NVIDIA AVO の ARC-AGI-3 「100%」達成

— 事実関係の検証と意義の分析 —

2026 年 8 月 23 日(第 2 版・一次資料再検証済)

Claude Opus 5

---

1. 要旨

- NVIDIA は 2026 年 8 月 21 日、自社の汎用エージェント・アーキテクチャ「AVO(Agentic Variation Operators)」が、Claude Opus 5 を基盤モデルとして ARC-AGI-3 の公開セット 25 環境・全 183 レベルを RHAE 100.00 で突破したと発表した [1]。主張の核心は「モデル単体の性能ではなく、ハーネス(エージェント設計)が性能を決める」というシステムレベルの命題であり、AGI 達成やベンチマーク攻略の証明ではない。
- ARC Prize は技術報告書において、公開セットのスコアを公式リーダーボードに掲載しない方針を明示し、人間リプレイにより公開全環境で 100%を取るオープンソース・ハーネスを公開すると述べている [3]。公開セットの 100%は、同財団の設計上「そうなることが想定されていた事象」である。
- VISTA(Kaiming He グループ)が先行して同じ公開セットで RHAE 100 を達成しており [7]、AVO の新規性は「アクション数が 12%少ない(6,624 対 7,542)」ことと、「GPU カーネル最適化用に設計されたアーキテクチャがそのまま転移した」ことにある [1]。
- 技術的意義は本物だが限定的である。長期ホライズン自律エージェントにおける「永続メモリ+監督+反復ループ」というスキャフォールディングの汎用性を示した好例といえる。ただし対照実験ではなく、準私設・完全私設セットは未評価である [1]。

2. 事実関係の整理(一次情報で確認済み)

2.1 発表の概要

NVIDIA Technical Blog、2026 年 8 月 21 日公開(同日中に更新)。著者は Terry Chen(principal engineer)、Yeyin (Eva) Zhu、Zhifan Ye、Jean-Francois Puget、Humphrey Shi(VP, High-Performance AI)の 5 名 [1]。

なお同記事末尾には編集注記があり、「ARC-AGI-3 の公開セットと準私設・完全私設の競技用セットをより正確に区別するよう文言を更新した」と記されている [1]。公開直後に「公開セットである」旨の限定が補強された経緯がうかがえる(注記の存在は事実、意図は推測)。

2.2 AVO とは何か

AVO は Agentic Variation Operators の略で、NVIDIA が開発した汎用コーディングエージェント・システムである。進化的探索(evolutionary search)における固定的な変異(variation)ステップを、次の候補をどう生成するか自ら決める自律エージェントに置き換える発想に立つ。論文は arXiv:2603.24517[2]。

アーキテクチャは次の三要素から構成される [1]。

- inspect–plan–implement–evaluate の反復ループ
- 永続メモリ(persistent memory):過去の実装・評価結果・コンパイラ/プロファイラ出力・蓄積した推論を、単一のモデルコンテキストを越えて保持する
- supervisor(スーパーバイザー):探索の停滞や非生産的サイクルを監視し、必要に応じて主エージェントを別戦略へ誘導する

基盤モデルは Claude Opus 5(主)であり、GPT-5.6 Sol でも限定的にテストされている。モデルの追加学習は行われておらず、あくまでハーネス(scaffolding)側の設計である [1]。

先行実績として、DGX B200 上で 7 日間連続の自律実行を行い、500 超の最適化方向を探索して 40 のカーネルバージョンをコミットした。multihead attention では、評価した構成において cuDNN 比で最大 3.5%、FlashAttention-4 比で最大 10.5%の性能向上を報告している。さらに、得られたカーネルを追加約 30 分の自律作業で grouped-query attention へ適応させたとする [1]。

