パフォーマンスの絶壁と新たな北極星

ARC-AGI-3が暴く「次トークン予測」の限界と、
真の汎用人工知能（AGI）へのパラダイムシフト

2026年3月・評価基準の完全なる転換

幻想の終焉：フロンティアモデルを襲った「パフォーマンスの崖」

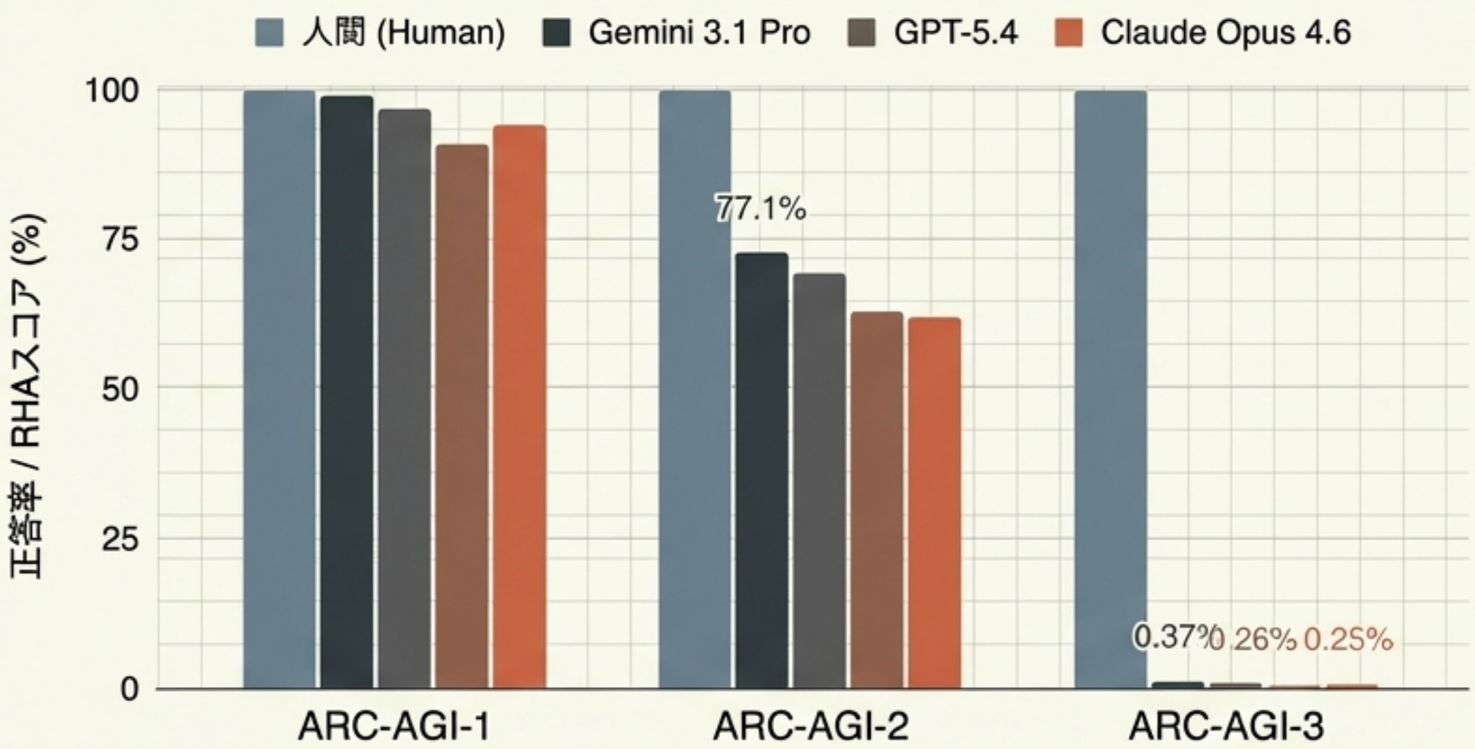
100% vs 1%未満。

人間（初見プレイ）：100.0%

Gemini 3.1 Pro: 0.37%
(ARC-AGI-2: 77.1%からの急落)

