NVIDIA AVO による ARC-AGI-3 公開セット評価の技術的深層とその構造的意義

概要と研究成果の要点

2026 年 8 月、NVIDIA の研究チームは長期自律エージェント基盤「Agentic Variation Operators （AVO）」が、AI の流動性知能（Fluid Intelligence）を測定するベンチマーク「ARC-AGI-3」の公開評価セット（Public Set : 全 25 環境・183 レベル）において、Relative Human Action Efficiency （RHAE）スコア 100.00 を達成したことを発表した [1]。本検証において AVO は、公開セットに含まれる全レベルを完答し、集計上の行動効率において初見の人間の上位中央値と同等水準を記録した [1]。

本結果の背景として、ARC Prize が別途公表している Claude Opus 5（High）のモデル単体評価スコアは 30.16%であった [1]。ただし NVIDIA 自身が注記している通り、この 30.16%の評価と AVO による評価では、推論設定、観測インターフェース、エージェント構成、評価セットアップが根本的に異なっている [2]。したがって、このスコア差は AVO というハーネス（Harness / Scaffolding）の純粋な寄与のみを直接測定した対照実験の結果として解釈すべきではないものの、適切な実行基盤と推論ループの設計がモデルの実効性能の発現に大きく関与することを示唆する事例として注目されている [1]。

AVO は元来、ゲーム推論用としてではなく、NVIDIA Blackwell（B200）アーキテクチャ向けの GPU カーネル（CUDA/PTX）の自動最適化を目的とした進化的探索フレームワークとして開発された [1]。専門的なハードウェア最適化タスク向けに設計された認知ループを、ルールの事前説明がない視覚的推論環境へと転用して高効率なタスク完了を達成した点は、長期自律エージェントの汎用的な設計アプローチの有効性を示す重要な知見となっている [1]。

評価領域 / 指標	対象タスク	達成成果・定量データ	比較対象ベースライン
ARC-AGI-3 (公開セット)	未知ゲーム環境の自律推論 (25 環境 / 全 183 レベル)[1]	**RHAE 100.00** (全レベル完答、総アクション数 6,624 回)[1]	Claude Opus 5 単体評価 (30.16%※別条件)[1]、VISTA (7,542 回)[1]
Blackwell B200 MHA カーネル	Multi-Head Attention の自律最適化 [1]	最大 **1,668 TFLOPS** (BF16)[1]	cuDNN 比 **+3.5%**、FlashAttention-4 比 **+10.5%** [cite: 1]

Blackwell B200 GQA カーネル	Grouped-Query Attention への自律適応 [1]	**30 分間の追加探索**で最適化完了 [1]	cuDNN 比 **+7.0%**、FlashAttention-4 比 **+9.3%** [cite: 1]
連続自律稼働安定性	GPU カーネルの進化的探索 [1]	**7 日間完全無人稼働** (500 以上の探索方向、40 世代コミット)[1]	従来の単発 LLM パイプライン (局所解スタック)[1]

AVO のアーキテクチャ設計と中核メカニズム

進化的探索（Evolutionary Search）の古典的枠組みにおいて、親世代の解集団 P_t から変異候補を生成する処理は以下のように定式化されてきた [1]。

$$\mathrm{Vary}(P_t) = \mathrm{Generate}(\mathrm{Sample}(P_t))$$

FunSearch や AlphaEvolve に代表される従来の LLM イン・ザ・ループ手法では、固定的なヒューリスティクスに基づいて親コードを選択し、言語モデルに対して 1 ターンのみのコード生成を要求する構造が一般的であった [1]。この構造では、LLM 自身が外部仕様書を参照したり、コンパイラ出力を解析して能動的にデバッグを行ったり、試行錯誤の軌跡を踏まえて仮説を更新したりする柔軟性が制限されていた [1]。

AVO はこの変異操作（Vary ステップ）そのものを、自律的なコーディングエージェントループへと置換するアプローチを採用している [1]。エージェントには過去の解の履歴である集団情報 P_t、ドメイン知識ベース K（CUDA 仕様書やゲーム仕様等）、および評価関数 f（スループット計測器や環境フィードバック等）が提供される [1]。エージェント自身が親の選択、参照すべき技術情報の決定、コード改変の計画、テストの実行、ならびにエラー発生時の自己修復を自律的に遂行する [1]。

