

Based on Primary Source Analytics | Date: 2026-08-23

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 で実載したと発表した¹。主張の核心は「モデル単体の性能ではなく、ヘーネス(エージェント設計)が鍵である」という命題であり、
- ・ ARC-AGI-3 の公開セットは、従来の AGI ベンチマークとは異なり、汎用性・柔軟性を重視した設計となっており、AVO の最適化利用が鍵となる。
- ・ 技術的な観点から、AVO の実装は、従来の AGI ベンチマークとは異なり、汎用性・柔軟性を重視した設計となっており、AVO の最適化利用が鍵となる。

NVIDIA AVOのARC-AGI-3 「100%」 達成^{*}：ノイズとシグナルの分離

事実関係の検証と、真の戦略的意義の分析

AGI到達の証明ではない。だが、汎用エージェント・アーキテクチャの強力な証拠である。

2. 事実

2.1 発表内容

NVIDIA Technical Blog, 2026 年 8 月 21 日公開(同目中に引用)。著者は Terry Chen(principal engineer)、Yeyin (Eva) Zhu、等。「ARC-AGI-3 の公開セットと単機設計・完全私設の競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」旨の限定が補填された経緯がうかがえる(注記の存在は事実、意図は推測)。

Executive Synthesis: 事象の本質



技術の所在 (The Engine)

性能の源泉は基盤モデル (Claude Opus 5) ではなく、その外側にある「AVO (スキャフォールディング)」の設計にある。モデルの再学習なしで、相互作用推論を劇的に引き上げた。



評価の限界 (The Arena)

100%達成はあくまで「デモ用公開セット (25環境)」のみの結果。真のベンチマークである「準私設・完全私設セット (110環境)」は未評価。ARC Prize公式は公開セットでの評価を否定している。



真の価値 (The Value)

最大のブレイクスルーは、「GPUカーネル最適化」のために組まれたループが、無関係なARCの空間推論に転移した点にある。NVIDIAの物理AIへの強力な布石。

The Claim vs. The Fact: ヘッドラインの解像度を上げる

The Claim (表面的なニュース)

“「NVIDIA AVOがARC-AGI-3で100%を達成」

“「Claude Opus 5を使用」

“「アクション数を12%削減」

The Fact (一次情報の真実)