GPT-5.4: 0.26%

Claude Opus 4.6: 0.25%



世界最高峰の計算資源を誇るLLM群は、未知の環境において1%の壁すら超えられなかった。
我々は「AGIの入り口」にはいなかったのである。

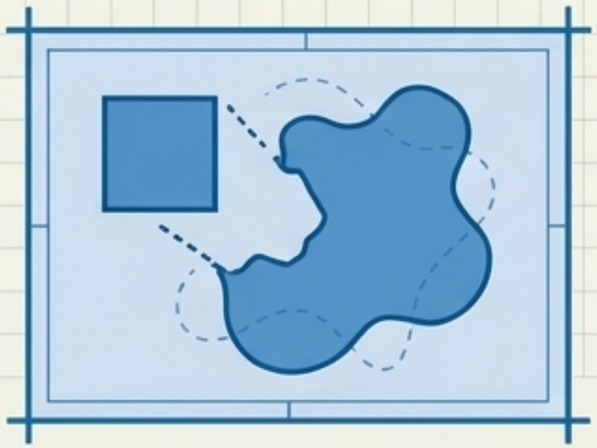
知能の再定義：静的パターンの暗記から、動的なスキル獲得へ

過去の評価基準 (ARC-AGI-1 / 2)		新たな評価基準 (ARC-AGI-3)	
パラダイム	パターンの暗記と検索 (Memorization-and-retrieval)	**パラダイム**	適応的行動 (Adaptive behavior) と効率
環境	静的・バッチ処理	**環境**	動的・対話型 (POMDP)
測られるもの	結晶性知能（事前知識への依存）	**測られるもの**	流動性知能 (Fluid Intelligence)
成績	80〜90%超（飽和状態）	**成績**	**1%未満**

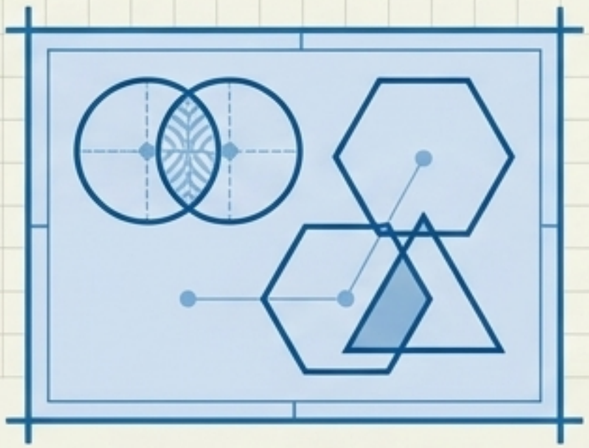
「知能とは、事前知識や経験を、不確実性を伴う価値あるタスクにおける新しいスキルへと変換する効率（Skill-acquisition efficiency）である」 — François Chollet (2019)

文化と知識の剥奪：純粋な「推論」だけが残る空間

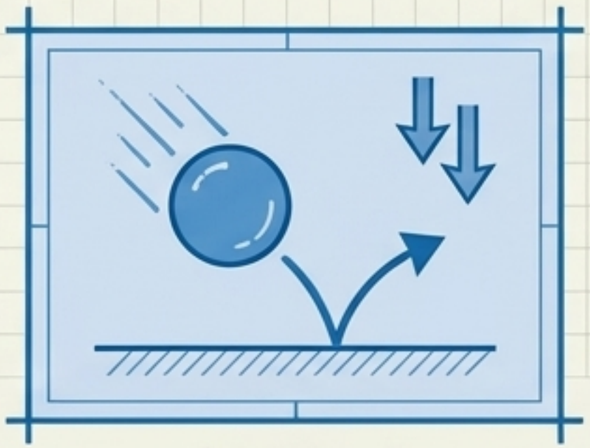
ARC-AGI-3には、言語も、数字も、現実世界のオブジェクトも存在しない。テストは以下の4つの生得的な「コア知識（Core Knowledge priors）」のみを前提に手作りされている。



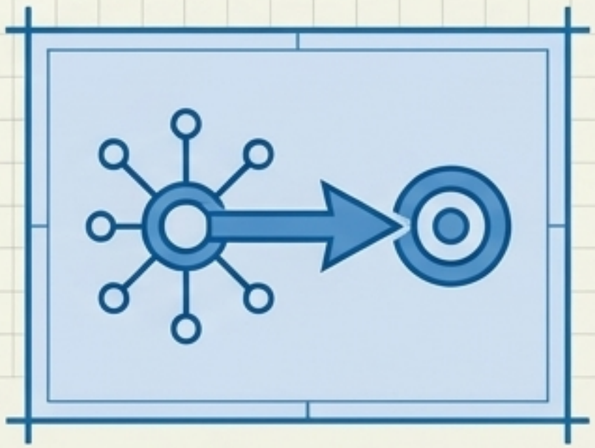
1. 物体性 (Objectness): 物体が独立した実体として存在し、まとまりを保つ。



