74タスク・7領域のagentic terminal coding	43.3	GPT-5.6 Sol 34.4、Fable 5 33.7前後、Opus 4.8 21.1	Opus 5値は二次再掲、既存比較値は公式 [29]
AutomationBench	47ツールを跨ぐ業務ワークフロー完遂	26.0	GPT-5.6 Sol 18.1、Fable 5 17.4	Opus 5値は二次再掲、ベンチ定義は公式 [30]

ベンチマーク	何を測るか	Opus 5	比較対象	ソースの強さ
OSWorld 2.0	108本の長時間コンピュータ操作ワークフロー	70.6	Fable 5 66.1、GPT-5.6 Sol 62.6、Opus 4.8 55.7	Opus 5値は二次再掲、ベンチ定義は原論文 [31]

日付で見ると、ARC-AGI-3の顕著な変化は次のように並びます。 [32]

日付	出来事	意味
2019-11-05	Cholletが“On the Measure of Intelligence”公開	ARC系列の理論基盤を提示 [33]
2026-03-25	ARC-AGI-3公開、当時のフロンティアAIは0.51%	インタラクティブ推論への移行点 [34]
2026-05-01	GPT-5.5 0.43%、Opus 4.7 0.18%の失敗分析公表	失敗モードが可視化される [11]
2026-05-28	Claude Opus 4.8公開	Opus系の実務・エージェント能力を強化 [35]
2026-06ごろ	Opus 4.8がARC-AGI-3で1.5% SOTA	ただし依然として低水準 [22]
2026-06-09	Claude Fable 5公開	最上位一般公開モデルの基準点 [36]
2026-07-09	GPT-5.6 Solが7.78%	初めて“実用的に見える”水準へ上昇 [21]
2026-07-24	Claude Opus 5が30.2%	段差的なジャンプ、ただし未飽和 [20]

方法論上の注意点と未解決問題

最大の注意点は、**ARC-AGI-3の“何点”と“何で取ったか”を分けて考えること**です。ARC Prizeは、公開セットでの100%攻略ですらAGI進歩の妥当な指標ではないと明記し、公開セット特化・ARC専用ハーネス・人手で埋め込まれた戦略は公式スコアから切り離しています。したがって、今後もし30.2%を再現できない追試が出たとしても、それは必ずしも「Anthropicが誇張した」ことを意味せず、**測定条件の違い**を先に疑うべきです。 [37]

第二に、**ベンチマーク妥当性は高いが完全ではない**という点です。ARC-AGI-3は人間100%可解、未知環境、行動効率重視という点で強い設計ですが、依然として人工的・ターン制・抽象的です。実務上のエージェント障害は、認可、曖昧な指示、社会的駆け引き、長期メモリ管理、エラー回復、ユーザ確認不足などから起きます。OSWorld 2.0の論文も、最良モデルですら500ステップで20.6%しか完遂できず、しばしば制約追跡や隠れ状態の回復に失敗すると報告しています。つまり、ARC-AGI-3が改善しても、**実世界エージェントの完全自律運用**はまだ別問題です。 [38]

第三に、**Fable 5との比較は未整理な部分が残る**ことです。Anthropicのモデル概要ではFable 5がなお「最も高能力な一般公開モデル」とされつつ、Opus 5は一部ベンチマークでFable 5級または上回る値を示しています。さらにFable 5は30日保持要件つき、Opus 5は半額という運用差もあるため、実務での“最強”は単純なベンチスコアでは決まりません。ARC-AGI-3でも、公開確認できるのはOpus 5の30.2と、ARC Prize側の「Fable-class models are approximately 20% on public demo environments」という断片であり、**同一条件での厳密なA/B表はまだ足りない**と考えるべきです。 [39]