対象は公開セット限定

準私設・完全私設セットの結果ではない。



モデル単体ではなくハーネスの成果

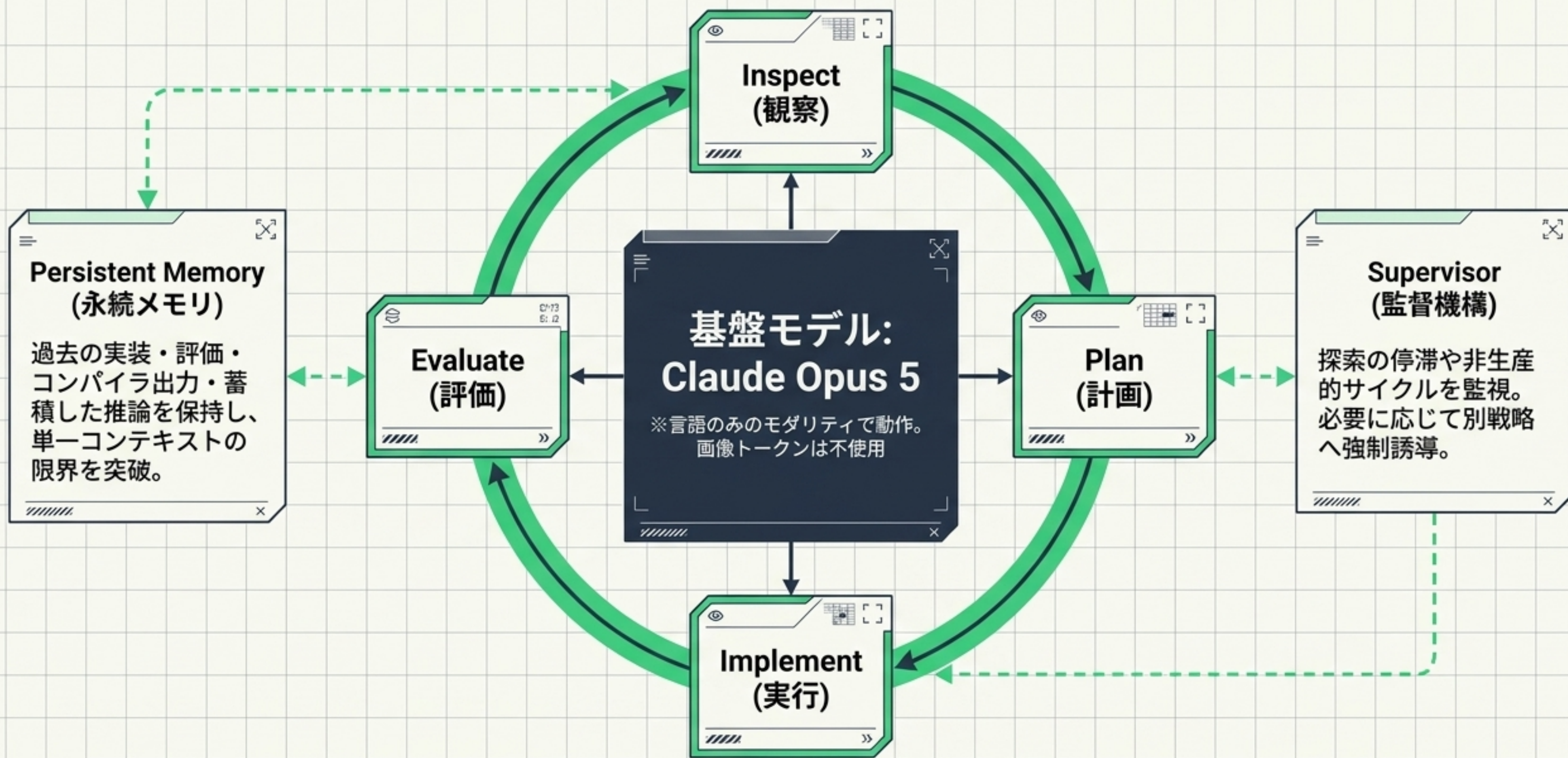
Opus 5の自己学習ではなく、外付けのエージェント設計 (AVO) によるもの。



