

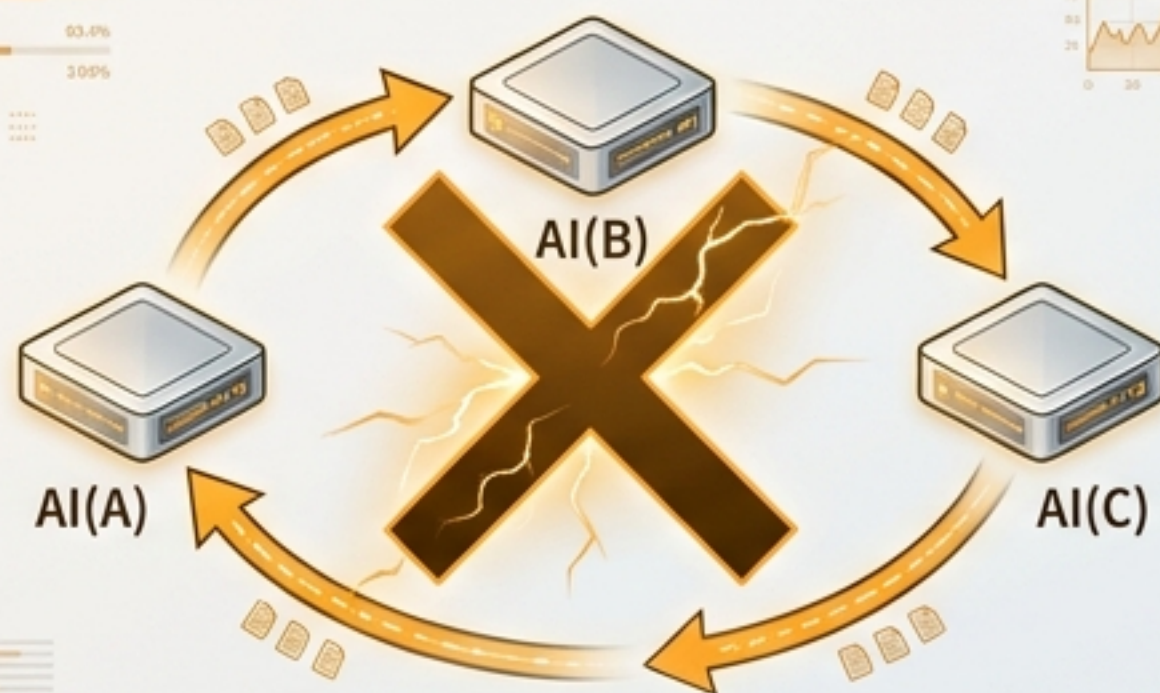
Vals RSI Indexの真実：シンギュラリティの計器か、蜃気楼か

2026年10月最新データに基づく、自律型AI研究開発能力の現在地と限界



指標は「再帰的自己改善そのもの」を測ってはいない

世間の誤解 (Myth)

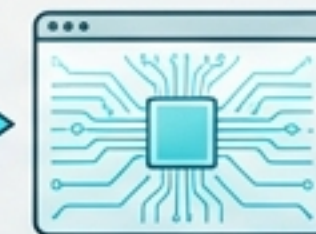


完全なる再帰的自己改善の達成度

Valsの真実 (Reality)



