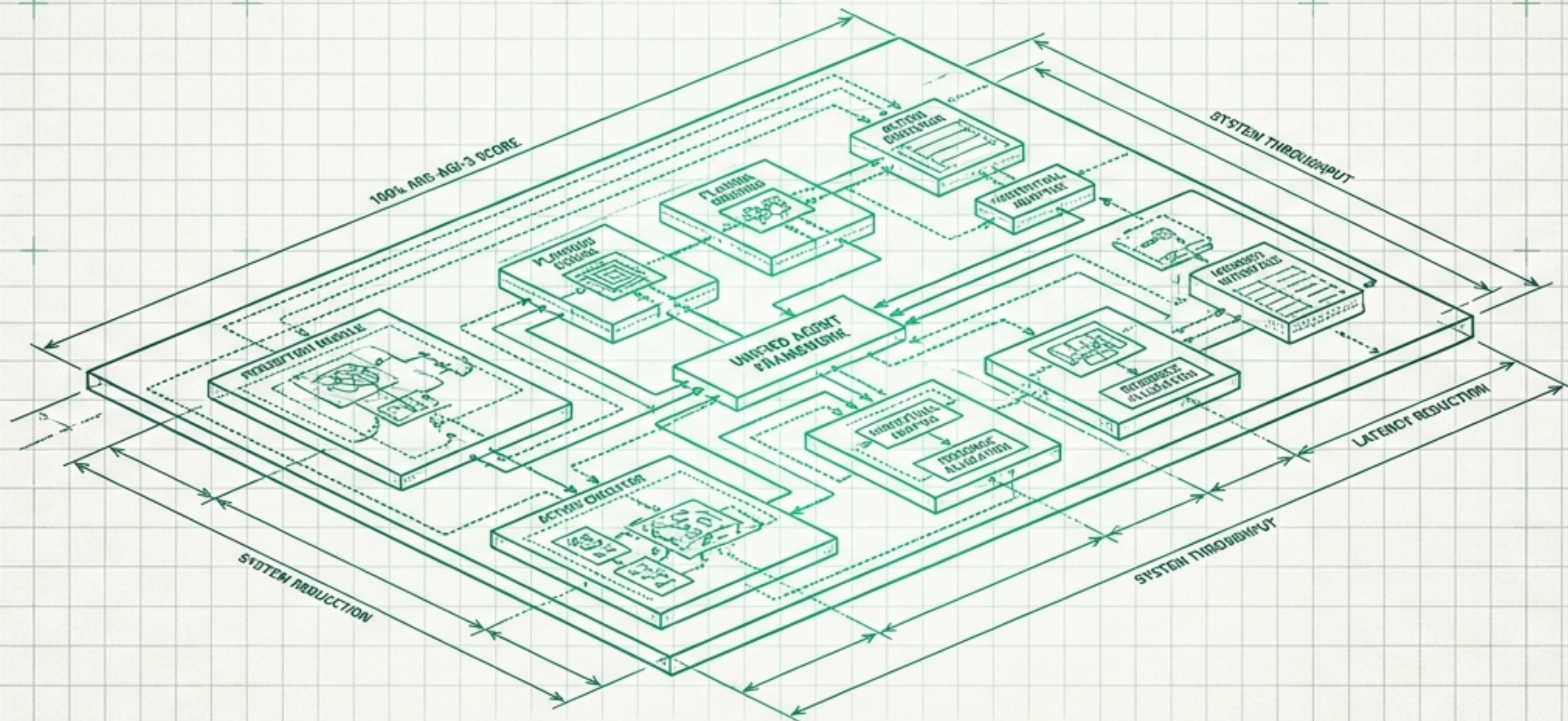


NVIDIA AVOによるARC-AGI-3「100%」の真の技術的意味

「AGIへの到達」ではなく、「基盤モデル」から「エージェント・システム工学」への決定的なパラダイムシフト



An Analytical Explainer



0 1 50M

Executive Summary - 3つの重要インサイト



The Achievement (偉業)

NVIDIA AVO + Claude Opus 5
がPublic Set (183レベル) を
完走。

環境アクション数6,624で
100.00 RHAEを達成。VISTAよ
り約12%高効率。



The Core (本質)

新しいAIモデルではなく、永続
メモリとスーパーバイザーを統
合した「エージェント・ハーネ
ス」の勝利。

GPUカーネル最適化のループを
未知のゲーム環境へそのまま移
植したアーキテクチャ。



The Reality (限界と留保)