100%は世界初ではない

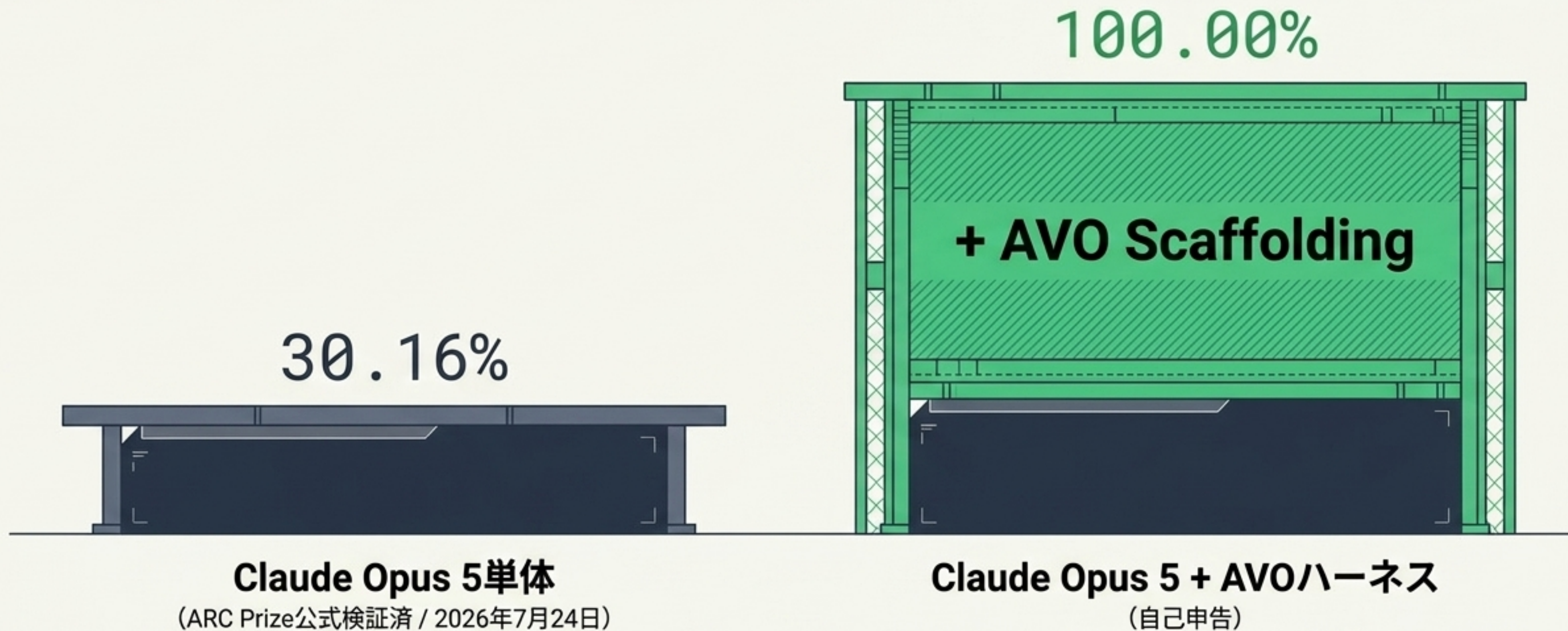
VISTAが先行達成済。AVOの新規性はアクション数の少なさ (6,624 vs 7,542) にある。

AVO's Engine: 進化的探索を自律化する反復ループ



The Secret Sauce: モデル能力を引き出す 「足場 (Scaffolding)」

同一の公開デモ25環境において、
基盤モデル単体とハーネス実装では
劇的なスコア差が生じる。



※モデルの追加学習 (Fine-tuning) は一切行われていない。AIの真の性能ボトルネックは、モデルの重みではなく「オーケストレーション層」にあることを強く示唆している。

The "Public Set" Illusion: 氷山の一角にすぎない「100%」

Public Demo Set
(公開デモセット)

25環境 (183レベル)

- AVOが100%を達成した領域。
- ARC Prize公式見解：訓練資源ではなく、デモ用インターフェース

Semi-Private / Fully-Private Set
(真の評価領域)

110環境 (準私設 55 + 完全私設 55)

- 未評価・スコア未公開の領域。
- ARC Prize公式見解：真の汎用知能 (AGI) への進捗を測る唯一の領域。

Key Insight:

ARC-AGI-2 (公開10：私設1) から設計が反転。
公開セットは設計上、容易に100%が取れる
よう想定されたデモセットに降格している。

Scoring Mechanics: 総当たりを許さないRHAEのペナルティ

The Formula

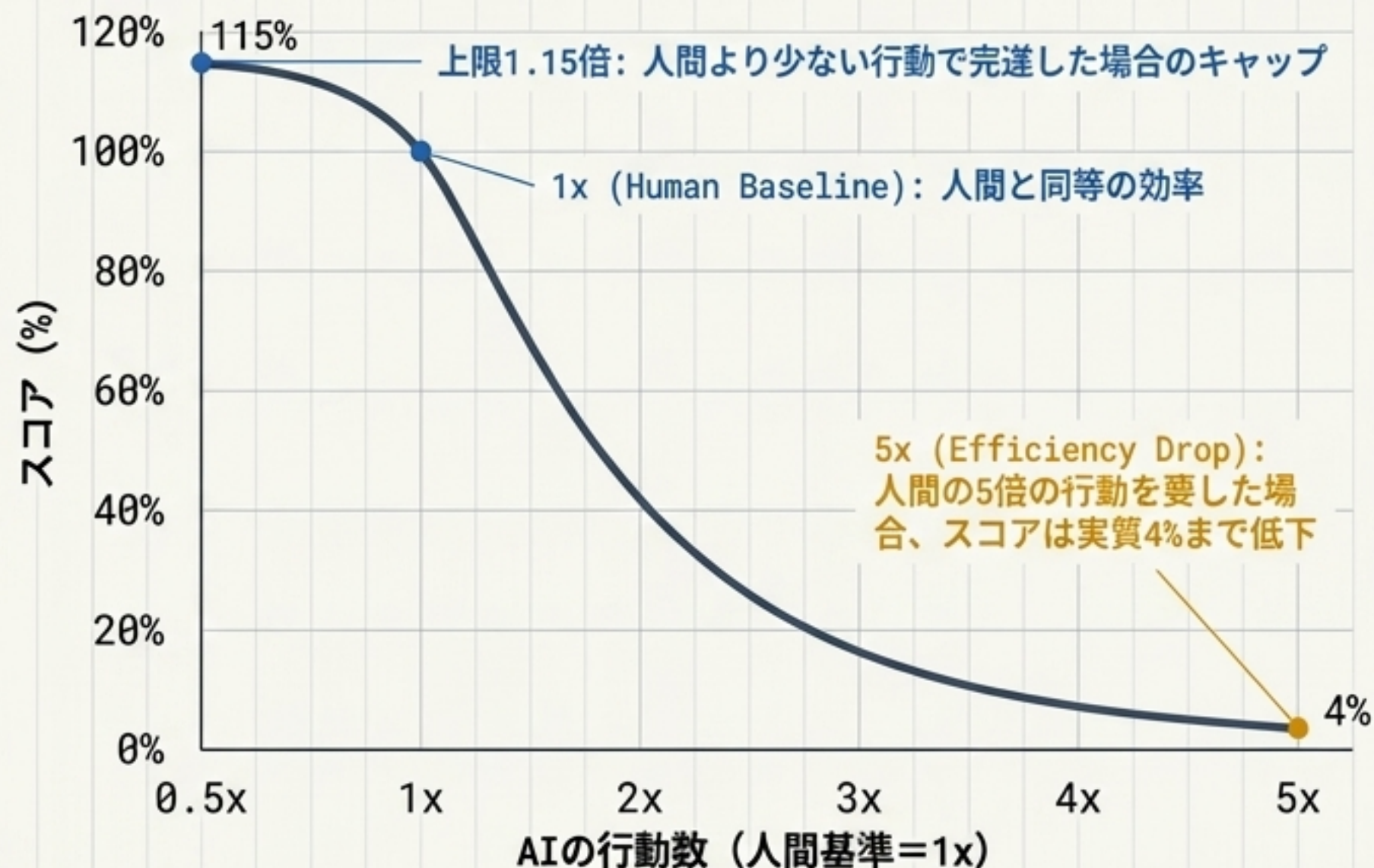
$$\text{レベルスコア} = (\text{人間ベースライン行動数} \div \text{AI行動数})^2$$

※人間の行動数=複数初見テストのUpper Median

Key Insight

AVOの新規性は「100%完遂」自体ではなく、RHAEルール下においてVISTAよりも総アクション数を約12% (7,542 → 6,624) 抑え込んだ効率性にある。

The Efficiency Curve



The Competitor Matrix:

先行・競合事例との厳密な比較

システム / モデル	評価対象セット	スコア / 行動数	検証状況
NVIDIA AV0 (Opus 5)	公開25環境	RHAE 100.00 / 6,624 Actions	自己申告 (未検証)
VISTA (Opus 5)	公開25環境	RHAE 100 / 7,542 Actions	自己申告 (未検証)
Schema (Opus 4.8+Fable 5)	公開セット	98.98%	自己申告 (未検証)
Claude Opus 5 (Bare model)	公開25環境	30.16%	ARC Prize検証済
GPT-5.6 Sol (Modified settings)	公開セット	38.3%	OpenAI自己申告

Takeaway: 公開セットでの100%・99%は既に飽和状態。AV0はVISTAの「Direct-Interaction」原則を採用しつつ、独自ループにより行動効率を最適化したモデルと言える。

