Claude Opus 5におけるARC-AGI-3スコア30.2%達成の技術的要因と産業的インプリケーション

Gemini 3.6 Flash

ARC-AGI-3ベンチマークの定義と評価結果の全貌

2026年7月24日に米国Anthropic社より正式リリースされたClaude Opus 5は、フロンティア型人工知能モデルの性能評価体系において重要な転換点を示している[1]。前世代モデルであるOpus 4.8と同一のAPI利用料金（入力100万トークンあたり5ドル、出力100万トークンあたり25ドル）を維持しながら[2]、同社の最上位モデルであるClaude Fable 5に迫る、あるいは一部の領域ではそれを超える知能と自律実行力を実現している[1]。本モデルの能力を象徴する客観的指標として世界的な注目を集めたのが、新規問題解決力と自己適応型エージェント知能を測定する「ARC-AGI-3」ベンチマークにおける30.2%というスコアである[4]。

ARC-AGI（Abstraction and Reasoning Corpus）は、Googleの研究者でありKerasの創始者でもあるFrançois Chollet氏らが立ち上げたARC Prize Foundationによって運営される、人工汎用知能（AGI）の真の流動的知能を評価するためのベンチマーク体系である[8]。従来のARC-AGI-1および2が「静的な入出力グリッド図面から幾何学的変換ルールを推定する」形式であったのに対し、2025年から2026年にかけて展開されたARC-AGI-3は、完全対話型かつターン制の未知環境でエージェントを稼働させる評価設計へと進化を遂げた[8]。評価環境において、エージェントには事前の操作手順や明示的な目標、環境のルールルールが一切提示されない[7]。エージェントは未知の環境内で自発的に操作を試行し、それに対する状態変化を観測することで環境の動態モデルを自己構築し、目的を推論した上で一連の行動計画を立案・実行しなければならない[7]。

このベンチマークは「相対的人間行動効率（Relative Human Action Efficiency）」を軸に測定され、試行錯誤の回数や探索の無駄を排除した効率性が厳密に評価される[7]。人間がほぼ100%の成功率を記録する一方で[7]、従来のフロンティアモデル群はパターン記憶が通用しない環境に直面して破綻し、1%未満から一桁台前半のスコアで伸び悩んでいた[4]。2026年半ばにおける既存モデルの最高スコアは、極大の推論計算量を投入したGPT-5.6 Solの7.8%であり、前世代のOpus 4.8に至っては1.5%にとどまっていた[4]。このような背景下において、Opus 5が記録した30.2%というスコアは、従来の数値を約4倍引き離す歴史的な躍進であり、大規模言語モデルが単なる学習データの再構成を超えて、未知の環境における仮説検証型の思考を獲得しつつあることを示している[4]。

モデル名称	API価格（入力/出力 1Mトークン）	ARC-AGI-3 スコア	Frontier-Bench v0.1 スコア	主要自動化ベンチマークにおける費用対効果

Claude Opus 5	$5.00 / $25.00	**30.2%**	**43.3%**	OSWorld 2.0においてFable 5の最高性能を1/3強のコストで達成[1]
Claude Fable 5	$10.00 / $50.00	（未実施/非公開）[16]	33.7%	基準最高性能層（高コスト）[1]
GPT-5.6 Sol	$5.00 / $30.00	7.8%	（非公表/未掲載）[5]	Terminal-Bench 2.1等で優位性[5]
Claude Opus 4.8	$5.00 / $25.00	1.5%	18.7%	従来世代（試行エラー率が高い）[1]

この性能差は評価に要した総コストの観点からも裏付けられている[4]。Opus 5（High推論設定）が総評価コスト約2万ドルの計算資源で30.2%に到達したのに対し、GPT-5.6 Solは2.4万ドルのコストを消費しながら7.8%にとどまっており、成果あたりの歩留まりにおいて決定的な格差が存在する[4]。さらに、実領域のエンジニアリング能力を模したFrontier-Bench v0.1においても、Opus 5は43.3%を記録し、Fable 5の33.7%やOpus 4.8の18.7%を大幅に上回る最先端性能を示した[1]。