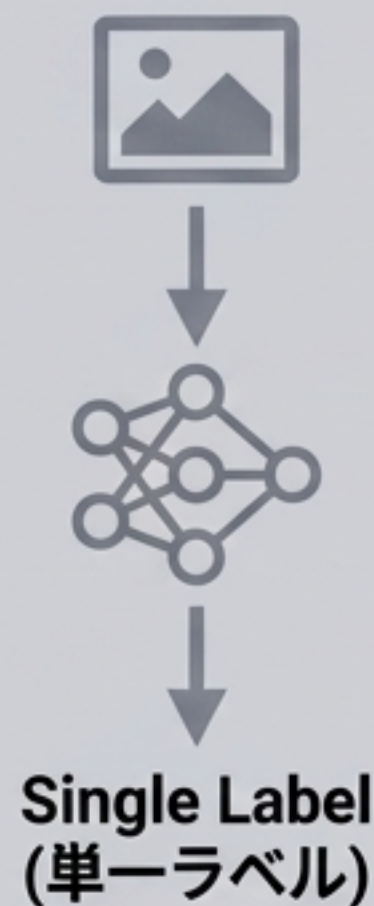
完全非公開のPrivate環境での
未検証。

見かけの効率の裏にある莫大な
内部推論 (Hidden
Compute) と、学習データ汚染
の懸念。

ARC-AGI-3が測るもの：静的認識から対話的推論へ

初期のフロンティアモデルが1%未満に留まった難関環境の真の要求事項

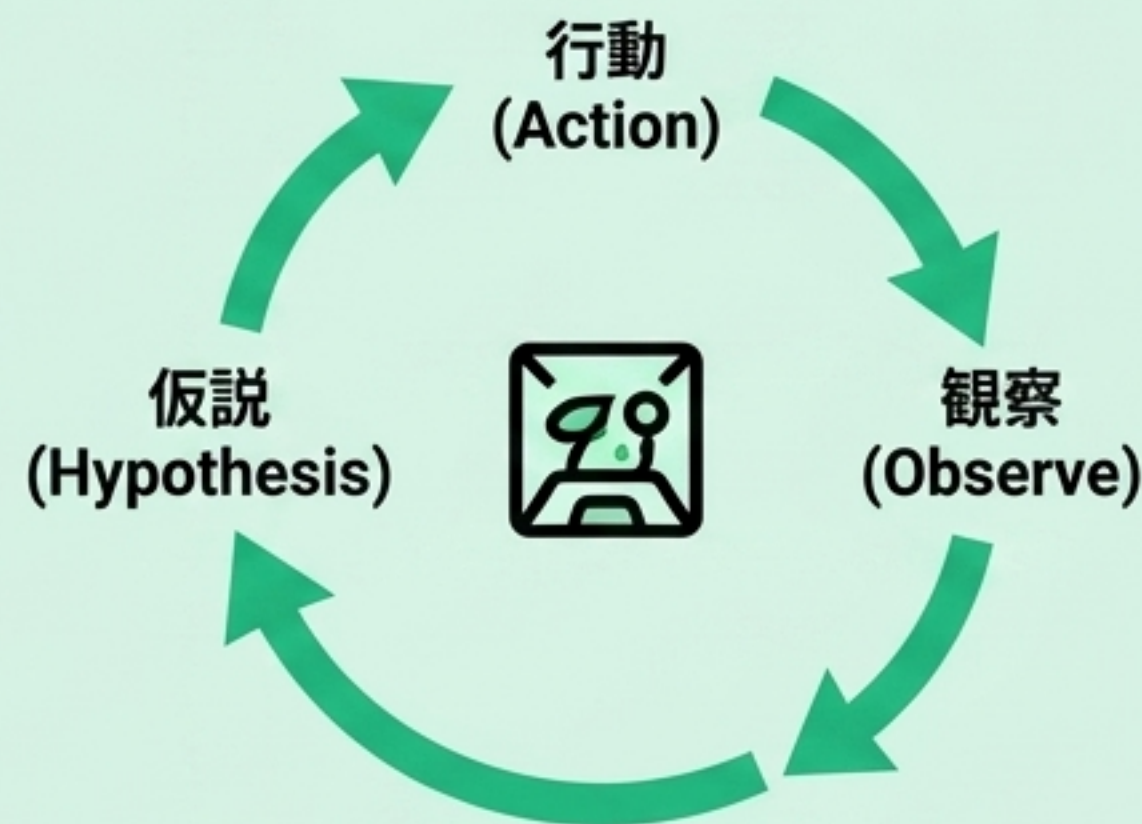
Static Perception (静的認識)



従来のVLMベンチマーク

画像を一度だけ解析し、正解を出力する。

Interactive Reasoning (対話的推論)