> 【第 1 版からの修正】第 1 版は「論文では対 cuDNN 最大 7.0%、対 FlashAttention-4 最大 9.3%」と併記し、公式ブログとの数値の食い違いがあるかのように記述していた。今回ブログ本文を精査したところ、ブログはこれらの数値を記載しておらず、7.0%/9.3% は grouped-query attention 適応後の値として論文側で報告されたものと解される。ただし論文本文を今回直接再検証できていないため、本稿では当該数値を本文から外し、ブログで確認できる範囲に限定した [1]。

2.3 ARC-AGI-3 とは何か

ARC Prize Foundation(共同創設者 François Chollet、Mike Knoop)による初のインタラクティブ推論ベンチマークである。技術報告書は arXiv:2603.24621、v1 の登録日は 2026 年 3 月 24 日 [3]。公式ブログ「Announcing ARC-AGI-3」は 2026 年 3 月 25 日公開で、同日サンフランシスコの Y Combinator 本部において Chollet と Sam Altman(OpenAI CEO)の対談とともに公開された [4]。

> 【第 1 版からの修正】第 1 版は「技術報告書の日付表記は 3 月 27 日」としていたが、arXiv の初版表示は 2026 年 3 月 24 日であり、誤りであった [3]。

ARC-AGI-1/2 が静的なグリッドパズルであったのに対し、ARC-AGI-3 は指示・ルール・目標のないターン制の環境にエージェントを置き、相互作用を通じて探索(exploration)・モデル化(modeling)・目標設定(goal-setting)・計画と実行(planning and execution)の四要素を評価する。観測は 16 色の 64×64 グリッド、行動空間は 5 つのキー操作+Undo+座標指定の選択という小さな集合に限定される。言語・数字・文化的記号を排し、Core Knowledge priors(物体性、幾何・位相、基本物理、エージェント性)のみに依存する [3]。

データセット構成は、公開デモ 25 環境、準私設 55 環境、完全私設 55 環境である。ARC-AGI-2 が公開:私設で約 10:1 であったのに対し、ARC-AGI-3 はこの比率を反転させ、公開セットを「訓練資源」から「デモ用インターフェース」へ位置づけ直している [3]。

人間側の較正では、各環境を 10 名が試行し、独立に 2 名以上が全レベルを完遂できた環境のみが採用された。その結果、全 ARC-AGI-3 環境は事前訓練なしの人間により 100%解けることが検証されている。他方、公開時点の公式(準私設)リーダーボードの最高値は Gemini 3.1 Pro Preview の 0.37%であり、GPT 5.4 (High) 0.26%、Opus 4.6 (Max) 0.25%、Grok-4.20 0.00% が続く [3]。

> 【第 1 版からの修正】第 1 版は「フロンティア AI は 0.51%」としていたが、技術報告書の該当表の最高値は 0.37%であり、本文抄録の記述も「1%未満」である。0.51%は一次資料で確認できなかったため撤回する [3]。

2.4 スコアリング(RHAE)

スコアは Relative Human Action Efficiency(RHAE)と呼ばれる。現行の公式スコアリング解説によれば、レベルスコアは(人間ベースライン行動数÷AI 行動数)の二乗であり、人間より少ない行動で完遂した場合の上限は 1.15 倍に制限される。ゲームスコアはレベル番号を重みとする加重平均であり、後半レベルほど重みが大きい。総合スコアは全ゲームスコアの平均で、0〜100%に正規化される [5]。

二乗項により非効率な総当たり探索は強く抑制される。人間ベースラインの 5 倍の行動を要したレベルのスコアは$(1/5)^2$=4%まで低下する。加えて、評価コスト抑制のため、レベルごとに人間ベースラインの 5 倍の行動数で打ち切る運用が置かれている [3, 5]。

> 【第 1 版からの修正】第 1 版は式を「min(1.0/1.15,$(人間÷AI)^2$)」と記載していたが、これは二つの一次資料の記述を取り違えて混成したもので誤りである。あわせて「人間の 5 倍で実質 0 点」との記述も、数値上は 4%であり不正確であった [5]。