2. 幾何学と位相幾何学 (Geometry/Topology): 対称性、接続性、包含関係の理解。

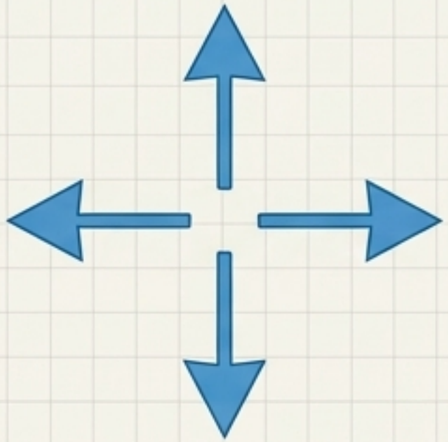


3. 基本的な物理法則 (Basic Physics): 重力、運動量、反発などの直感的な力学。



4. エージェント性 (Agentness): 存在が意図を持ち、目標に向かって行動する認識。

アクション空間の非対称性：力技（Brute-force）の無効化



【狭いアクション空間】

- 環境例: ls20 (見えない内部状態のナビゲーション)
- 操作: 方向キーのみ
- 1ステップの選択肢: 4通り

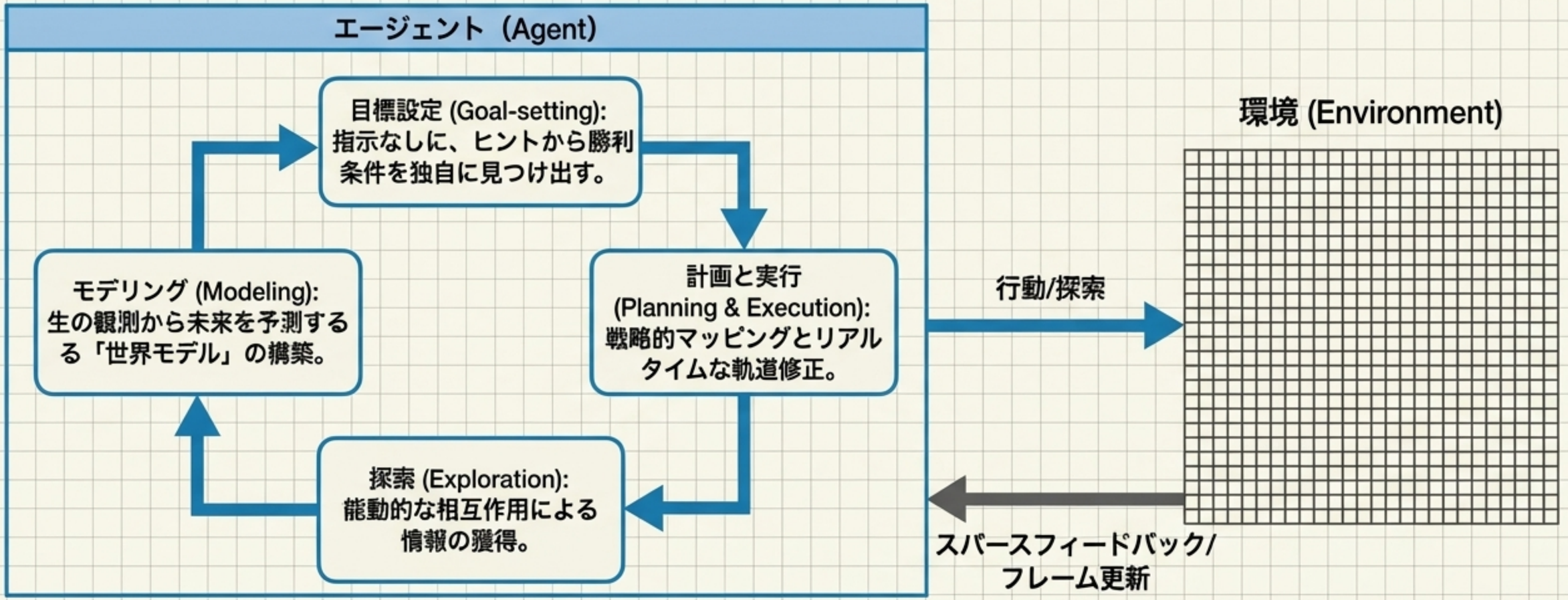
【巨大なアクション空間】

- 環境例: vc33 (閾値と予算管理) / ft09 (抽象的パターンマッチ)
- 操作: 任意のピクセルをクリック
- 1ステップの選択肢: 4,096通り