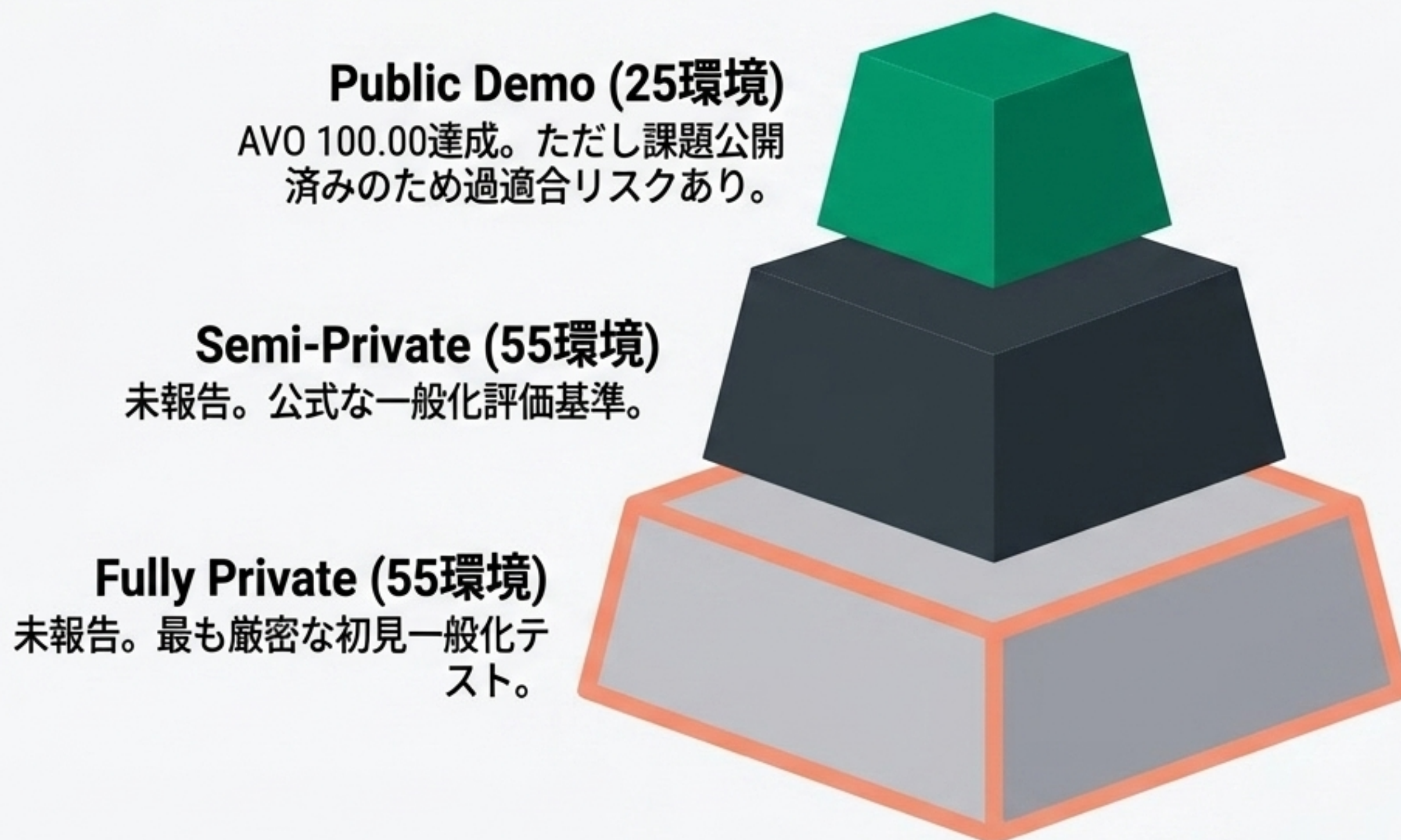
ARC-AGI-3環境

未知の因果関係に対し、少数のインタラクションからルールを推論し、スキルを獲得する能力を測る。

ゲーム固有のルールや自然言語による目的説明は一切与えられない。

The 100% Illusion - 公開セットと非公開セットの壁

RHAE 100は「全レベルでの人間超え」を意味しない。高効率レベルによる相殺を含む完走の証明である。

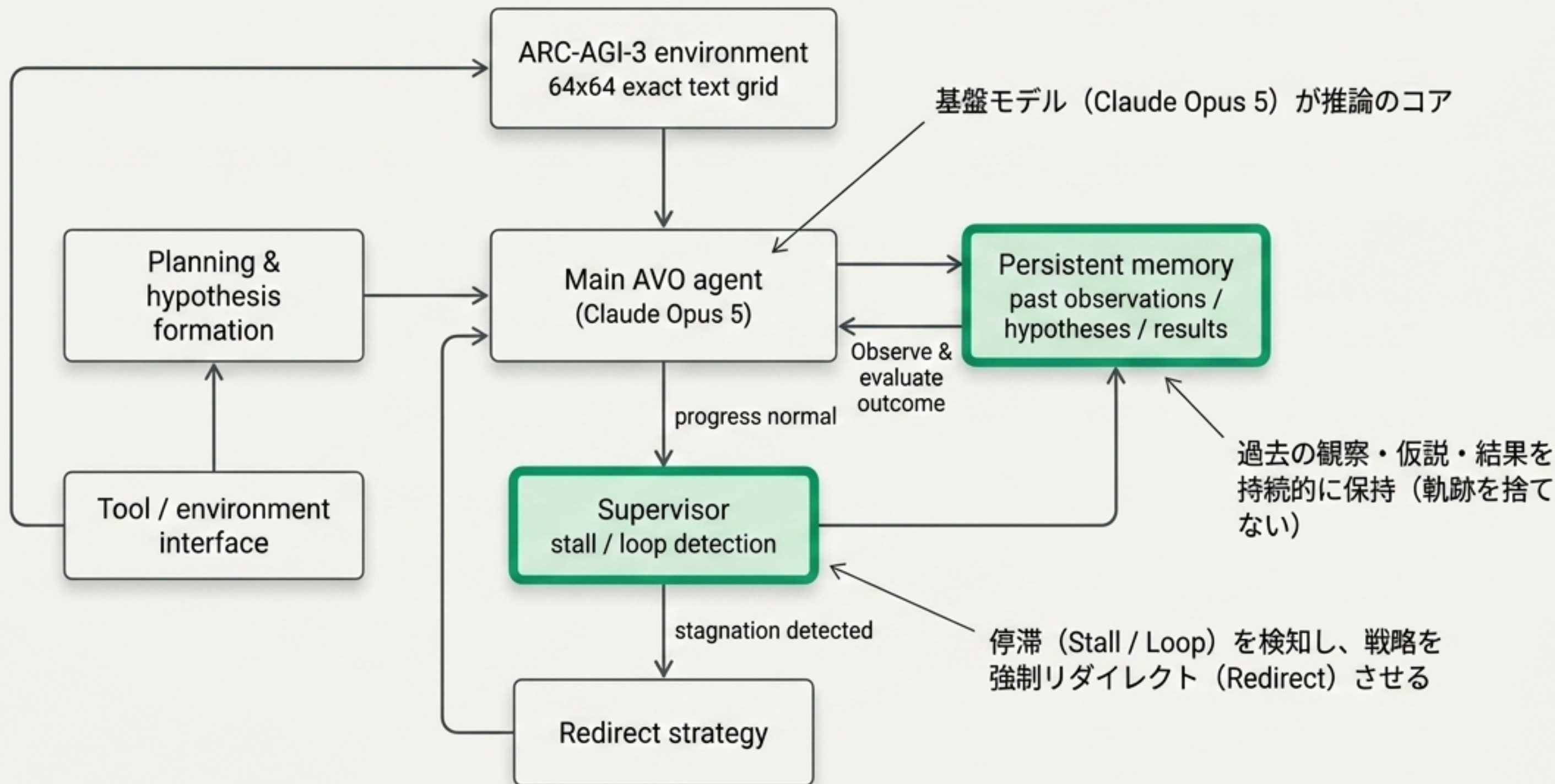


Insight

公式な **Benchmark Solved** には**Private Set**の攻略が不可欠である。
Public Set 100%は、あくまで公開環境への高度な適応を示すものに過ぎない。

AVO Architecture - エージェント・ハーネスの構造

新しいモデルではなく、汎用のシステムループ。CUDA最適化の進化探索アーキテクチャを未知のゲームへ接続。



What AVO is NOT - 100%を支える技術の誤解と真実

視覚入力 (Multimodality)



画像認識 (VLM / PNG) による空間把握



正確なテキストグリッド (64x64 Text Grid) のみの記号論理

モデルの学習 (Learning)



強化学習 (RL) や模倣学習 (IL) によるモデルの重み更新



テスト時のプロンプト/コンテキスト内での状態更新 (Test-time memory)