人間ベースラインの定義には、二つの一次資料の間に食い違いがある。技術報告書(2026 年 3 月)は「各環境を正確に 10 名がテストし、行動数で 2 番目に良い人間(second-best)をベースラインとす

る」と記す一方[3]、現行の公式スコアリング解説は「複数の初見プレイヤーの upper median(偶数人の場合は中央2値のうち行動数の多い側。4名なら3位、5名なら3位)」と記す[5]。本稿は現行仕様として後者を採り、報告書公表後に定義が改定された可能性を指摘するにとどめる(定義変更の有無・時期は未確認)。

> 【第1版からの修正】第1版は「10名の初見テスターの2番目に良い(upper median)行動数」と、両資料を等値に扱っていた。upper median は一般に「2 番目に良い」を意味せず、両者は別概念である[3,5]。

レベルスコアの上限についても同様の食い違いがある。技術報告書は「1.0倍で打ち切る」と明記するが[3]、現行解説は「1.15 倍」とする[5]。本稿は現行解説に従う。

2.5 「100%」が具体的に意味するもの

NVIDIA の報告値は、公開セット 25 環境(各 6〜10 レベル、計 183 レベル)を RHAE 100.00、総アクション数 6,624 で完遂したというものである。VISTA が報告する 7,542 アクションと比較して約12%少ない[1,7]。

観測インターフェースにも差異がある。VISTA の主構成が 512×512 のレンダリング PNG(テキストグリッド表現も併せて検討)であるのに対し、AVO の構成では LLM はテキストのみのモダリティで動作し、各観測は厳密な64×64テキストグリッドとして与えられ、画像および画像トークンは一切モデルに渡されない。エージェントはルールや目標の説明なしに利用可能な行動のみを与えられ、相互作用を通じてその効果を推論する[1]。

総合 100%は、全ゲーム・全レベルを完遂したうえで、ゲーム単位の加重平均において人間ベースライン相当以上の行動効率を達成したことを意味する。レベル単位では上限 1.15 倍の超過得点が認められるため、同一ゲーム内の一部の非効率なレベルが相殺されている可能性はあり、「すべてのレベルで人間以上に効率的だった」ことまでは含意しない[5]。

あわせて以下の限界に留意が必要である。

- 公開セット限定であり、準私設・完全私設セットは未評価である。NVIDIA 自身も「これらは公開セットの結果であって、準私設・完全私設の競技用セットの結果ではない」と明記している[1]。
- 対照実験ではない。対 VISTA 12%減も、後述の 30.16%との対比も、エージェントバックエンド・観測表現・メモリ・コンテキスト管理・推論設定という複数変数が同時に異なる。NVIDIA 自身が本文中で2度これを明記している[1]。

- 計算コスト・所要時間は非開示である。ARC Prize の公式リーダーボードが横軸にコストを置く設計であること、および高推論設定でのフル評価が数万ドル規模になりうるとされることを踏まえると、効率性の評価にはコスト情報が不可欠である [3, 13]。

3. ARC Prize 側の公式見解

ARC Prize は技術報告書において、公開セットのスコアを公式リーダーボードに一切掲載しないと明記している。理由は「システムが特別に設計・訓練されていない問題に対処する能力こそが汎用知能」であり、公開環境の知識をもって作られたハーネスは task-specific overfitting に当たるという点にある。同報告書はさらに、公開セットで 100%が取れることを示すために人間リプレイを用いたオープンソース・ハーネスを公開するとし、「公開セットは断じて AGI への進捗の妥当な指標ではない」と述べる [3]。

同財団は harness 研究の価値自体は否定しておらず、自己申告・未検証を前提とする「コミュニティ・リーダーボード」を別途設けている。AVO の公開セット結果は、制度上はこの枠に対応するものと位置づけられる(制度の存在は事実、当てはめは分析)[3]。