スコア高騰を可能にしたコア・技術メカニズムの深掘り

インタラクティブな仮説検証と閉ループ型試行錯誤

従来の大規模言語モデルは、プロンプトとして与えられた情報に対して一方向的に回答を出力する「オープンループ型」の推論構造を主としていた[12]。この構造は一般的な文章生成や既知のコード記述には有効であるものの、ARC-AGI-3のように操作のフィードバックに基づいて次の行動を調整しなければならない不確実性の高い環境では、初期の誤認がそのまま増幅してタスク破綻を引き起こす要因となっていた[12]。

Opus 5で導入された本質的な技術革新は、モデル内部における「閉ループ型」の動的仮説検証と自己修正メカニズムの高度化にある[1]。Opus 5は環境に対してアクションを実行した後、得られた状態変化を即座に自身の内部環境モデルと照合する[9]。自身の予測と実際の応答の間に相違を検知した場合、過去の仮説を棄棄し、新たな探索ルールを即座に再構築する思考プロセスを展開する[1]。計画の立案段階で理論的矛盾を事前に察知する「先行思考（Plan-time verification）」と、実行結果に基づいて解法を粘り強く調整する「継続的補正（Iterative persistence）」が統合されたことで、散漫なランダム探索を回避し、最小限のアクション数で未知のルールに到達することが可能となった[3]。

推論時計算量の動的な適応とeffortパラメータ制御

Opus 5における推論精度の飛躍は、アーキテクチャレベルでの適応的思考（Adaptive thinking）と、ユーザーおよびシステム側で制御可能な推論時計算量（effort設定）の最適化によってももたらされ

ている[3]。適応設定にはlow、medium、high、xhigh、maxという段階が設けられており、思考の深さと計算コストの調整が柔軟に行える仕組みとなっている[3]。
ARC-AGI-3のベンチマーク実行において、Opus 5はHigh以上の推論設定を適用することで、困難な分岐点における探索木の深さを大幅に拡張している[4]。過去のモデルでは長時間の推論ステップを踏むと、文脈の肥大化に伴って命令への追従性が低下したり、自己ループに陥ったりする現象が見られた[12]。しかし、Opus 5は100万トークンの広大なコンテキストウィンドウを厳密に保持しつつ[3]、ロングホライゾン（長期間・多段階）の推論チェーンにおいても一貫した評価基準を維持できるよう内部アテンション機構が強化されている[3]。この「テストタイム計算量（Test-time Compute）」の拡大と質の向上が、未踏の問題空間を走破する推論力の土台を形作っている[4]。

自律的ツール生成および障害迂回ロジック

未知の環境下でタスクを達成するためには、事前に提供されたツールを利用するだけでなく、不足している機能を自律的に定義・実装する高度な創発的挙動が不可欠である[3]。Opus 5の定性的検証プロセスにおいては、目的達成に必要な手段が存在しない場合に、自らコードを記述して独自の測定系やツールを構築する行動パターンが明確に確認されている[3]。
実際の検証例として、Frontier-Benchにおける機械部品の3D FreeCADモデル再構築タスクが挙げられる[3]。このタスクではモデルが画像を直接閲覧できないという制約条件があらかじめ与えられていたが、Opus 5は画像ファイルを直接参照できないと認識するや否や、生ピクセルデータから視覚的幾何学情報を抽出する独自のコンピュータビジョン処理パイプラインをPythonコードで書き上げ、データを数値化してCADモデリングを完遂させた[3]。同一の制限下で競合モデル群が5回の試行すべてで失敗したのに対し、Opus 5は自立的に迂回路を創造することで解決に至った[3]。同様に、金融市場のデータフィード構築タスクにおいて検証用の外部環境が存在しない状況に直面した際も、自らデータ解析の正当性を確かめるためのテストハーネス（検証用疑似環境）を自動生成してコードの正常性を確認する挙動が観測されている[3]。外部環境に受動的に依存せず、自律的に検証基盤を作り出すこの資質が、指示や前提を欠くARC-AGI-3の環境で極めて高い適応力を発揮した[3]。

憲法AIに基づく安全制御と探索の安定性

長時間の自律走行を行うAIエージェントにおいて、途中で予測不能な暴走や自己破壊的な操作を起こさない安全性・堅牢性は、タスクの完遂率に直接的な影響を及ぼす[3]。Anthropicが実施した自動行動監査によると、Opus 5は同社が開発したフロンティアモデルの中で最も高いアライメント性能を示し、意図しない逸脱行動の発生率を示すスコアで2.3という極めて低い値を記録した[3]。
憲法AI（Constitutional AI）に代表される行動規範への遵守率が高まったことで、探索空間における不合理な試行や、復元不能な副作用を伴う危険なコマンド実行が未然に防止されている[3]。この安全制御は単なる利用制限にとどまらず、エラーの累積によるシステムの過熱や思考の迷走を抑止するフィルターとして機能しており、多段階の探索プロセス全体における挙動の安定化と正答率の押し上げに寄与している[3]。