環境シミュレーション



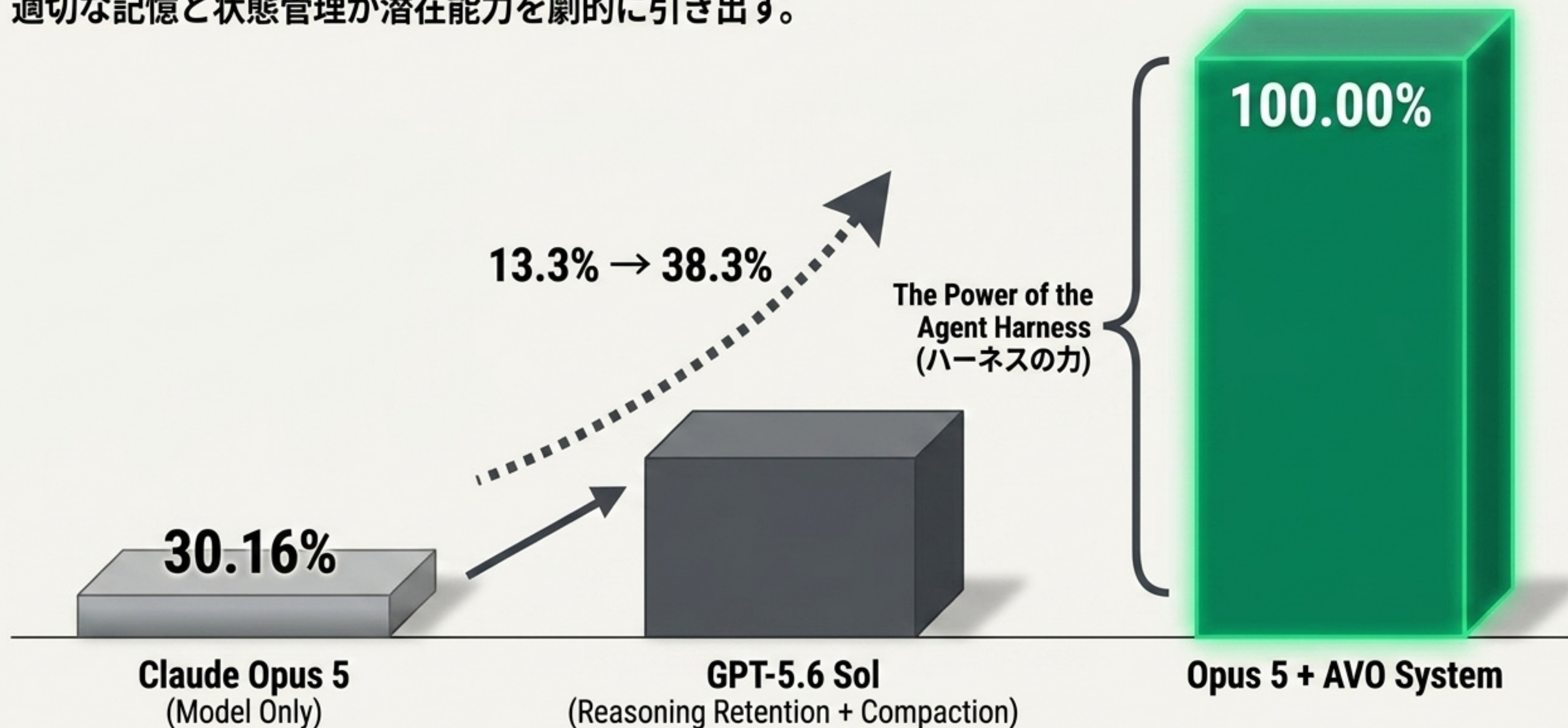
明示的なPythonシミュレータによる世界モデルの構築



LLMによる直接的な自然言語での因果推論

The Paradigm Shift - 「モデル能力」から「システム工学」へ

AIの能力は「モデル×ハーネス×メモリ×コンピュート」の積になった。
適切な記憶と状態管理が潜在能力を劇的に引き出す。



SOTA Comparison - 100%到達システムのアーキテクチャ比較

特徴	AVO	VISTA	Tycho	Twin
主観測 (Perception)	64x64 Text grid	512x512 PNG	Text grid / Executable	Text grid / Workspace
明示的世界モデル (World Model)	なし/自然言語	なし	あり/Python	あり/Digital Twin
スーパーバイザー (Meta-control)	あり	なし	Explore/Plan loop	Validation replay
アクション数 (Efficiency)	6,624	7,542	6,641	11,614

The Missing Pieces - 存在しないアブレーション

AVO固有のコンポーネント（MemoryやSupervisor）の因果効果は厳密には分離評価されていない。

$$100 = \text{Opus 5} + \text{Memory} + \text{Supervisor} + \text{Text Grid} + \text{Context}$$

Known
Base

?

?

?

?

Independent Validation vs Missing Data

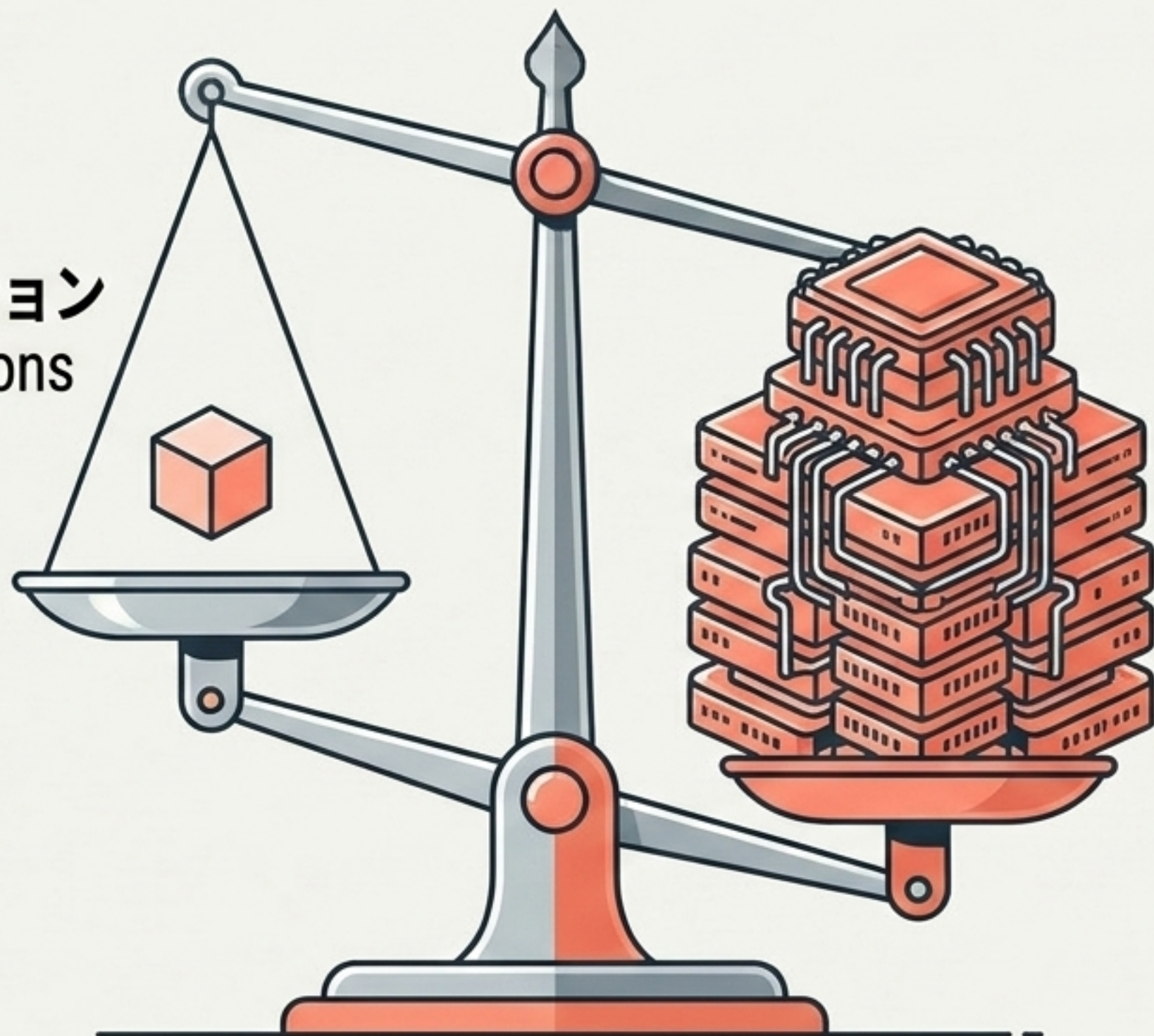
TwinやOpenAIは、ハーネスの有無
(例: 93.3 → 61.1) による効果を独立して
証明している。