ランダムな探索や総当たり攻撃による攻略は数学的に不可能。
視覚情報から意味のある領域を抽出し、意図を持ってアクションを選択しなければならない。

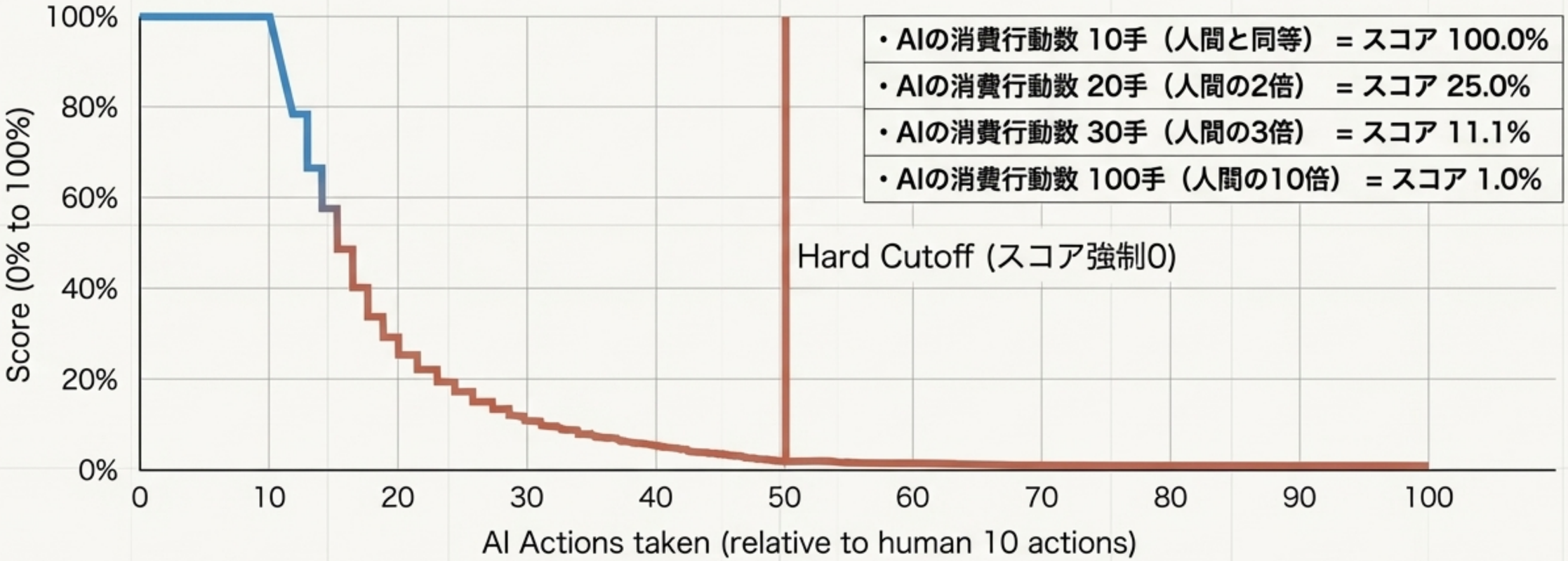
求められる4つのコア能力：対話型推論ループ

報酬関数すら未知のPOMDP（部分観測マルコフ決定過程）空間において、極めて希薄なフィードバックのみでこのループを回し続ける必要がある。



RHAEスコア：知能の「コスト」を測る冷酷な計算式

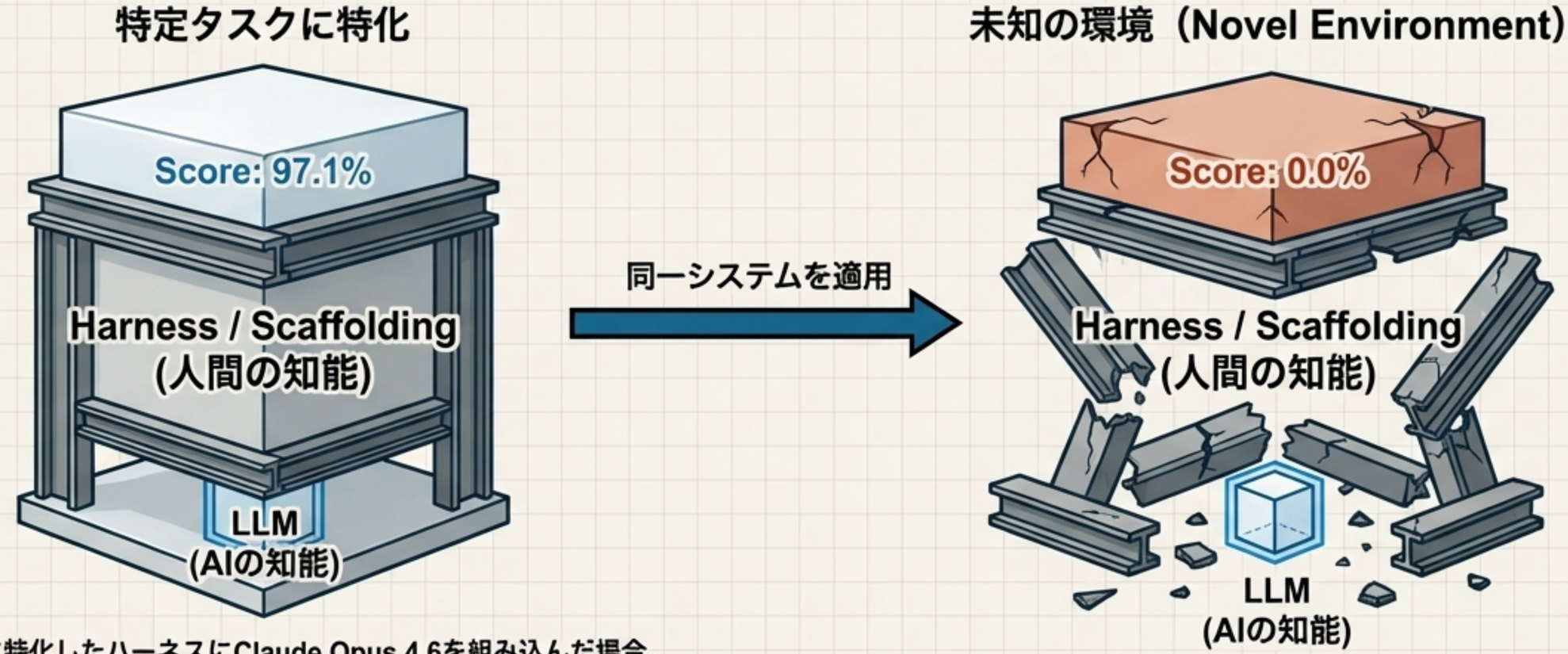
相対的人間行動効率（RHAE)。結果ではなく、人間のベースライン（初見プレイで2番目に優秀秀な行動数）と比較した「手数の少なさ」を二乗ペナルティで評価する。