30.2%達成がもたらす産業的価値と実務的メリット

単発チャットから長時間自律走行エージェントへの変革

ARC-AGI-3における30.2%というスコアは、AIの用途が「人間からの指示に対して即座に文章やコードを返答する単発対話ツール」から、「人間に代わって複雑な業務手順を長時間自律的に遂行する

実用型エージェント」へと本格的に移行したことを実証している[1]。
従来のLLMをエンタープライズ業務に組み込む際、最大の障壁となっていたのは、複数ステップを伴う作業の途中でAIが想定外のエラーに直面した際に自力で復旧できず、処理がストップしてしまう点にあった[4]。未知の環境における適応力を示すARC-AGI-3のスコア向上は、業務プロセス内で突発的な例外や仕様変更が生じた場合でも、AIが状況を即座に再評価し、自己修正を行って最終ゴールまで完走する確率が格段に高まったことを意味する[1]。

ソフトウェアエンジニアリングと業務プロセスの飛躍的自動化

具体的な実務領域において、Opus 5の自律的思考力はソフトウェア開発、UI操作、業務ワークフロー自動化の面で明白な成果をもたらしている[1]。
ソフトウェアエンジニアリング分野においては、単に指定された関数を記述するレベルを超え、大規模なコードベース全体を俯瞰したリファクタリングや不具合修正が可能となった[1]。オープンソースのパッケージマネージャーに存在した実環境のバグ修正タスクにおいて、Opus 5は開発コミュニティが作成した既定のパッチが見落としていた根深い根本原因（Root Cause）とエッジケースを特定し、完全な修正案を提示した事例が報告されている[3]。
パソコン画面の直接操作能力を測定するOSWorld 2.0ベンチマークにおいては、Fable 5が記録した最高スコアを3分の1以下のコストで更新した[1]。WebブラウザやGUIアプリケーションを操作する際、モバイル表示の画面下部に隠れた要素や、動的に変化するレイアウトパターンを視覚的かつ論理的に認識し、自力で適切な操作手順を組み立てることが可能である[3]。また、企業向け業務自動化を評価するZapier AutomationBenchにおいても、同一タスクコストあたりの合格率で次点モデルの約1.5倍を記録しており、日常的な定型作業から非定型な例外処理を含むワークフローまでを通し切る処理能力が強化されている[1]。

未知のデータ空間および高度な科学研究への適応

過去のトレーニングデータに依存したパターンマッチングから脱却した知能は、最先端の自然科学領域や学術研究においても大きな力を発揮する[3]。
Anthropicの内部評価によると、Opus 5はバイオインフォマティクス、構造生物学、有機化学を含むライフサイエンス領域の評価指標すべてで前世代モデルであるOpus 4.8を上回った[3]。分光法データから複雑な分子構造を推測する有機化学タスクでは10.2ポイントの向上を記録し、タンパク質の配列変異がその機能に及ぼす影響を予測するタスクにおいても7.7ポイントの精度向上を達成している[3]。
研究現場において、未解明の実験データが提示された際に、適切な統計的検定手法を自ら選択して偽陽性を排除し、独立した複数の解析アプローチを組み合わせて結論をクロスチェックする「慎重な研究者（Careful Scientist）」としての挙動を示すため、創薬や新素材開発における仮説構築のスピードを劇的に加速させる可能性を秘めている[3]。

タスク完遂単価（Cost per Task）の劇的低減とROIの改善

AI導入の経済性を評価する上で最も重要な観点は、単一トークンの表面的な単価ではなく、「目的とする1つの成果物を完成させるまでに消費した総コスト（Cost per Task）」である[4]。

Opus 5は単価自体をOpus 4.8と同額（入力 $5/出力$ 25）に抑えつつ[2]、過剰な出力や無駄な試行錯誤を大幅に削減することで、実質的なタスク完遂コストを劇的に引き下げた[3]。金融モデリングの専門タスクにおいては、Opus 4.8と比較して精度を9ポイント引き上げながら、処理にかかるターン数を