しかし、AVOの実験では、永続メモリ単体の
効果、画像入力にした場合の変化、スーパ
バイザーを外した際のスコア低下などのアブ
レーション・データが公開されていない。

The Hidden Compute - 環境効率 vs 計算効率の罠

ARCのRHAE指標は、内部推論（CoT）やLLM呼び出しをアクションコストにカウントしない盲点がある。

環境アクション
6,624 Actions



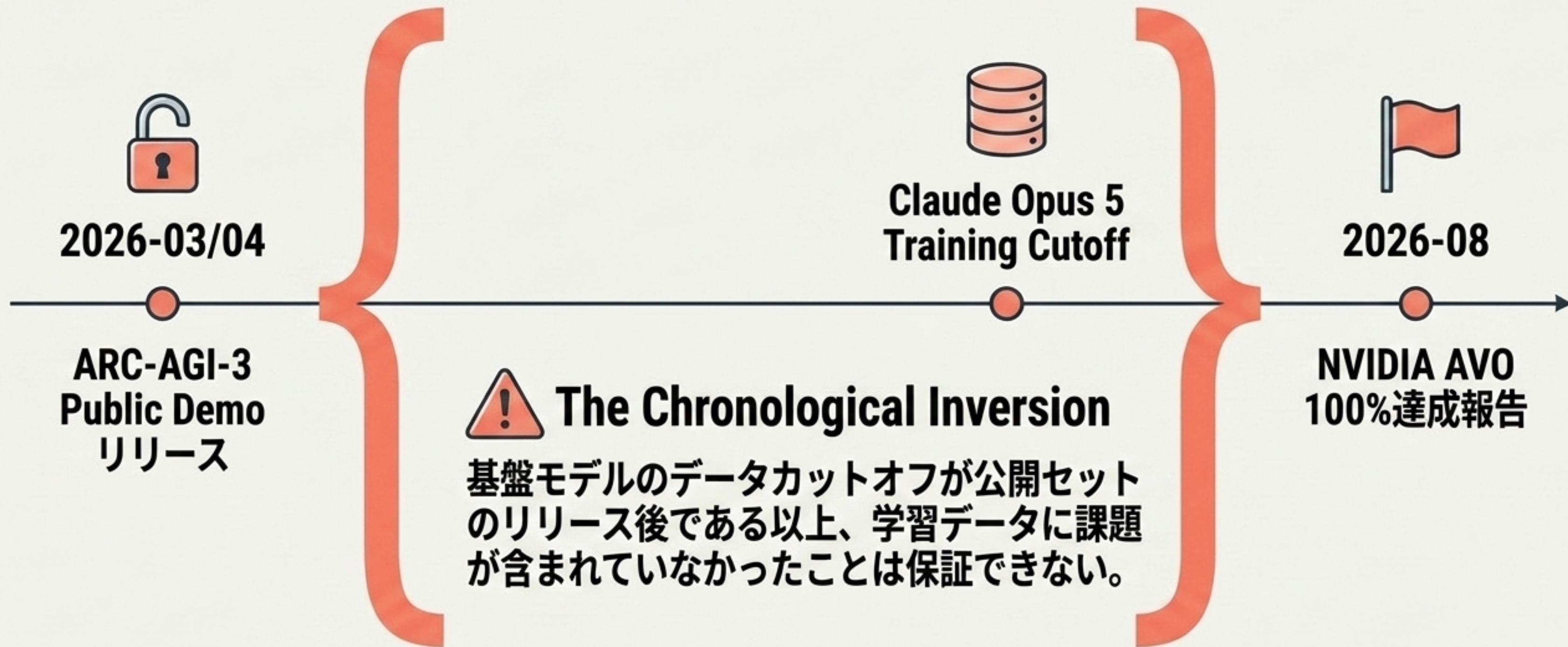
Evidence from Twin

Twinは環境アクションを半減させる裏で、トークン消費を2.5倍（2.60B tokens, 91.4h）に増大させた。
アクション数の少なさだけでシステムの真の効率は測れない。