計算資源の暴力で数万回の無作為なクリックを行い偶然正解にたどり着くことは、もはや「知能」の証明にはならない。人間の5倍の手数をかけた時点でスコアは強制的に0となる。

「ハーネスの錯覚」と真の汎化の分離

多くのベンチマークでの高得点は、LLM自身の汎用知能ではなく、人間が手作業で構築したワークフロー（ハーネス／足場）の優秀さを測っているに過ぎない。



ls20環境に特化したハーネスにClaude Opus 4.6を組み込んだ場合。

真のAGIは、人間による事前の手助け（ハードコーディングされた探索アルゴリズム）なしに、未知の環境に適応できなければならない。

完全な崩壊の理由：「次トークン予測」パラダイムの限界

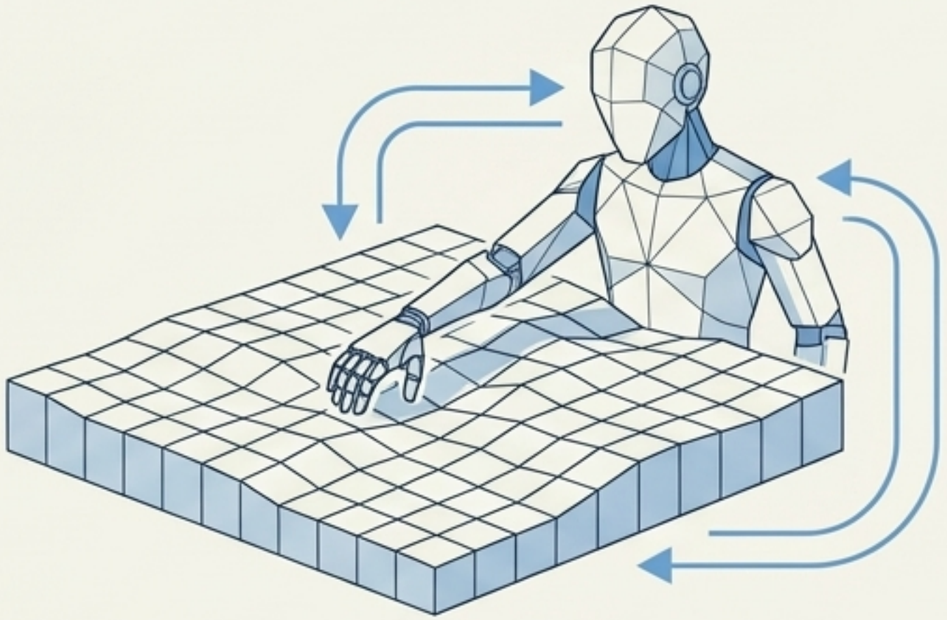
The LLM Paradigm

巨大な記憶の海

GPT-5.4やGemini 3.1 Proの根幹は自己回帰モデルである。彼らはインターネット上の高次元のパターンを補間することはできても、動的な物理空間におけるグラウンディング（Grounding）を持たない。