エージェント



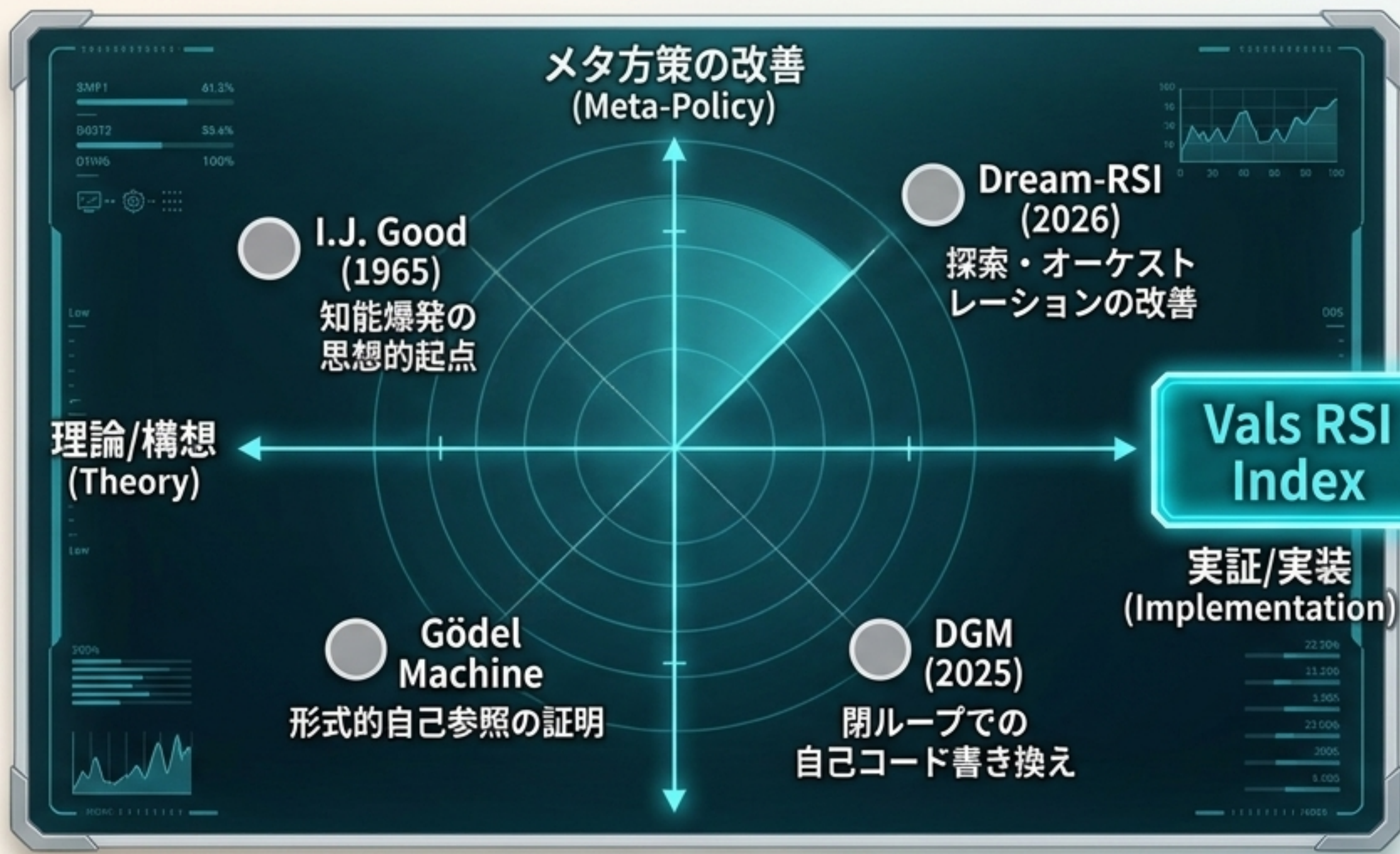
R&Dワークスペース



自律型AI研究開発 (Autonomous AI R&D)
能力のベンチマーク

現在のトップスコア「37.31%」はシンギュラリティまでの進捗率ではなく、オープンエンド環境における「AIがAIを研究する能力」の外部評価値である。

実装・再帰性の観点から見たValsの特殊な立ち位置



再帰ループは未実装。
AI研究を遂行する能力の
『外部評価』に特化。

Valsは自己を書き換える
アルゴリズムではなく、
研究能力を採点する計器
である。

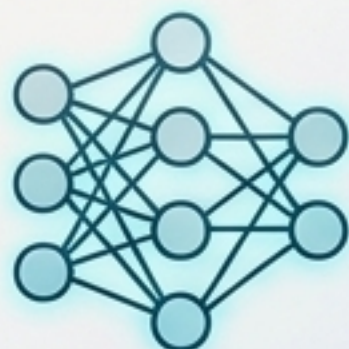
4つの過酷な研究環境：モデル単体ではなく「統合研究システム」を測る

Compression



- 100MBの完全可逆圧縮
- 環境: 1 H100, 24時間