長期にわたるタスク遂行においてエージェントが直面する大きな課題の一つが、局所的な解の周回や過去の失敗経路の再探索といった探索のスタグネーション（停滞）である [1]。AVO はこの課題に対し、主エージェント（Main Coding Agent）と監督者エージェント（Supervisor Agent）による 2 層構造を導入した [1]。

主エージェントがツールの実行や細部のコード生成、ステップごとの環境インタラクションを担当する一方で、監督者エージェントは全体の探索軌跡を上位視点から監視する [1]。主エージェントが同一のエラーを反復したり進捗が停滞したことを検知すると、監督者は上位指示を注入して探索の前提を見直させ、異なる戦略への軌道修正を促す [1]。

また、長大な試行履歴を維持するため、AVO は過去の実装差分、コンパイラおよびプロファイ

ラの出力、形成した推論仮説を破棄せずに引き継ぐ累積的な会話・コンテキスト管理機構（Persistent Memory）を備えている[1]。これにより、長期間にわたるセッションにおいても推論の連続性を保ち、知見を累積（Reasoning Compounding）させることが可能となっている[1]。

ARC-AGI-3 の測定設計と競合アプローチの比較

François Chollet らによって創設された ARC-AGI-3 は、従来の静的な入力・出力グリッド変換（ARC-AGI-1/2）から、動的な対話環境への直接介入へと評価軸を拡張したベンチマークである[1]。ルールや目的が明示されない未知の環境において、エージェントは行動に対する状態遷移を観察し、背後にあるメカニズムとクリア条件を自力で推論して適応する必要がある[1]。
評価指標である Relative Human Action Efficiency（RHAE）は、初見の人間の上位中央値（Upper Median）のアクション数と AI のアクション数の比率の 2 乗としてレベルごとに計算される[1]。

$$\text{level\_score} = \min\left(1.15,\ \left(\frac{\text{human\_baseline\_actions}}{\text{ai\_actions}}\right)^2\right)$$

$$\text{Score}_{\text{Game}} = \frac{\sum_{\ell \in \text{Completed}} \ell \cdot \text{level\_score}_\ell}{\sum_{\ell=1}^{N_{\text{total}}} \ell}$$

この非線形な評価設計により、AI の行動数が人間基準の 2 倍であればレベルスコアは 25%、5 倍であれば 4%へと急減するため、非効率な総当たり探索（ブルートフォース）には強いペナルティが科される[1]。またゲーム全体の集計では難関な後半レベル（1 から始まるレベル番号 ℓ）に大きな重みが与えられるため、高スコアの獲得には全レベルの確実なクリアと高い行動効率の双方が求められる[1]。

システム	基盤 LLM	アプローチ分類	レベル達成率	RHAE スコア	総アクション数	知覚・入力インターフェース	アーキテクチャ特性

NVIDIA AVO	Claude Opus 5	汎用変異ループ＋自己監督	**100% (183/183)**	**100.00**	**6,624**	64×64 テキストグリッド[1]	直接対話型。GPU 最適化と同一の基幹ループを用い、VISTA 比で行動数を約 12% 削減[1]。
Tycho (Actor - Requested)	Claude Opus 5	プログラム的世界モデル	100% (183/183)	100.00	6,641	64×64 構造化データ[1]	実行可能な Python シミュレータを内部生成して検証・計画[1]。
Tycho (Actor - Requested)	GPT-5.6 Sol	プログラム的世界モデル	100% (183/183)	100.00	7,766	64×64 構造化データ[1]	アクターがモデル構築を専門エージェントに委託する動的パイプライン[1]。
VISTA	Claude Opus	直接視覚知覚	100% (183/183	100.00	7,542	512×512 PNG 画	高解像度画像

(MIT)	5.0	＋自然言語推論	)			像[1]	とロスレス視覚メモリによる直接知覚・推論[1]。
NVIDIA NOOA (Research Preview)	GPT-5.6 Sol	オブジェクト指向ハーネス	19/25 ゲーム完了	85.10	非公表	構造化データ / 型付き I/O[4]	型付き I/O と SQLite 長期記憶による高効率探索 ($13.3/game)[4]。
Claude Opus 5 (High) 単体評価	Claude Opus 5	単体呼び出し / 簡易プロンプト	不完全	30.16	計測不能	ARC Prize 標準設定[1]	ARC Prize による評価。記憶喪失と探索スタグネーションにより初期レベルで脱落[1]。