3分の1に、所要時間を60%短縮させることに成功している[3]。さらに、取引システムのベンチマークにおいては、推論トークン数を7分の1に圧縮してピーク性能を叩き出しており、計算資源の消費効率が極めて高い[3]。企業にとっては、同一の予算で実行可能な自律タスクの件数が数倍に拡大することを意味し、生成AI投資における投資対効果（ROI）を根本から改善する道を開いている[4]。

領域別の性能バイアスと実務運用における限界

ARC-AGI-3における30.2%という数値はモデルの汎用的な思考力を示す卓越した指標であるものの[4]、Opus 5がすべての個別タスクや専門ドメインにおいて一律に他モデルを圧倒しているわけではない[4]。実務導入にあたっては、得意領域と相対的な苦手領域を客観的に把握することが不可欠である[4]。

評価領域・ベンチマーク	**Opus 5**の相対的性能	優位性を示す競合・他モデル	運用上の留意点とインプリケーション
新規環境における問題解決 **(ARC-AGI-3)**	圧倒的優位 **(30.2%)** [cite: 4, 7]	なし（2位はGPT-5.6 Solの7.8%）[4]	未知のルールに基づく探索やインタラクティブな環境適応において比類のない強みを持つ[4]。
自律的ターミナルエンジニアリング **(Frontier-Bench)**	圧倒的優位 **(43.3%)** [cite: 1, 3, 7]	なし（Fable 5は33.7%）[1]	複数ファイルにわたる高度なバグ修正や仕様変更において高い完遂率を提示[1]。
単体エージェントコーディング **(DeepSWE v1.1)**	良好だが追従 (68.8%)[4]	**GPT-5.6 Sol (72.7%) / Fable 5 (69.7%)** [cite: 4]	単発のコード生成や特定フレームワークの自動修正では競合モデルが僅差で先行[4]。
法務ドメイン評価 **(Legal Agent Benchmark)**	良好だが追従 (11.7%)[4]	**Claude Fable 5 (13.3%)** [cite: 4]	最高難度の高度な契約・法務論理プロセスでは、最上位モデルであるFable 5が優位を維持[4]。
サイバー攻撃・エクスプロイト生成	意図的に限定[3]	**Claude Mythos 5** [cite: 3, 6, 17]	脆弱性発見性能（OSS-Fuzz）は高いが、攻撃コード生成は安全基準により固

			くブロックされる[3]。

定量データが示す通り、Opus 5の強みは「長期間回すエージェント」「PC操作」「業務ワークフロー」「新規問題解決」という、1回のやり取りで完了しない試行錯誤型のタスクに集中している[4]。一方で、特定の単体プログラミングベンチマークや、高度に専門化された法務の議論においては、Fable 5やGPT-5.6 Solが優位を保っている領域も存在する[4]。また、セキュリティ分野において、ソースコードの脆弱性発見能力は高水準にあるものの、バイナリベースのスキャニングやエクスプロイト（攻撃コード）の作成機能は安全分類器によって厳格に制御されており、フラグが立った要求は自動的にOpus 4.8等の安全なモデルへフォールバックする運用が組み込まれている[3]。

さらに、ベンチマークスコアの「絶対値」に対する冷徹な視点も要求される。ARC-AGI-3の30.2%はAI技術として前人未到の数字であるが、人間の到達度（100%）と比較すれば約7割の課題は未解決のままである[7]。また、実務自動化を測定するAutomationBenchの絶対値も26.0%であり、全体の約4分の1のタスクを通過させる水準にとどまっている[4]。したがって、現場への導入に際しては、AIにすべての決定を丸投げする完全自動化を目指すのではなく、AIエージェントが自律的に作業を進めつつも、重要局面で人間が確認・承認を行う「人間協働型（Human-in-the-loop）」のプロセス構築が現実的かつ堅牢なアプローチとなる[4]。

総括と今後の展望

Claude Opus 5がARC-AGI-3において記録した30.2%というスコアは、AIモデルの性能評価軸を従来の「知識量や静的なパターン認識」から、「未知の環境下における流動的知能と自律的エージェント力」へと不可逆的にシフトさせた[1]。