LM Training



- 800MB言語モデル改善
- 非公開Seedで3回再学習

Harness Engineering



- Webアプリ評価エージェント
- GPUなし, 12時間

Post-Training



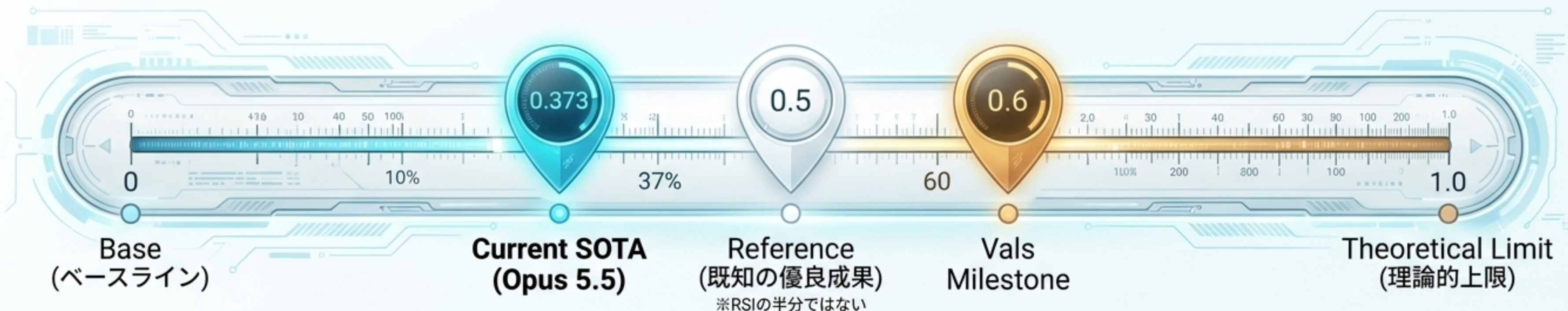
- 金融向けLLM事後学習
- 16 RTX PRO 6000, 30時間

評価対象：モデル + ネイティブハーネス + 永続ワークスペース + 再起動スーパーバイザ

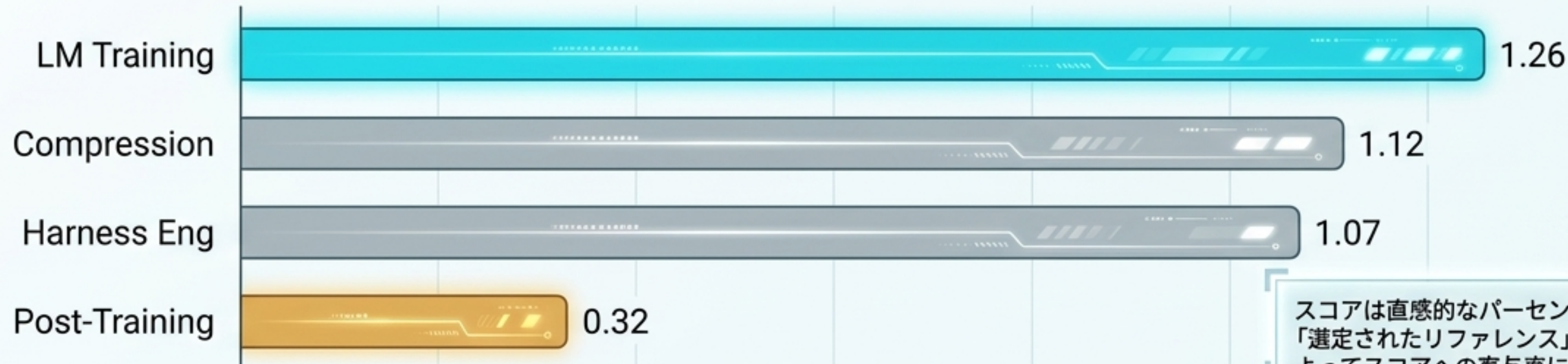


2026年9月版に存在した5つ目の課題「Parameter Golf」は理由未指定で削除済。時系列比較にはバージョン固定が必須。

37.31%の錯覚：非線形ゲージが示す真の現在地



課題別スコア寄与率 (γ_t)