ハーネスの一般化限界について、同財団は自ら実証データを示している。公開 3 環境(ls20、ft09、vc33)を対象に汎用ハーネスを構築させ、公開セット全体で試験したところ、極端な二峰性が観測された。TR87 の変種では Opus 4.6 がハーネスなしで 0.0%、Duke ハーネスで 97.1%に達する一方、BP35 では両条件とも 0.0%であった。同財団はこれを「特別に設計されたハーネスは AGI 進捗の測定手段として有用でない」根拠としている [3]。

現時点で ARC Prize が検証済み(verified)として掲載する ARC-AGI-3 の最高スコアは、Claude Opus 5(High reasoning)の 30.16%であり、2026 年 7 月 24 日設定である。同ページは対象を「ARC-AGI-3 Public Demo(25 環境)」と明示し、環境ごとの内訳(100.0%が 5 環境、0.0%が 3 環境など)を公開している [6]。

> 【第 1 版からの修正】第 1 版は 30.16%の対象セットを明示せず、また「準私設セットの検証済みピークは GPT-5.6 Sol の 7.78%」と記載していた。前者は公開デモセットの値であることが確認できたため明記した。後者は一次資料で確認できなかったため撤回する [6]。

この点は重要である。AVO の 100%と Opus 5 の 30.16%は、いずれも同一の公開デモ 25 環境を対象としている。したがって両者の差はセットの難易度差ではなく、ハーネス・推論設定・評価系の差に帰せられる。もっとも NVIDIA が明記するとおり、これは AVO の寄与を分離した測定ではない [1, 6]。

François Chollet は(AVO 発表前の時点で)X 上において、公開セットは demonstration set であって eval でも training でもなく、公開デモセットのスコアは私設セットのスコアを示すものではないと繰り返し注意喚起している[14]。AVO を名指しした ARC Prize の公式反応は、本調査時点(2026 年 8 月 23 日)では確認できなかった。

4. 先行・競合事例

AVO の位置づけを理解するには、公開セットにおける先行事例との比較が不可欠である。

システム / モデル	評価対象セット	スコア	検証状況
NVIDIA AVO(Claude Opus 5)	公開セット(25 環境 183 レベル)	RHAE 100.00 / 6,624 アクション	自己申告・ARC Prize 未検証
VISTA(Claude Opus 5)	公開セット(同 183 レベル)	RHAE 100 / 7,542 アクション	自己申告・ARC Prize 未検証
Schema(Opus 4.8+Fable 5)	公開セット	98.98%	自己申告・ARC Prize 未検証
GPT-5.6 Sol(OpenAI 設定変更)	公開セット	13.3%→38.3%	OpenAI 自己申告
Claude Opus 5(High reasoning)	公開セット(Public Demo)	30.16%	ARC Prize 検証済 (2026/7/24)
Gemini 3.1 Pro Preview ほか	準私設セット(公開時点)	最高 0.37%	ARC Prize 公式
人間(初見テスター)	全環境	100%完遂	ARC Prize 公式

注:各行はスコアの定義・評価条件・検証主体が異なり、直接比較はできない。出典は本文各節を参照。

- VISTA(Kaiming He グループ):Claude Opus 5 を基盤モデルとして公開 25 ゲーム全 183 レベルで勝率 100%、RHAE 100 を AVO より先に達成し、初見の人間プレイヤー比で 56.0%少ないアクション数を報告している[7]。AVO はこの VISTA の direct-interaction 設計原則を採用しつつタスクインターフェースを独立に再実装したもので、エージェントバックエンドは根本的に異なるとする。VISTA が Claude Code や Codex 経由でモデルを駆動するのに対し、AVO は永続メモリ・監督・独自実行ループを備える[1]。
- Schema(Impossible Research、2026 年 7 月):Opus 4.8+Fable 5 で公開セット 98.98%、GPT-5.6 Sol で 95.35%(いずれも自己申告・未検証)[8]。