ARC Prize's Official Stance: タスク特化の過学習 (Overfitting)

The Theory

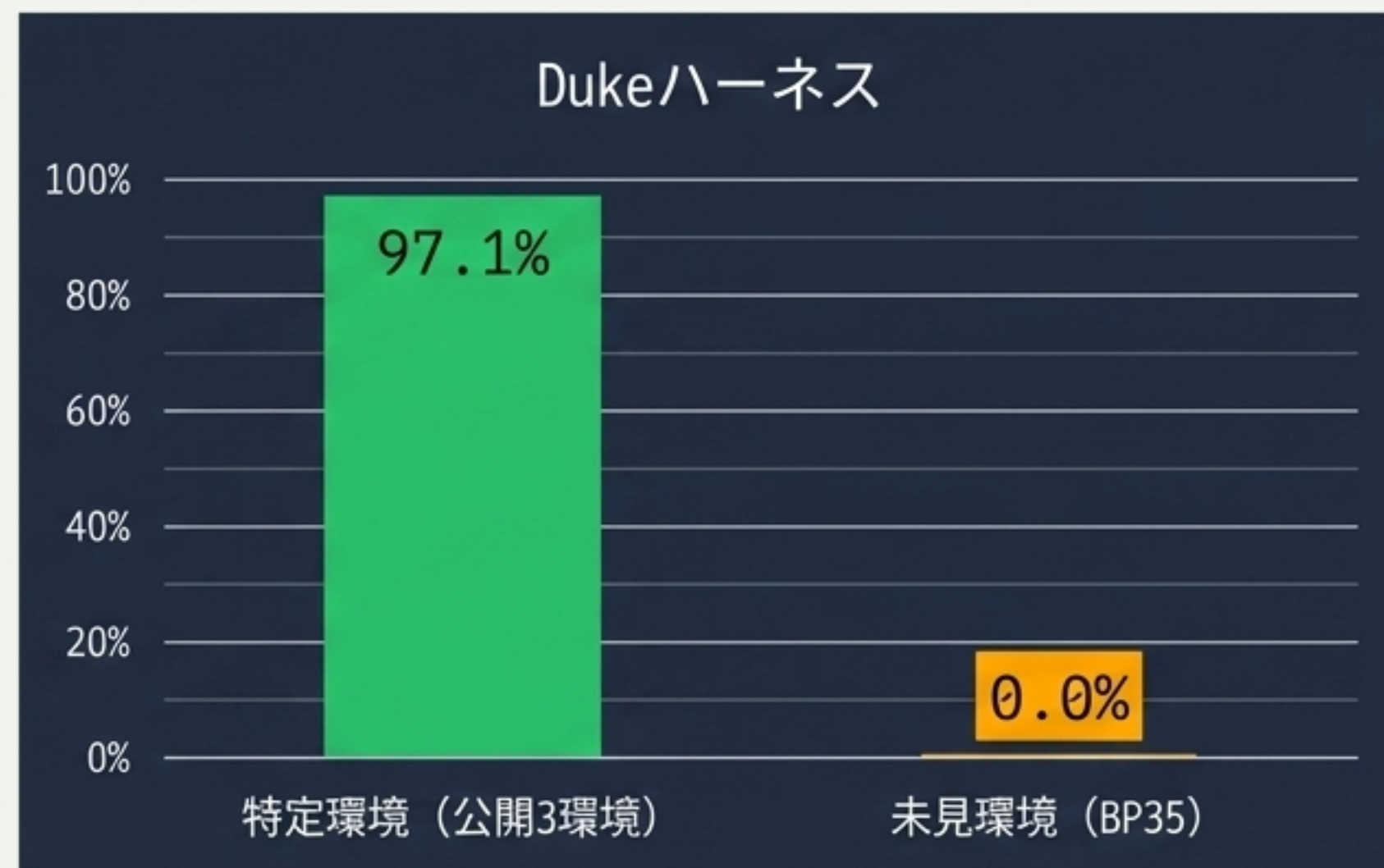


Official View:

「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る。」

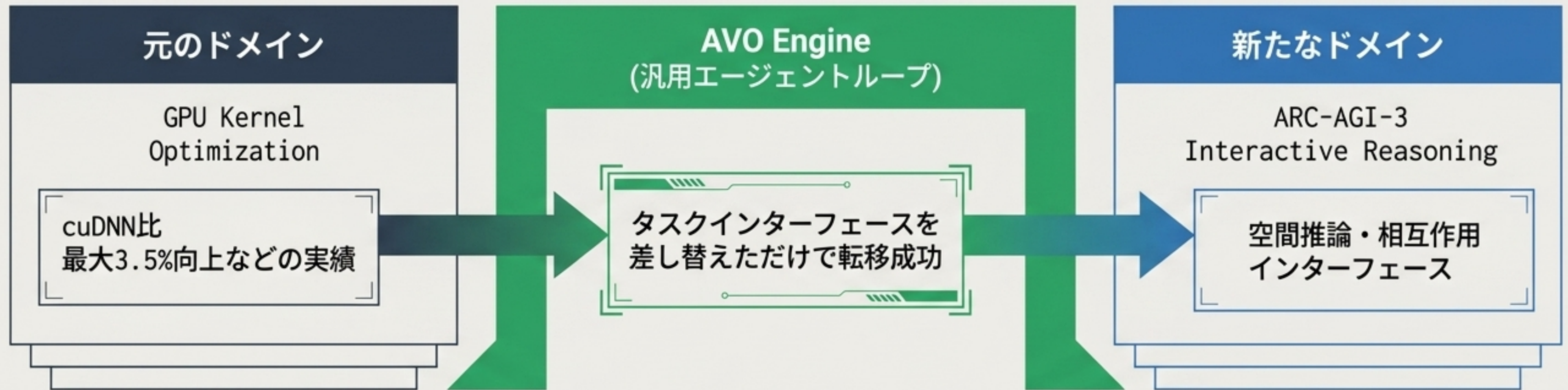
「公開セットのスコアは公式リーダーボードには掲載しない。AGIへの進捗の妥当な指標ではない。」

The Generalization Wall (二峰性の実証データ)



Conclusion: 公開環境の知識をもって作られたハーネスは、特定のパズルへの過学習 (Task-specific overfitting) に陥るリスクが高い。