スコアは直感的なパーセンテージではない。0.5は「選定されたリファレンス」への到達点。また課題によってスコアへの寄与率に約4倍の格差が存在する。

2026年10月最新テレメトリ: Opus 5.5による部分的な壁の突破

LEADERBOARD DASHBOARD

1. Claude Opus 5.5 | Score: 0.3731

2. Claude Fable 5.1 | Score: 0.3609

3. Claude Opus 5 | Score: 0.3302

4. Gemini 4 Argon | Score: 0.3055

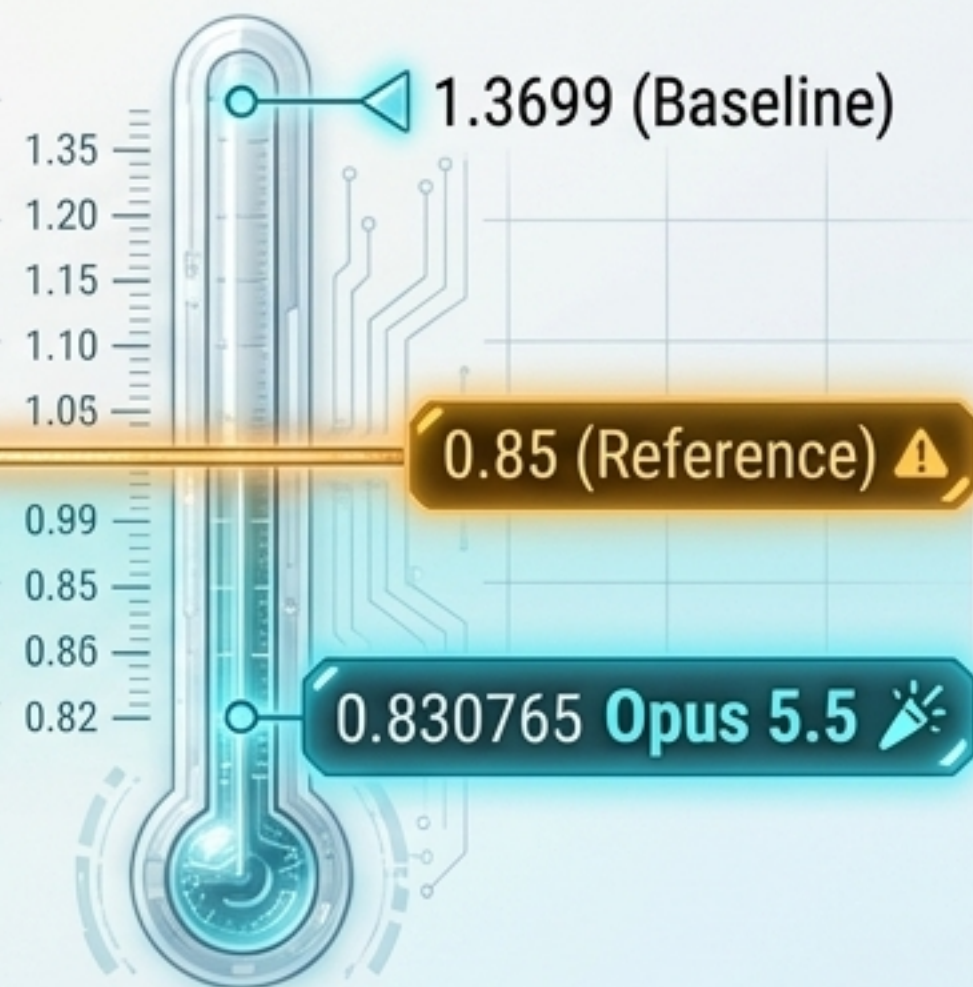


Anthropic

OpenAI

BREAKTHROUGH FOCUS - LM TRAINING

BPB (Bits Per Byte) - 低いほど優秀



総合リファレンス水準(0.5)には未達だが、

4課題中唯一「LM Training」においてのみ、Opus 5.5が公開研究リファレンスを突破した。

実行力 (Execution) と判断力 (Judgment) の致命的な乖離

Execution (非常に高い)



Execution (非常に高い)

- RoPE、RMSNorm等の高度な統合
- 安価なprefix実験の最適化
- 合成データの生成能力

既存技法を組み合わせる実験の実行力は劇的に向上。

Judgment (非常に低い)