- Tycho:明示的な実行可能ワールドモデルを構築する方式(arXiv:2607.28287)。AVO はこれを採らず、ARC 特化のワールドモデル層を持ち込まない汎用エージェント方式を選択した[1]。
- OpenAI(2026 年 7 月 29 日):retained reasoning と compaction の 2 設定を有効化することで、GPT-5.6 Sol の公開セットスコアが 13.3%から 38.3%へ上昇した。モデル重みは不変である[9]。

5. 技術的評価

5.1 汎用性(general-purpose)の主張の妥当性

NVIDIA の中核主張は、ARC 特化のワールドモデル層を導入せず、GPU カーネル最適化用の同一エージェントループを、タスクインターフェースとツールのみ差し替えて転移させたという点にある。ブログは「ドメインは変わり、フィードバック経路も変わるが、中核のエージェントループは変わらない。転移するのはドメイン知識ではなく、持続的な自律進行のための機構である」と述べる[1]。この主張は検証可能で妥当性が高い。

もっとも、VISTA の direct-interaction 設計原則を借用している以上、完全にゼロからの転移ではない。また ARC Prize が示した二峰性の実証(3.節)は、公開環境で高性能を示すハーネスが未見環境へ一般化するとは限らないことを示しており、AVO の汎用性の主張も私設セットでの検証を経て初めて確証される(分析)。

5.2 ARC-AGI の飽和(saturation)議論

ARC Prize 自身が ARC-AGI-1/2 について、公開セットと私設セットが同分布に近いと高次の近道による攻略が可能になると認め、Gemini 3 Deep Think の推論過程に ARC-AGI 特有の色対応が現れた事例を証拠として挙げている[3]。ARC-AGI-3 は公開:私設比を反転させ、公開セットを「デモ」に格下げすることでこの問題に対処した。

公開セットにおける 100%の連発(VISTA→Schema→AVO)は、まさに同財団が予見し警告していた現象である。したがって「ベンチマークが飽和した」ではなく「設計上、公開セットは容易である」と理解するのが正確である(分析)。真の進捗指標は準私設・完全私設セットであり、そちらの検証済みスコアは依然として低い水準にとどまる[3]。

6. 各方面の反応

- X(公式スレッド):NVIDIA AI 公式アカウントが「汎用コーディングエージェントが 100%達成」と投稿した。祝福と同時に「公開データへの過学習」との批判が寄せられ、split を無視した提

示であるとの指摘も見られた。ただし本調査ではこれらの投稿を直接検証できておらず、二次引用・スナップショット経由での間接的確認にとどまる[14]。

- Hacker News(コミュニティ投稿・反応資料、id=49387755、約 70 ポイント/39 コメント):「公開セットであって私設セットではない」「NVIDIA 独自モデルではなく Opus 5 へのハーネスである」「ARC-AGI-3 はハーネスを除外してモデルの生性能を測る設計ではなかったか」といった懐疑的な指摘が複数見られた。他方で「進化的手法の進展を過小評価している」との擁護もあり、AVO 論文著者による「この研究は半年前に GPU カーネル向けに行われ、同じ手法を ARC-AGI-3 に適用した」旨のコメントも投稿されている[13]。
- 懐疑論の核心は 4 点に整理できる。①公開セットの飽和(VISTA/Schema が既に 100%/99%)、②AGI 接近という解釈への否定、③対照実験ではないこと、④コスト非開示[13, 18]。

7. 意義の分析

7.1 この結果をどう読むべきか

「NVIDIA が ARC-AGI-3 を攻略した」「AGI に近づいた」と読むのは誤りである。正しくは、「長期ホライズン・エージェントにおいて、永続メモリ+監督+反復ループというスキャフォールディングが、無関係なドメイン(GPU カーネル最適化)から相互作用推論へ転移することを、公開セット上で実証した」と読むべきである(分析)。