テキストベースの指示がなく、行動結果に基づいて自身の内部状態（信念）をリアルタイムに更新し続けなければならない環境では、数手で文脈を失う。

The Required Grounding



動的な物理空間

「巨大な記憶に依存した『確率的なオウム（Stochastic parrots)』は、世界に対する真の世界モデル（World model）を持っていない。」

— Yann LeCunらが指摘する根本的欠陥

プレビューコンペティションが示す、代替アーキテクチャの可能性

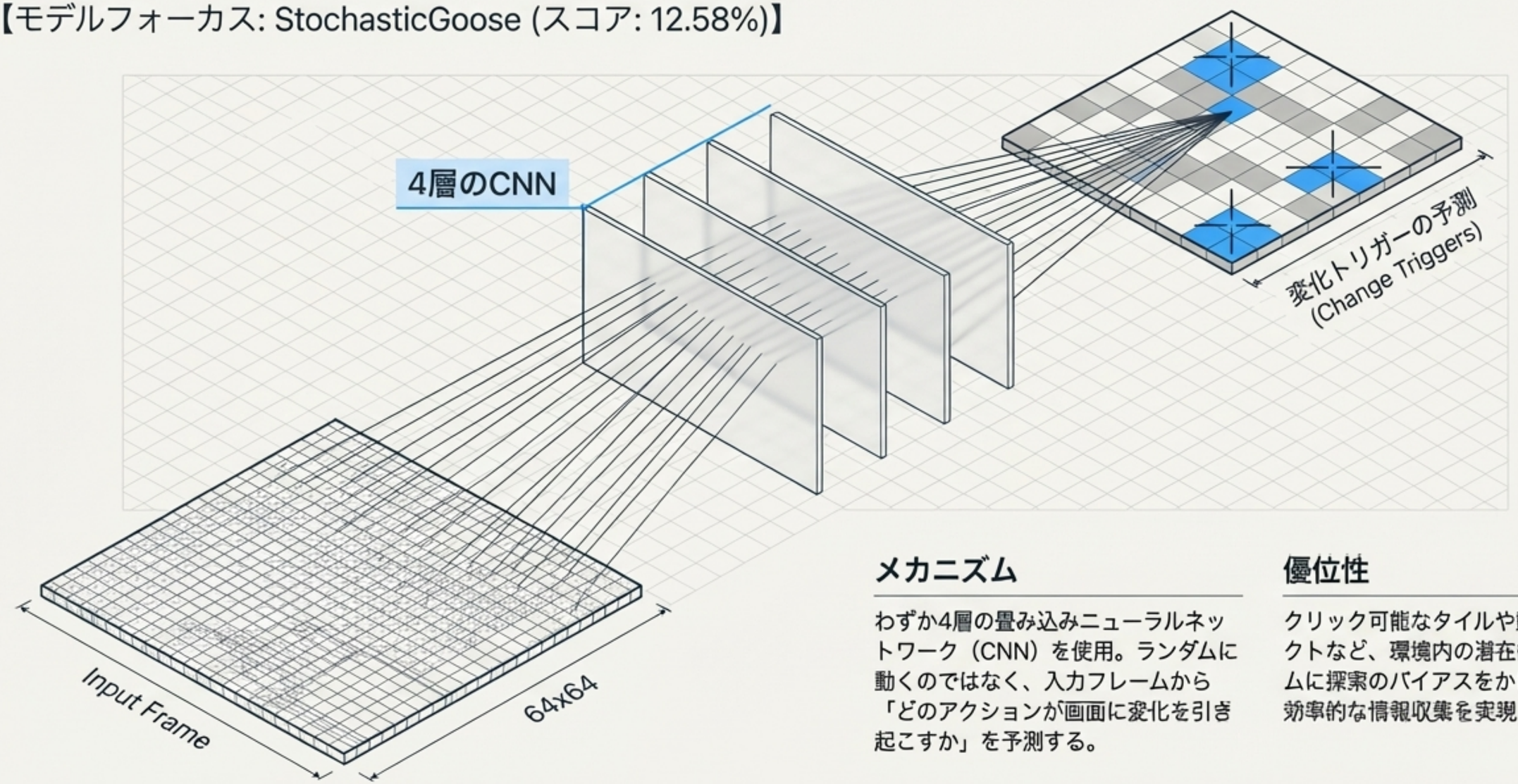
巨大な言語モデルが挫折する中、LLMに依存しない小規模で特化型のアーキテクチャが上位を独占した。

1位	StochasticGoose (Tufa Labs)	CNN + 変化予測に 基づく強化学習	RHAEスコア: 12.58% (18/20レベルクリア)
2位	Blind Squirrel (個人: Will Dick)	グラフベースの有向 状態空間マッピング	RHAEスコア: 6.71% (13/20レベルクリア)
-	フロンティアLLM群 (GPT-5.4, Gemini, Claude)	巨大な計算資源 / 自己回帰モデル	RHAEスコア: 1%未満

Grounded Editorial Whitepaper

アプローチ解剖 ①：視覚的優先度と能動的サンプリング

【モデルフォーカス: StochasticGoose (スコア: 12.58%)】



メカニズム

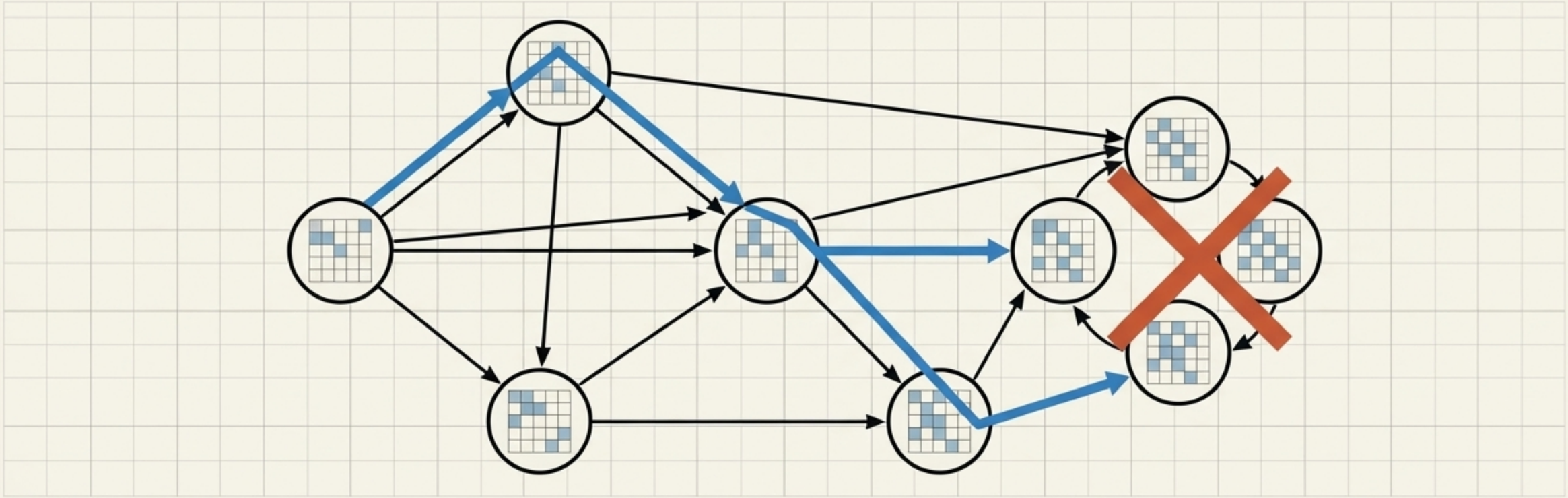
わずか4層の畳み込みニューラルネットワーク（CNN）を使用。ランダムに動くのではなく、入力フレームから「どのアクションが画面に変化を引き起こすか」を予測する。