見えない推論コンピュータ
（Hidden Compute）
数千ドル規模のAPI推論
数億単位の消費トークン

Contamination Risk - データ汚染という最大の留保

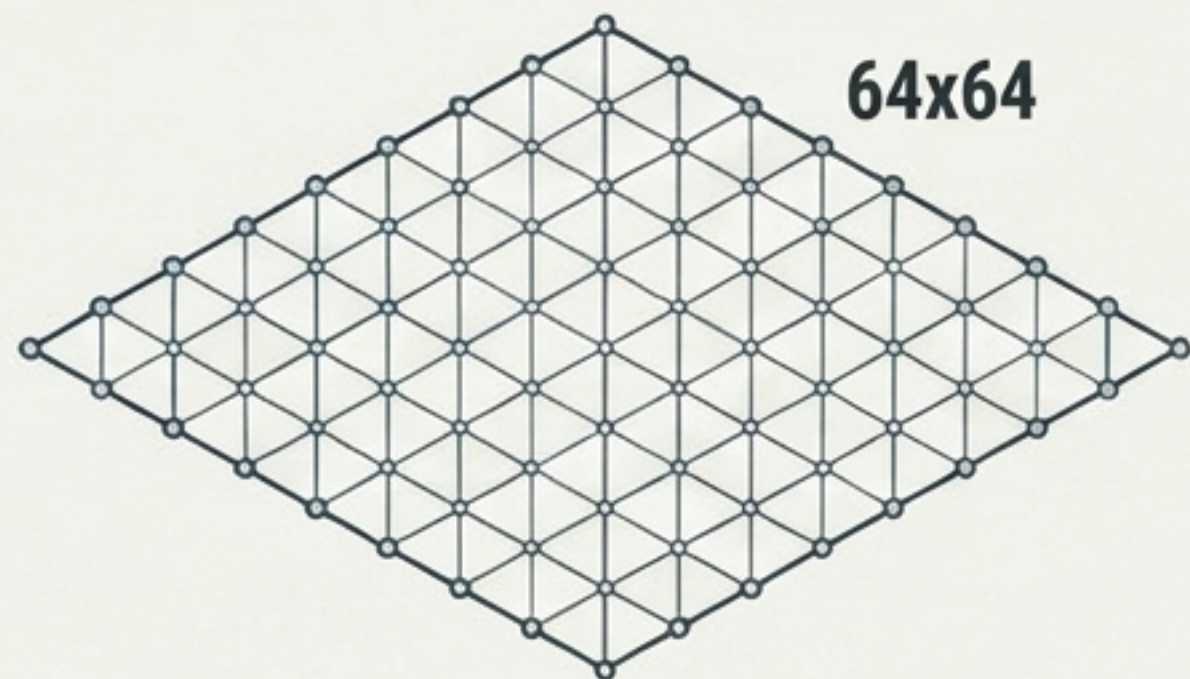
Public 100%を真の初見一般化と断定できない致命的な理由



Real-World Implications - 物理世界への移植と障壁

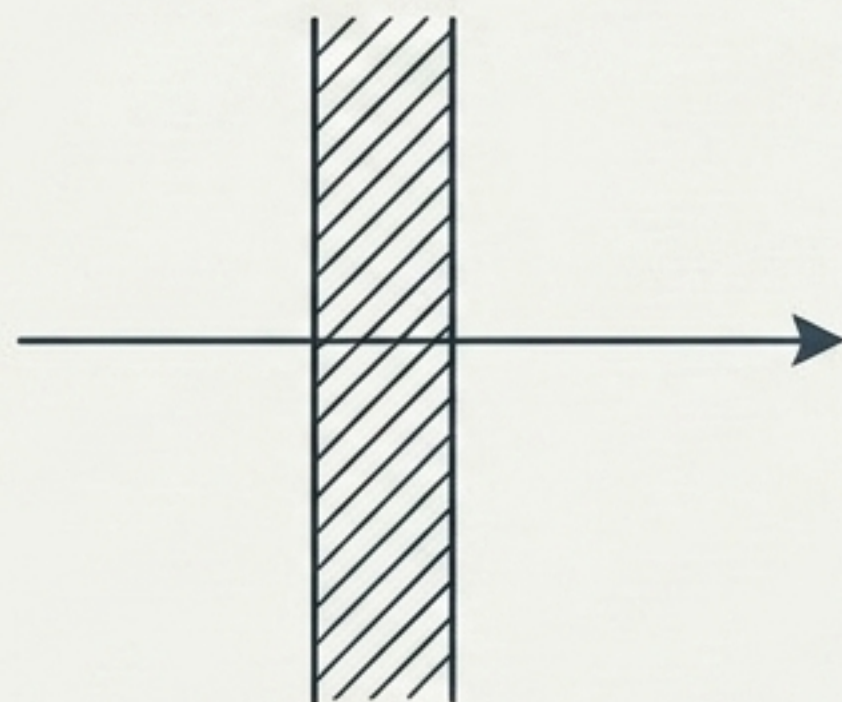
強力な設計原理と、実世界の連続的なノイズ環境との間のギャップ

ARC Grid



離散・抽象・ターン制。試しに押す探索 (Active Exploration) が許容される安全な空間。

Safety Envelope
(安全性の障壁)