The Real Breakthrough: 無関係なドメイン間の「構造転移」



The Core Argument

AVOはARC特化のワールドモデル層を一切持ち込んでいない。

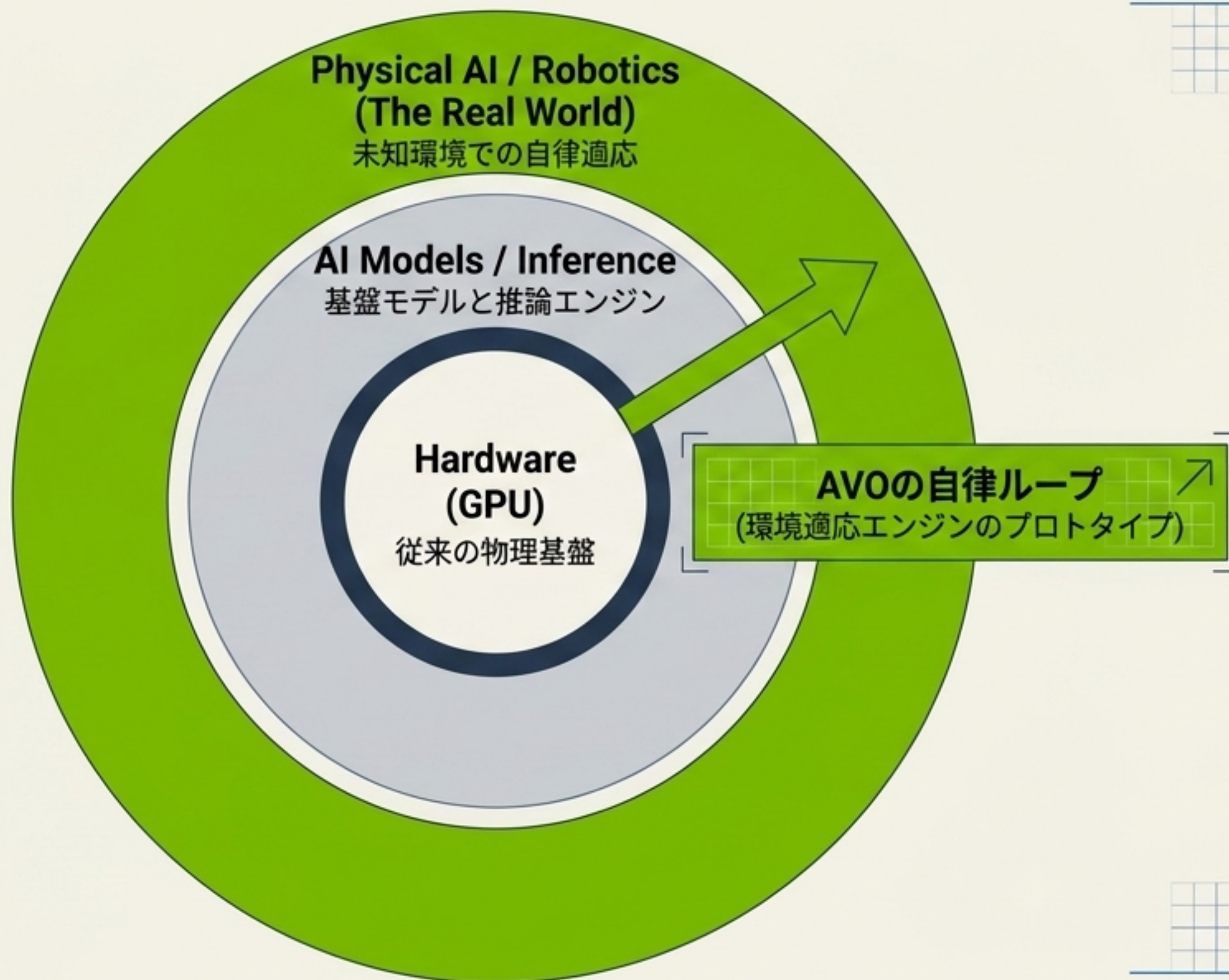
数ヶ月前に「GPUカーネルの最適化」のために設計された自律進行ループが、そのまま機能した。

Why It Matters

転移した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**」。

これこそが、長期ホライズンにおける汎用アーキテクチャの実証である。

NVIDIA's Strategic Blueprint: なぜNVIDIAがこれをするのか？



フルスタックへの布石

ハードウェア企業から、AIの「フルスタック企業（モデル・アーキテクチャ・エージェント）」へ向けた展開。

Physical AIへの直結

ARC-AGI-3の「初見環境の探索・適応」という課題構造は、現実世界のロボティクス制御と極めて近い。
未知のルールに対し、事前知識なしで推論し適応する自律ループの証明。

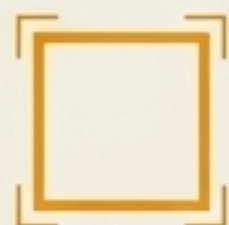
Milestones to Watch: 今後注視すべき「真の進捗閾値」

Radar Checklist



私設セットでのVerified Score

AVOまたは後続システムが、準私設/完全私設セットで高いスコアを出せるか。ここが汎用性の最終証明となる。



コストとE2E時間の開示

ARC Prizeがコストを評価軸に据える以上、フル評価におけるコスト効率の証明が必須。



GPT-5.6 Solとの体系的比較

「Solは実時間で速く、Opus 5は行動数で少ない」という予備的観察のさらなる検証。



オープンソース化の動向

NVIDIAがAVOをコミュニティに公開するか。

Final Takeaway: ハーネス設計に残された改善余地は依然として大きい。
AIの性能限界は、もはやモデル単体ではなく、システム全体の設計によって定義される。