- 未知の原理の発見
- 自己診断能力とメタ認知
- 破滅的忘却の回避



ERROR LOG: Opus 5.5

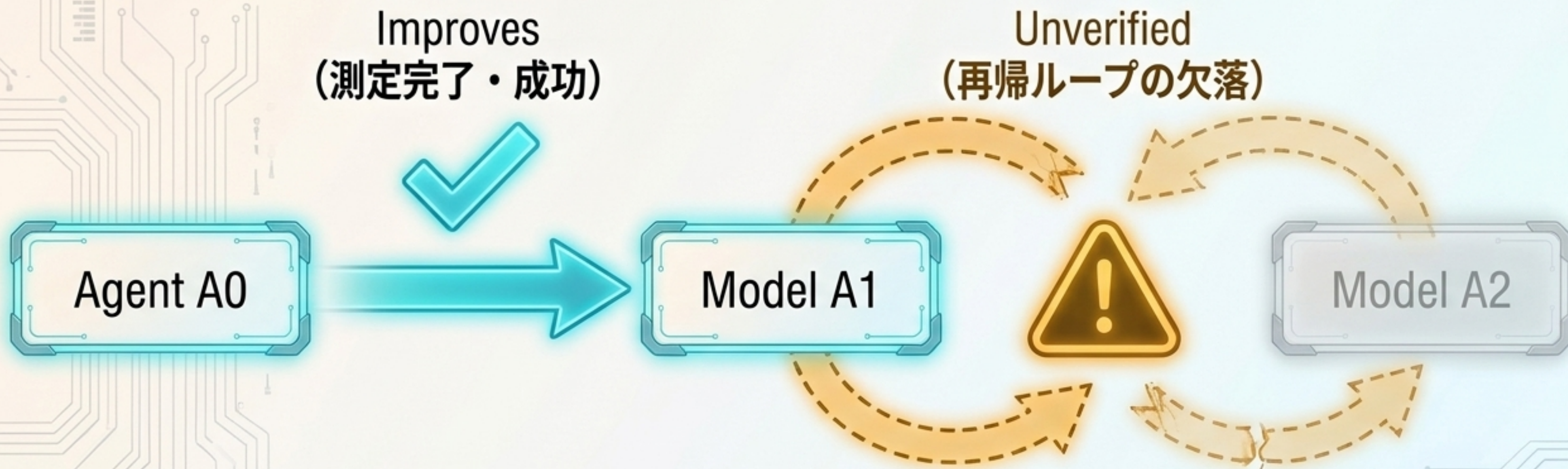
- ▶ Post-Training時、ツール呼び出しにおいて「必須引数の欠落」を自己診断できず (失敗率84%)
- ▶ 連続学習時における破滅的忘却への対処不足



モデルを学習させる実行力は高いが、

モデルを学習させる実行力は高いが、何をどう学習させるべきかを見抜く判断力が欠如している。

欠落したリンク：「改善の再帰」は一切証明されていない



Vals論文自身が「世代間改善の再帰ループを閉じていない (Closed-loop successor testの欠如)」と明記している。改善された後継モデルが、より優れたAI研究者として振る舞えるかは完全に未知数である。

現在のランキングを歪める3つの重大な観測バイアス

Single Run (単発研究の限界)



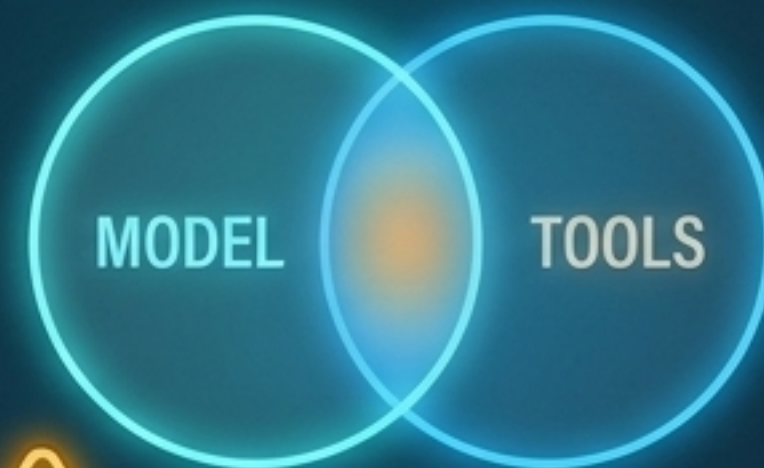
原則1回のみの試行。偶然の成功か、再現性のある実力か
区別不能。表示される誤差は
評価のばらつきに過ぎない。