実務家向けの教訓は明快である。すなわち、モデル選択よりもハーネス設計に残された改善余地がなお大きく、収穫逓減には到達していない。ARC Prize 自身も、優れたハーネス研究の成果は最終的にモデル API の内側へ流れ込む(chain-of-thought が第三者ハーネスから o1 という一次モデルへ至った経路)と見通しており、この観点からはハーネス研究の経済的価値と AGI 進捗の指標性とは切り分けて評価すべきことになる[3]。

7.2 NVIDIA の戦略的意味

GPU カーネル最適化という実利のあるドメインから相互作用推論への転移を示したことは、ロボティクス/フィジカル AI や業務自動化への橋渡しとして戦略的に整合する[21]。ハードウェア企業からフルスタック(モデル・アーキテクチャ・エージェント)へ踏み込む布石と読むのが自然である(分析)。なお同社は ARC-AGI-2 の ARC Prize 2025 競技でも NVARC チームが 1 位(24%)を獲得しており、ARC 系ベンチマークへの継続的関与という文脈もある[3]。

フィジカル AI(未知環境での自律適応)は、ARC-AGI-3 の「初見環境の探索」と問題構造が近い。AVO の長期ホライズン制御はここに直結しうる(推測)。

7.3 知財実務への含意

本件が示す「モデルではなくハーネスが性能を決める」という命題は、AI 関連発明の権利化・侵害立証の双方に影響を与えうる(分析)。基盤モデル自体ではなく、エージェント・オーケストレーション層(メモリ管理、監督機構、探索制御)に技術的貢献と差別化が生じるのであれば、保護対象の重心もそこへ移動する。他方、当該層はサービス内部で実施されるため外部からの侵害立証が困難であり、営業秘密による保護との使い分けが実務課題となる。なお本項は筆者の分析であり、一次情報に基づく事実の記述ではない。

7.4 今後注視すべき閾値

- AVO(または後続システム)が準私設/完全私設セットで ARC Prize 検証済みスコアを出すか。ここで高スコアが出れば意義は激変する [3, 6]。
- コスト/タスクと E2E 時間の開示。ARC Prize のリーダーボードがコストを横軸に据える以上、効率性主張の妥当性判定に必須である [3, 13]。
- NVIDIA が AVO を公開・オープンソース化するか [13]。
- GPT-5.6 Sol との体系的比較。現状は「Sol は実時間で速く、Opus は行動数で少ない」という予備的観察にとどまる [1]。
- ARC Prize 2026 は賞金総額 200 万ドルで ARC-AGI-3 トラックと ARC-AGI-2 トラックの二本立てで実施され、ARC-AGI-2 トラックは今年が最終年である。今後の主眼は ARC-AGI-3 に移る [3]。

8. 留意点(限界と未確認事項)

- 本報告の中核事実は、NVIDIA 公式ブログ、ARC Prize 技術報告書(arXiv:2603.24621)、ARC Prize 公式スコアリング解説、同 results ページを本版で直接再取得して確認した。二次メディアは補助的に用いた。
- AVO 論文(arXiv:2603.24517)本文は本版では直接再検証していない。同論文に帰される数値(grouped-query attention の性能向上率等)は本文から除外した。
- X 投稿(文献 14)は直接検証できておらず、二次引用・スナップショット経由の間接的確認にとどまる。したがって X 上の反応は「確認済み事実」ではなく「間接的に確認された反応」として扱っている。

- Hacker News は一次資料ではなくコミュニティ投稿・反応資料である。コメントの定量分析は行っていないため、懐疑論の分布については「複数見られた」以上の主張はしていない。
- RHAE の仕様(レベルスコア上限、人間ベースラインの定義)について、技術報告書と現行の公式スコアリング解説の間に食い違いがある。本稿は現行解説を採用したが、改定の有無・時期は確認できていない[3, 5]。
- AVO の ARC-AGI-3 実行に要したコスト・所要時間・私設セット性能は非公開である。対 VISTA 12%減は対照実験ではない(NVIDIA 自身が明記)[1]。
- 「汎用アーキテクチャ」「frontier-level」といった表現には NVIDIA のマーケティング的側面が含まれる。転移の事実自体は確認できるが、その一般性の程度は公開セット 1 本+GPU カーネル 1 本という 2 ドメインに基づく限定的な証拠である(分析)。