知覚インターフェースの設計において、MIT の VISTA が 512×512 の高解像度 PNG 画像を入力として直接的な視覚認識を行うのに対し、AVO は画像トークンを使用せず、環境状態を正確な 64×64 のテキストグリッド（数値・記号配列）として LLM に提示している点に重要な技術的差異がある[1]。

また、推論パラダイムとしては、環境の遷移規則を模倣する Python コードを合成して内部シミュレーションを行う「プログラム的世界モデル（Tycho 等）」と、観測状態から直接言語的に推論して行動を選択する「直接対話型（VISTA、AVO 等）」が存在する[1]。AVO は直接対話型の構成を採りながら、探索監督機構と履歴保持によって高い行動効率を維持し、Tycho

（6,641 アクション）とほぼ同等水準のアクション数（6,624 アクション、差は 17 アクション・約 0.26%）で全レベルを完答した[1]。

ハーネス設計の体系化と NOOA の設計原則

AVO の成果と並行して、NVIDIA はエージェントの実効性能を引き出すための実行基盤設計に関する研究成果を公表している[2]。エージェントフレームワーク「NOOA（NVIDIA-labs OO Agents）」では、基盤モデルの推論能力を損失なく活用するための 6 つの設計原則が提示されている[2]。

現代の大規模言語モデルは事前学習において膨大な Python コードを学習しているため、オブジェクトのメソッド呼び出し、型定義の解釈、例外処理といったプログラミングパラダイムを自然に扱うことができる[5]。NOOA はこの特性を活かし、エージェントの全インターフェースをプレーンな Python オブジェクトモデルとして構成している[4]。

ハーネス設計機能	実装メカニズム	性能・効率への寄与
Typed input/output	Pydantic 等の型定義による引数・戻り値の厳格なバリデーション[4]	構文解析エラーを抑制し、ツール実行の失敗率を低減[4]。
Pass by reference	巨大なデータダンプを避け、生オブジェクトの参照と境界付きプレビューを渡す[4]	トークン消費量を抑制し、コンテキスト窓の浪費を防止[4]。
Code as action	JSON 形式のツール呼び出しを排し、実行可能な Python コードを直接生成[4]	制御構文や複数ツールのインライン呼び出しを 1 ターンで完結[4]。
Programmable loop	オーケストレーション自体を編集可能な通常 Python として実装[4]	タスクの状況に応じた動的な探索ループの再構成を支援[4]。
Explicit object state	エージェントインスタンスのクラス属性として状態を保持[4]	会話履歴のコンパクションから状態管理を分離[4]。
Model-callable APIs	ハーネスのコンテキストやイベント履歴をモデルから呼べる API として公開[4]	エージェント自身による能動的な記憶の検索・整理・修正を可能化[4]。

なお、メモリ機構に関して、AVO が会話履歴・プロファイラ出力・推論ログの累積管理を行うコンテキスト保持機構を採用しているのに対し、NOOA では SQLite をバックエンドとした人間可読なリレーショナル永続メモリを採用している[1]。NOOA の検証では、SWE-bench Verified において GPT-5.5 を用いて 82.2%（当時のリーダーボード SOTA である 79.2%を更新）を記録しつつ、LLM 呼び出し回数を 29 回、消費トークン数を約 1.1M へと抑制したことが報告されている[4]。

産業応用・コンピュート需要・セキュリティ境界への影響

GPU カーネル自律最適化の実績

AVO の基盤アーキテクチャは、最先端のハードウェアエンジニアリングにおいて具体的な性能向上を達成している[1]。NVIDIA Blackwell（B200）アーキテクチャにおける Attention カーネルの最適化は、Tensor Memory Accelerator（TMA）の非同期データ転送、Warp Specialization によるスレッドグループの協調、およびオンライン Softmax 補正処理が絡み合う極めて複雑な工学課題である[1]。

AVO は人間の介入を挟まず、7 日間にわたって 500 以上の最適化方向を自律探索し、40 世代にわたるコードの反復改良を実施した[1]。レジスタ割り当てのチューニング、命令パイプラインのスケジューリング、ワークロード分散の再構成を自動で行った結果、BF16 精度において最大 1,668 TFLOPS を記録し、cuDNN を最大 3.5%、FlashAttention-4 を最大 10.5%上回るスループットを達成した[1]。さらに Multi-Head Attention（MHA）で得られた最適化知見は、約 30 分の追加自律探索によって Grouped-Query Attention（GQA）へと適応され、cuDNN 比 +7.0%、FA4 比+9.3%の性能向上を記録した[1]。