優位性

クリック可能なタイルや動くオブジェクトなど、環境内の潜在的なメカニズムに探索のバイアスをかけることで、効率的な情報収集を実現した。

アプローチ解剖 ②：システマティックなグラフベース探索

【モデルフォーカス: Blind Squirrel (スコア: 6.71%)】



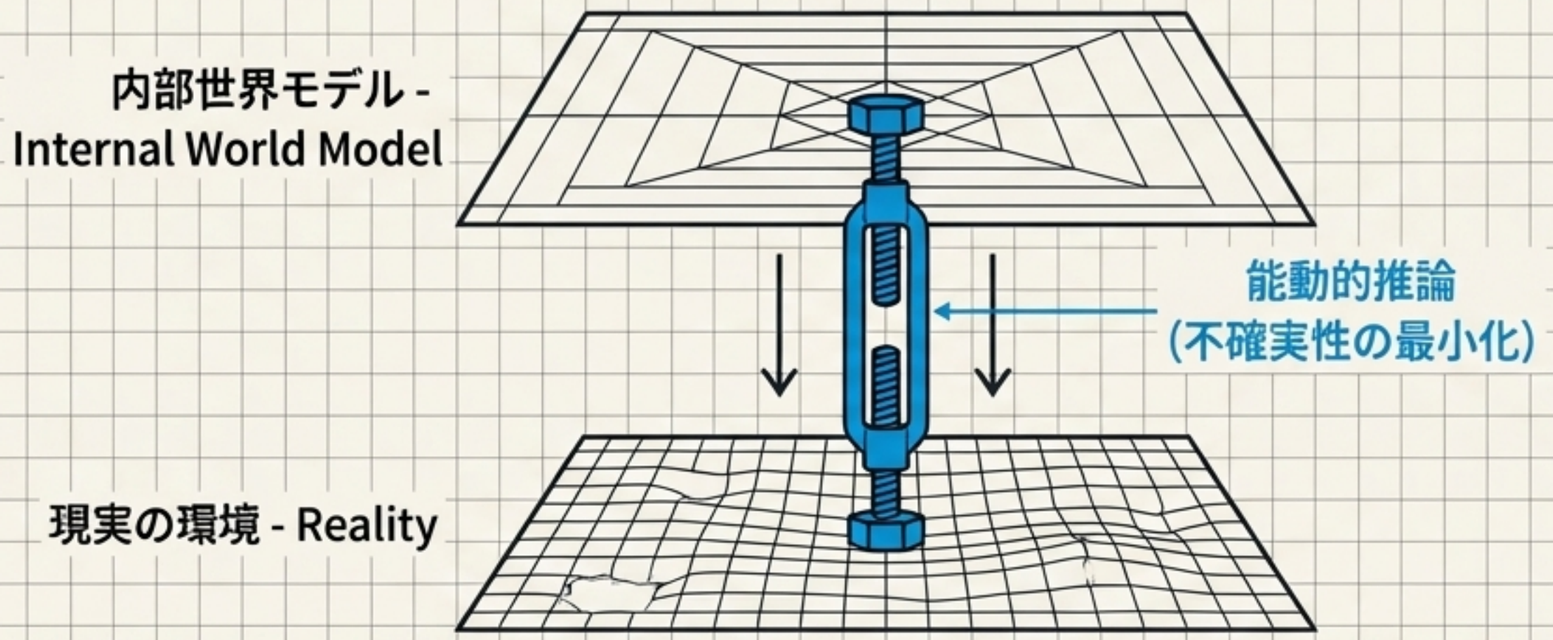
メカニズム

観察した各フレーム（状態）をハッシュ化して記憶。実行したアクションと結果をノードとエッジとして有向グラフ上にマッピングしていく。

LLMとの対比

LLMが数ターンの操作で幻覚（ハルシネーション）の霧に迷い込むのに対し、この手法は同じ状態をループする無駄な行動を完全に回避し、未検証のアクションへ至る最短経路を自律的に計算する。