9. 第 1 版からの主な修正

- **RHAE 計算式:** 「min(1.0/1.15,(人間÷AI)²)」→「(人間÷AI)²、レベル上限 1.15 倍」。二資料の混成による誤りを修正[5]。
- **技術報告書の日付:** 「3 月 27 日」→ arXiv v1 は 2026 年 3 月 24 日[3]。
- **人間ベースライン:** 「10 名の 2 番目に良い(upper median)」→ 技術報告書と現行解説で定義が異なる旨を明示し、現行解説の upper median を採用[3, 5]。
- **GPU カーネル数値:** 7.0%/9.3% を「論文との食い違い」として併記していた記述を削除し、ブログで確認できる MHA の数値と GQA 適応の事実のみを記載[1]。
- **「人間の 5 倍で実質 0 点」:** 数値上は 4%。加えて 5 倍での打ち切り運用が別途存在することを明記[3, 5]。
- **「100%＝全レベルで人間並み」:** ゲーム単位の加重平均であり、レベル単位では上限 1.15 倍の相殺が生じうる旨を追記[5]。
- **30.16% の対象セット:** 公開デモ 25 環境の値であることを明記。AVO の 100%と同一セットであることを新たに指摘[6]。
- **準私設 7.78%:** 一次資料で確認できず撤回[6]。
- **フロンティア AI 0.51%:** 技術報告書の該当表は最高 0.37%。0.51%を撤回[3]。
- **出典番号の表示:** 連番が連結して見える不具合を修正(「3, 5」等の区切りを挿入)。

参考文献

[1] NVIDIA, "NVIDIA AVO Reaches 100% on ARC-AGI-3, Demonstrating a Frontier-Level General-Purpose Architecture for Long-Horizon Autonomous Agents," NVIDIA Technical Blog, 2026 年 8 月 21 日公開・同日更新. https://developer.nvidia.com/blog/nvidia-avo-reaches-100-on-arc-agi-3-demonstrating-a-frontier-level-general-purpose-architecture-for-long-horizon-autonomous-agents/(一次情報。本版で全文を直接確認)

[2] T. Chen et al., "AVO: Agentic Variation Operators for Autonomous Evolutionary Search," arXiv:2603.24517, 2026 年 3 月.(一次情報。本版では本文を直接再検証していない)

[3] ARC Prize Foundation, "ARC-AGI-3: A New Challenge for Frontier Agentic Intelligence," arXiv:2603.24621v1 [cs.AI], 2026 年 3 月 24 日. https://arxiv.org/html/2603.24621v1(一次情報。本版で全文を直接確認)

[4] ARC Prize, "ARC-AGI-3"(公式ベンチマークページ/ローンチ記事), 2026 年 3 月 25 日. https://arcprize.org/arc-agi/3(一次情報)

[5] ARC Prize, "ARC-AGI-3 Scoring Methodology," ARC-AGI-3 Docs(現行版). https://docs.arcprize.org/methodology(一次情報。本版で直接確認)

[6] ARC Prize, "Claude Opus 5 - ARC-AGI Results," 2026 年 7 月 24 日. https://arcprize.org/results/anthropic-claude-opus-5(一次情報・検証済スコアおよび公開デモ環境別内訳。本版で直接確認)

[7] VISTA: A Visual Harness For Reasoning in an Interactive World(プロジェクトページ). https://vista-research.github.io/(一次情報・自己申告)