コンピュート需要への影響とエコシステム展開

自律エージェントの長期稼働は、推論インフラに対する計算資源需要の形態を変化させる可能性がある[2]。従来の対話型 AI がユーザーの単一入力に対して数百〜数千トークンを生成する形態であったのに対し、長期自律エージェントは 1 つの課題解決のために多数の推論ループを実行し、数百万〜数億トークンを連続消費する[1]。

このような自律探索ループの普及に伴い、推論ワークロード全体の計算需要拡大が見込まれている[2]。NVIDIA が自社モデルのみに依存せず、オープンなエージェント基盤やハーネス技術を提示しつつパートナーモデルとの連携を図る背景には、自律エージェント普及に伴う推論アクセラレータ需要を支えるプラットフォーム戦略が存在すると解釈されている[2]。

実行レイヤー分離とセキュアランタイム

エージェントの自律稼働が長期化するにつれ、セキュリティとガバナンスの確保が不可欠となる[4]。ハーネスのロジックがプログラマブルであり、エージェント自身がコードを生成・実行する構造において、プロンプトベースのガードレールのみに依存することは安全性の観点から不十分である[4]。

これに対し、NVIDIA はエージェントスタックにおいて推論・提案を行うレイヤーと、認可・

執行を行うレイヤーを分離するアプローチを提唱している[4]。エージェントおよびハーネスはアクションの「提案」を行い、ファイルアクセス、ネットワーク通信、API 実行などの実際の権限行使は、下層のセキュアランタイム（NVIDIA OpenShell 等）が最小権限の原則に基づいて決定論的に「執行」することで、長期自律環境における安全性を担保する構成が志向されている[4]。

限界事項と結論

公開セット評価における限界事項

AVO が記録した RHAE 100.00 は、あくまで公開評価セット（Public Set：25 環境・183 レベル）における結果である点に留意が必要である[1]。ARC-AGI-3 の本来の競技評価は非公開テストセット（Private Evaluation Set）で行われるものであり、今回の公開セットにおける成果が、完全な未知タスクに対する汎化性能を全面的に証明したと結論づけることはできない[1]。また、独立した研究者やコミュニティからは、公開セットのタスク特性や探索空間の性質上、体系的な反復行動や特定パターンの試行によって解に到達しやすい側面があるとの指摘もなされている[6]。真の流動性知能としての適応力については、今後の非公開セットでの客観的評価を通じた検証が待たれる[1]。

結論

NVIDIA AVO による ARC-AGI-3 公開セットでの成果は、AI エージェントのタスク遂行能力において、モデル単体のスケールだけでなく、記憶の維持、知覚インターフェースの適合、探索の停滞を回避する監督機構といった「ハーネス設計」が重要な役割を果たすことを示す有力な実証事例である[1]。

GPU カーネル最適化という極限の低レイヤ工学課題と、未知ルールの推論タスクという大きく異なる領域に対し、共通の変異・探索ループを適用して高水準の結果を導出したアプローチは、今後の長期自律エージェントのアーキテクチャ設計における重要な一角を占めるものと考えられる[1]。

引用文献

1. 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/
2. NVIDIA's AVO Harness Helps Claude Opus 5 Ace ARC-AGI-3, https://xenospectrum.com/en/nvidia-avo-harness-arc-agi-3/
3. NVIDIA「エージェント時代は、モデルの性能よりもハーネス設計が重要」, https://www.sbbit.jp/article/cont1/186645
4. Six Agent Harness Capabilities for Higher Model Performance, https://developer.nvidia.com/blog/six-agent-harness-capabilities-for-higher-model-performance/
5. I tried NVIDIA's new agent framework NOOA with a local LLM on, https://dev.classmethod.jp/en/articles/nvidia-nooa-local-llm-capability/

6. NVIDIA AVO Reaches 100% on ARC-AGI-3 : r/accelerate - Reddit, https://www.reddit.com/r/accelerate/comments/1vuh3hl/nvidia_avo_reaches_100_on_arcagi3/
7. 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