最後に、未解決問題は明確です。**どの能力が伸びたのか**。世界モデル形成か、長期メモリ圧縮か、自己修正か、単なる推論量増加か。**どの程度一般化するのか**。ARC-AGI-3で伸びた能力はWebエージェント、端末作業、業務自動化へどこまで移るのか。**どこまで安全に使えるのか**。より自律的なモデルは、防御的セキュリティ支援を強める一方で、境界を越えれば攻撃面も増やします。30.2%は、これらの問いを終わらせたのではなく、ようやく本格的に問える地点へ持ち込んだ、という理解がもっとも妥当です。 [40]

参考ソース

本稿の基礎になった一次ソースは、AnthropicのOpus 5公開記事、Claude Platformのモデル概要・移行ガイド・プロンプティング文書、ARC PrizeのARC-AGI-3技術報告・競技ページ・分析記事・verified resultsです。関連ベンチマークの定義には、Frontier-Bench公式告知、ZapierのAutomationBench公開記事、OSWorld 2.0論文、OpenAIのGDPval公開ページを優先しました。 [41]

日本語ソースとしては、Anthropic発表の速報を伝えたASCIIとPC Watch、およびAnthropic図表の絶対値を日本語で整理したテツメモや解説ノートを補助的に参照しました。特に、Anthropic本文が相対表現に留めた箇所の**数値再掲**にはこれらを使っています。そのため、本文中でも明示した通り、**関連ベンチマークの絶対値には一次・二次の混在がある**点に注意してください。 [42]

---

[1] [8] [12] [41] https://www.anthropic.com/news/claude-opus-5
https://www.anthropic.com/news/claude-opus-5

[2] https://arcprize.org/arc-agi/3
https://arcprize.org/arc-agi/3

[3] [9] [11] [40] https://arcprize.org/blog/arc-agi-3-gpt-5-5-opus-4-7-analysis
https://arcprize.org/blog/arc-agi-3-gpt-5-5-opus-4-7-analysis

[4] [13] [16] [39] https://platform.claude.com/docs/en/about-claude/models/overview
https://platform.claude.com/docs/en/about-claude/models/overview

[5] [19] [34] https://arcprize.org/blog/arc-agi-3-launch
https://arcprize.org/blog/arc-agi-3-launch

[6] [7] [10] [17] [18] [24] [37] https://arcprize.org/media/ARC_AGI_3_Technical_Report.pdf
https://arcprize.org/media/ARC_AGI_3_Technical_Report.pdf

[14] https://platform.claude.com/docs/en/about-claude/models/whats-new-opus-5
https://platform.claude.com/docs/en/about-claude/models/whats-new-opus-5

[15] https://platform.claude.com/docs/en/build-with-claude/prompt-engineering/prompting-claude-opus-5
https://platform.claude.com/docs/en/build-with-claude/prompt-engineering/prompting-claude-opus-5

[20] [26] [28] https://x.com/arcprize/status/2080716561539907928
https://x.com/arcprize/status/2080716561539907928

[21] https://arcprize.org/results/openai-gpt-5-6
https://arcprize.org/results/openai-gpt-5-6

[22] https://x.com/arcprize/status/2061512025638121516?lang=en
https://x.com/arcprize/status/2061512025638121516?lang=en

[23] https://arcprize.org/results/openai-gpt-5-6-terra
https://arcprize.org/results/openai-gpt-5-6-terra

25 https://arcprize.org/results/openai-gpt-5-6-luna
https://arcprize.org/results/openai-gpt-5-6-luna

27 29 https://www.frontierbench.ai/announcement
https://www.frontierbench.ai/announcement

30 https://zapier.com/blog/introducing-automationbench/
https://zapier.com/blog/introducing-automationbench/

31 38 https://arxiv.org/abs/2606.29537
https://arxiv.org/abs/2606.29537

32 33 https://arxiv.org/abs/1911.01547
https://arxiv.org/abs/1911.01547

35 https://www.anthropic.com/news/claude-opus-4-8
https://www.anthropic.com/news/claude-opus-4-8

36 https://www.anthropic.com/news/claude-fable-5-mythos-5
https://www.anthropic.com/news/claude-fable-5-mythos-5

42 https://ascii.jp/elem/000/004/422/4422202/
https://ascii.jp/elem/000/004/422/4422202/