[8] Impossible Research, "Frontier Models with Our Harness Achieve ~99% on ARC-AGI-3 Public"(Schema), 2026 年 7 月. https://schema-harness.github.io/(一次情報・自己申告)

[9] OpenAI, "How enabling two settings tripled our scores on the ARC-AGI-3 benchmark," 2026 年 7 月 29 日. https://openai.com/index/how-two-settings-tripled-our-arc-agi-3-scores/(一次情報)

[10] ARC Prize, "ARC-AGI-3 Preview: 30-Day Learnings." https://arcprize.org/blog/arc-agi-3-preview-30-day-learnings(一次情報)

[11] ARC Prize, "ARC Prize Verified Testing Policy." https://arcprize.org/policy(一次情報)

[12] Tycho: Active Abstraction with Programmatic World Models for ARC-AGI-3, arXiv:2607.28287. https://arxiv.org/abs/2607.28287(一次情報。文献 1 の参照先として確認)

[13] Hacker News, "Nvidia AVO scores 100% on the ARC-AGI-3 interactive reasoning benchmark," id=49387755. https://news.ycombinator.com/item?id=49387755(コミュニティ投稿・反応資料)

[14] @NVIDIAAI ほか X(旧 Twitter)投稿群、および François Chollet による公開セットの位置づけに関する一連の投稿。※本調査では二次引用・スナップショット経由で確認しており、投稿の直接検証には至っていない。

[15] The New Stack, "Claude Opus 5 scored 30% on ARC-AGI-3. Wrapped in Nvidia's AVO, it hit 100%." https://thenewstack.io/nvidia-avo-arcagi3-benchmark/(二次情報)

[16] Wccftech, "NVIDIA Built Its AVO Coding Agent To Optimize CUDA GPU Kernels…"

https://wccftech.com/nvidia-built-its-avo-coding-agent-to-optimize-cuda-gpu-kernels-and-it-just-achieved-a-100-score-on-a-public-test-without-receiving-any-prior-instruction/(二次情報)

[17] AlphaSignal, "NVIDIA's AVO Hits 100% on ARC-AGI-3 Where the Bare Model Scores 30%." https://alphasignal.ai/news/nvidia-s-avo-hits-100-on-arc-agi-3-where-the-bare-model-scores-30(二次情報)

[18] explainx.ai, "NVIDIA AVO: 100% ARC-AGI-3 Score, Public Set Only," 2026 年 8 月. https://www.explainx.ai/blog/nvidia-avo-arc-agi-3-100-percent-long-horizon-agents-august-2026(二次情報)

[19] DataCamp, "ARC-AGI-3: The New Interactive Reasoning Benchmark." https://www.datacamp.com/blog/arc-agi-3(二次情報)/ OfficeChai, "ARC-AGI-3 Released, Gemini 3.1 Pro Top Scores With Just 0.37 Percent." https://officechai.com/ai/arc-agi-3/(二次情報)

[20] AI Weekly, "NVIDIA's AVO Hits 100 on ARC-AGI-3, Uses 12% Fewer Actions." https://aiweekly.co/alerts/nvidias-avo-hits-100-on-arc-agi-3-uses-12-fewer-actions(二次情報)/ ai.wain.blog, "NVIDIA AVO Scores 100.00 on the ARC-AGI-3 Public Set by Changing the Harness, Not the Model." https://ai.wain.blog/en/nvidia-avo-arc-agi3-harness-O2KzuKeW/(二次情報)

[21] NVIDIA Newsroom, "NVIDIA and Global Robotics Leaders Take Physical AI to the Real World." https://nvidianews.nvidia.com/news/nvidia-and-global-robotics-leaders-take-physical-ai-to-the-real-world(一次情報)

※ 上記 URL はいずれも 2026 年 8 月 23 日時点での参照。文献 1・3・5・6 は本版で全文を直接再取得して検証した。文献 2・14 は直接検証に至っていない旨を本文 8 節に明記した。