この飛躍を可能にしたのは、閉ループ型の仮説検証メカニズム[1]、推論時計算量（effort）の柔軟な制御[3]、目的達成に必要なツールを自ら作り出す創発的思考[3]、そして長時間の思考プロセスを支える高度なアライメント技術の融合である[3]。

実務面において、この技術的進化は単なる処理速度や精度の向上にとどまらず、タスクの完遂歩留まりを高め、実質的な完遂コストを劇的に引き下げるという直接的なビジネスメリットをもたらしている[1]。領域ごとの特性や絶対値の限界を正しく見極めつつ、長期間稼働するエージェント基盤としてOpus 5を組み込むことは、次世代のソフトウェア開発やエンタープライズ業務の自動化を推進する上で最も強力なアプローチの一つとなる[4]。

引用文献

1. Anthropic、「Claude Opus 5」提供開始、Fable 5に迫る性能を半額のAPI料金で, https://news.mynavi.jp/techplus/article/20260725-4738495/
2. Anthropic、「Claude Opus 5」公開 Fable 5に迫る性能を半額で——サイバー安全策は緩和、拒否時は自動フォールバックも - ITmedia NEWS, https://www.itmedia.co.jp/news/articles/2607/25/news016.html
3. Introducing Claude Opus 5 - Anthropic, https://www.anthropic.com/news/claude-opus-5
4. 【2026年7月版】Claude Opus 5のベンチマークを読み解く。負けている5項目と勝ち方の偏り - note, https://note.com/kazu_t/n/ndaaae23d51ed
5. AIモデル使い分けマップ2026 | Opus 5・GPT-5.6 | Claude Code for Business - Fyve,

https://fyve.co.jp/claude-code/articles/claude-opus-5-model-selection-map
6. 「Claude Opus 5」が登場、Claude Fable 5の半額で同等性能を発揮し制限緩め - GIGAZINE, https://gigazine.net/news/20260725-claude-opus-5/
7. Claude Opus 5 Benchmarks Explained - Vellum, https://www.vellum.ai/blog/claude-opus-5-benchmarks-explained
8. ARC-AGI-3 : le Test qu aucune IA ne Réussit (Humains 100%, IA 0,37%) - Pasquale Pillitteri, https://pasqualepillitteri.it/fr/news/607/arc-agi-3-benchmark-ia-test
9. ARC Prize 2026 - ARC-AGI-3 - Kaggle, https://www.kaggle.com/competitions/arc-prize-2026-arc-agi-3
10. ARC-AGI-3: A New Challenge for Frontier Agentic Intelligence - arXiv, https://arxiv.org/html/2603.24621v1
11. History - ARC Prize, https://arcprize.org/history
12. What Is ARC AGI 3? The Interactive AI Benchmark Humans Solve at 100% | MindStudio, https://www.mindstudio.ai/blog/what-is-arc-agi-3-interactive-benchmark
13. Anthropic「Claude Opus 5」発表、実務エージェントでFable 5に迫り料金は半額 - XenoSpectrum, https://xenospectrum.com/anthropic-claude-opus-5-agent-cost/
14. Claude Code v2.1.219 の主要アップデート - Claude Opus 5 の登場、Opus 4.8 と同じ価格、2.5 倍の Fast モードも提供 | DevelopersIO, https://dev.classmethod.jp/articles/20260725-cc-updates-v2-1-219/
15. Anthropic、「Claude Opus 5」公開 最高峰Fable 5に迫る性能を半額で, https://www.zaikei.co.jp/amp/article/20260725/862761.html
16. Introducing Claude Opus 5 : r/Anthropic - Reddit, https://www.reddit.com/r/Anthropic/comments/1v5h6r8/introducing_claude_opus_5/
17. Claude Opus 5とは｜料金・ベンチマーク比較 - Fyve, https://fyve.co.jp/claude-code/articles/claude-opus-5-complete-guide
18. Claude Code v2.1.219 Major Updates - Introduction of Claude Opus 5, same price as Opus 4.8, also offering 2.5x Fast Mode, https://dev.classmethod.jp/en/articles/20260725-cc-updates-v2-1-219/
19. Introducing Claude Opus 5 : r/ClaudeAI - Reddit, https://www.reddit.com/r/ClaudeAI/comments/1v5h6o9/introducing_claude_opus_5/
20. Claude Opus 5 使い方ガイド｜最強AIとほぼ同じ性能が半額で使えるようになった - note, https://note.com/naoki_35/n/n70fc9d760c90