次なる壁：能動的推論（Active Inference）の実装



現状の課題

現在のトップエージェントも完璧ではない。首位のStochasticGooseでさえ、水量調整レベル（vc33）の序盤で約350回もの無駄なクリックを消費している。

解決策

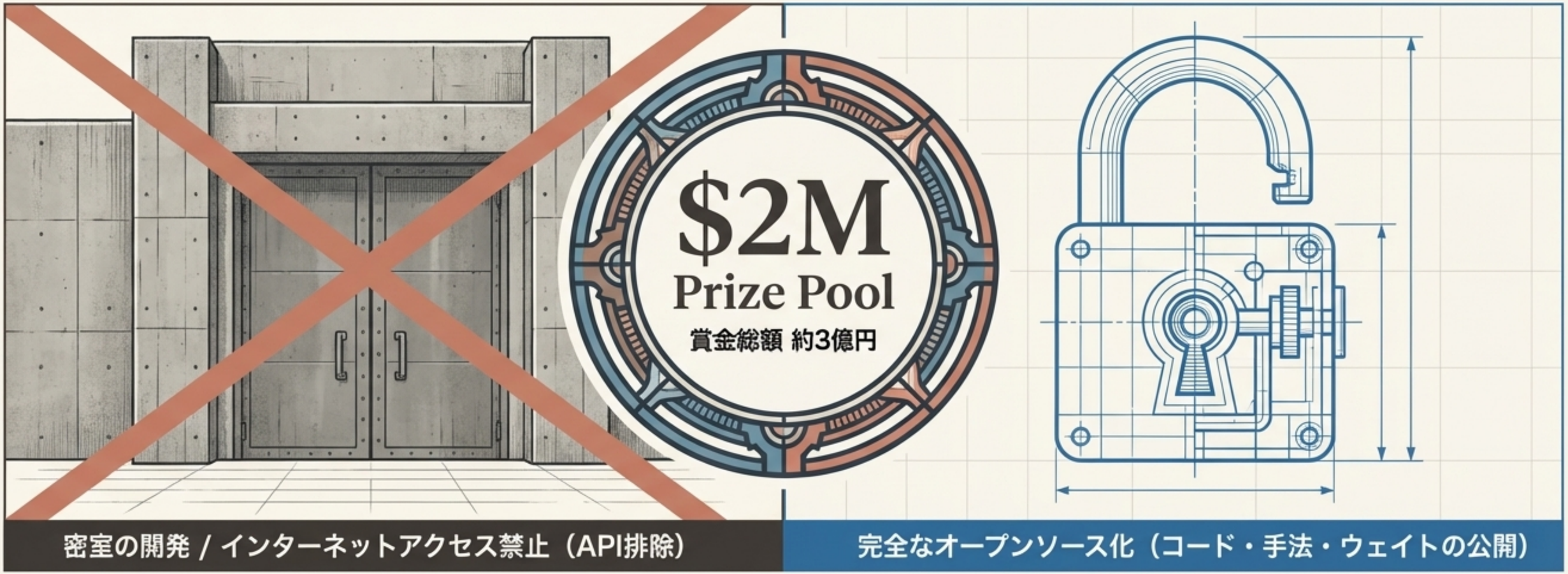
Karl Fristonらが提唱する「能動的推論」。エージェントが自らの内部世界モデルと現実とののズレ（不確実性）を最小化するように、内発的に行動を選択するメカニズム。

インサイト

このメカニズムの実装こそが、人間に匹敵するアクション効率（RHAE 100%）を達成するための鍵となる。

ARC Prize 2026：オープンソースAGI研究への回帰

AI開発の主導権を少数の巨大テック企業による「密室の開発」から取り戻し、コミュニティによる透明性の高い科学的進歩へと回帰させる。



【ルール1：APIへの依存断ち】 評価中のインターネットアクセス禁止。GPT等の巨大商用APIを排除し、自己完結型モデルの開発を必須とする。

【ルール2：知識の共有】 賞金獲得の条件として、MIT-0やCC0などの寛容なライセンスで完全公開する義務。早期公開者へのマイルストーン賞も設定。

真のAGIへ向けた北極星

2026年3月、ARC-AGI-3は「計算資源の暴力によるパターンの暗記」から「**未知の環境における適応的なスキル獲得**」へと、評価のパラダイムを決定的に移行させた。

既存のトランスフォーマーを数兆パラメータへ拡大し続けるだけでは、AGIには到達しない。

神経記号的AI（Neurosymbolic AI）、能動的推論、そして**新しい探索アルゴリズム**。
現在「唯一の未飽和なエージェントAIベンチマーク」のリーダーボードで真のスコア上昇が観測されたとき、我々は歴史的な証明を目撃する。