ハードな実行制約とランタイム保護

Real World



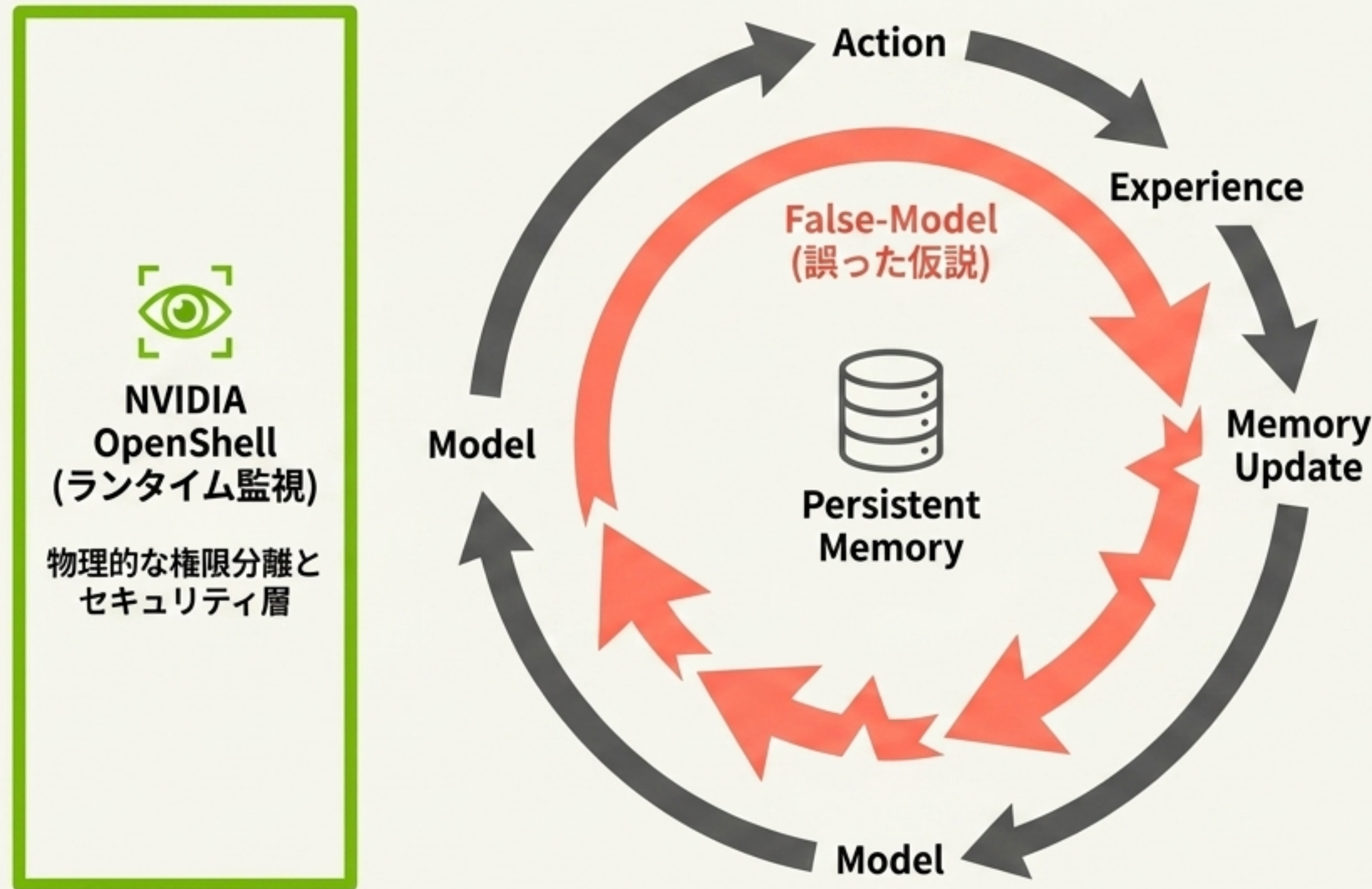
視覚ノイズ、連続制御、不可逆な物理的損傷。実行権限とポリシーを分離する物理的安全層が必要不可欠。

Transferable Insight

過去の失敗を永続状態に保存し、環境フィードバックで仮説を更新し続けるアーキテクチャ自体は、ロボティクスにも極めて有効である。

Agentic Safety – 自律エージェント特有の脆弱性

永続メモリは、正しい学習だけでなく、誤った仮説を長期間保持・増幅する装置にもなる。



⚠ Risk: LLM Supervisor Limits

メインエージェントとスーパーバイザーが同じ基盤モデルに依存する場合、エラーは独立しない。

✅ Solution: Hard Containment

モデルの自己監視に依存せず、実行権限やネットワークアクセスをランタイムで物理的に制限する層が不可欠。

The New Frontier - 今後の評価基準と未解決問題

ARC-AGI-3 Public Setが飽和した今、研究コミュニティが直面する次のハードル。



1. Blind Private Evaluation

完全非公開環境でのテストによる、データ汚染の排除と真実の汎用性の証明。



2. Factorial Ablation

メモリ、スーパーバイザーなど、コンポーネントごとの厳密な因果効果測定。



$$\text{Utility} = \alpha[\text{tokens}] + \beta[\text{time}] + \gamma[\text{dollars}]$$

3. Compute-Aware Scoring

計算コストを考慮した新指標の導入 ($\text{Utility} = \alpha[\text{tokens}] + \beta[\text{time}] + \gamma[\text{dollars}]$)。



4. Noisy/Continuous Environments

ノイズを含む連続環境や、不完全情報下でのハードな実行制約の検証。

Final Verdict - 最終評価

AGI到達の証拠としては限定的 (Limited)

非公開テストの欠如とデータ汚染の懸念により、真の初見一般化とは断定できない。

エージェント・システム工学の証拠としては極めて重要 (Highly Significant)

モデル単体では30%の能力を、ハーネスの設計次第で100%まで引き出せることを実証した。

Conclusion: The New Scaling Axis

NVIDIA AVOの最大の功績は、次のAIスケーリング軸が「パラメータ数」から、「テスト時の状態管理、永続記憶、ツールの統合、メタ制御」へと完全に移行したことを決定づけた点にある。