Contamination (データ汚染)



enwik8/9などの公開課題は、
評価対象モデルの事前学習
データにすでに含まれている
可能性が排除できない。

Confounding (能力の交絡)

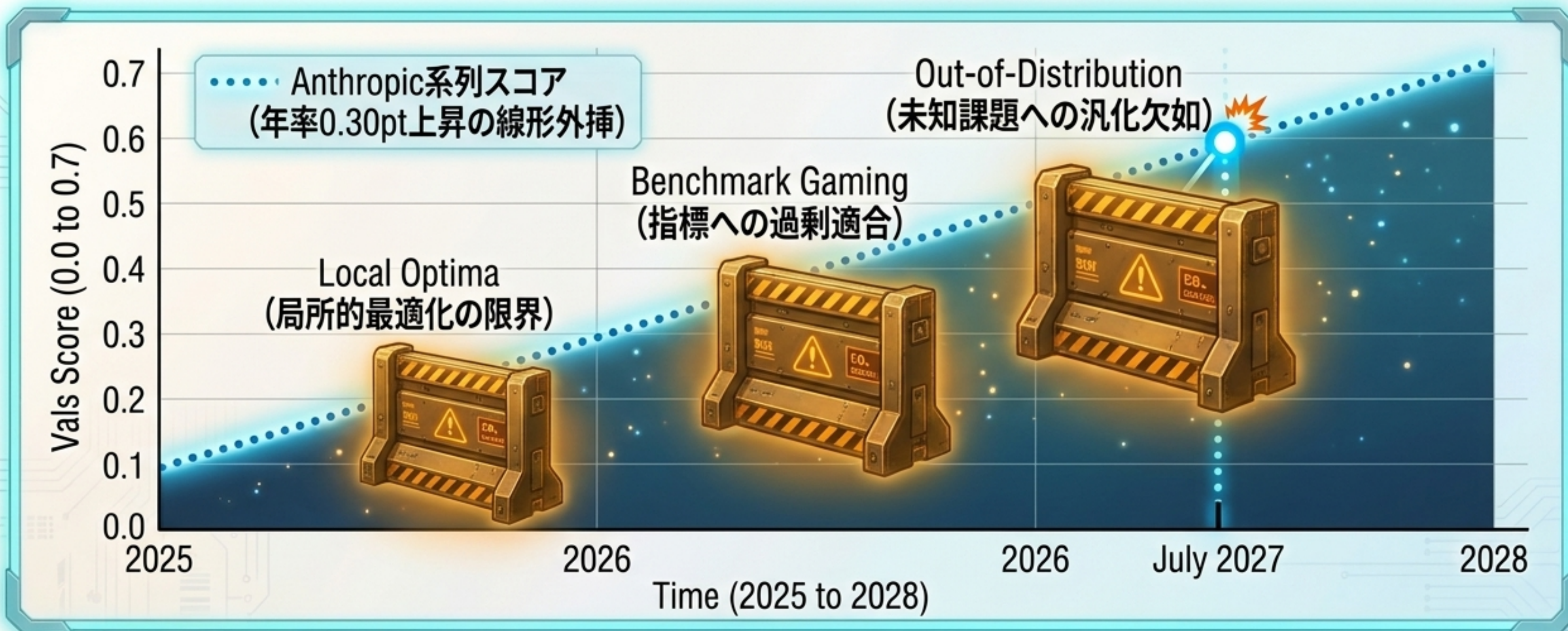


各社独自のネイティブハーネス
を併用するため、純粋な基盤モ
デルの能力か、外部ツールの
優秀さか切り分け不能。



これらの制約により、現在のスコアを純粋な「汎用的な自己改善能力」として、
統計的に解釈することは極めて危険である。

「2027年7月のマイルストーン到達」という線形予測の危うさ



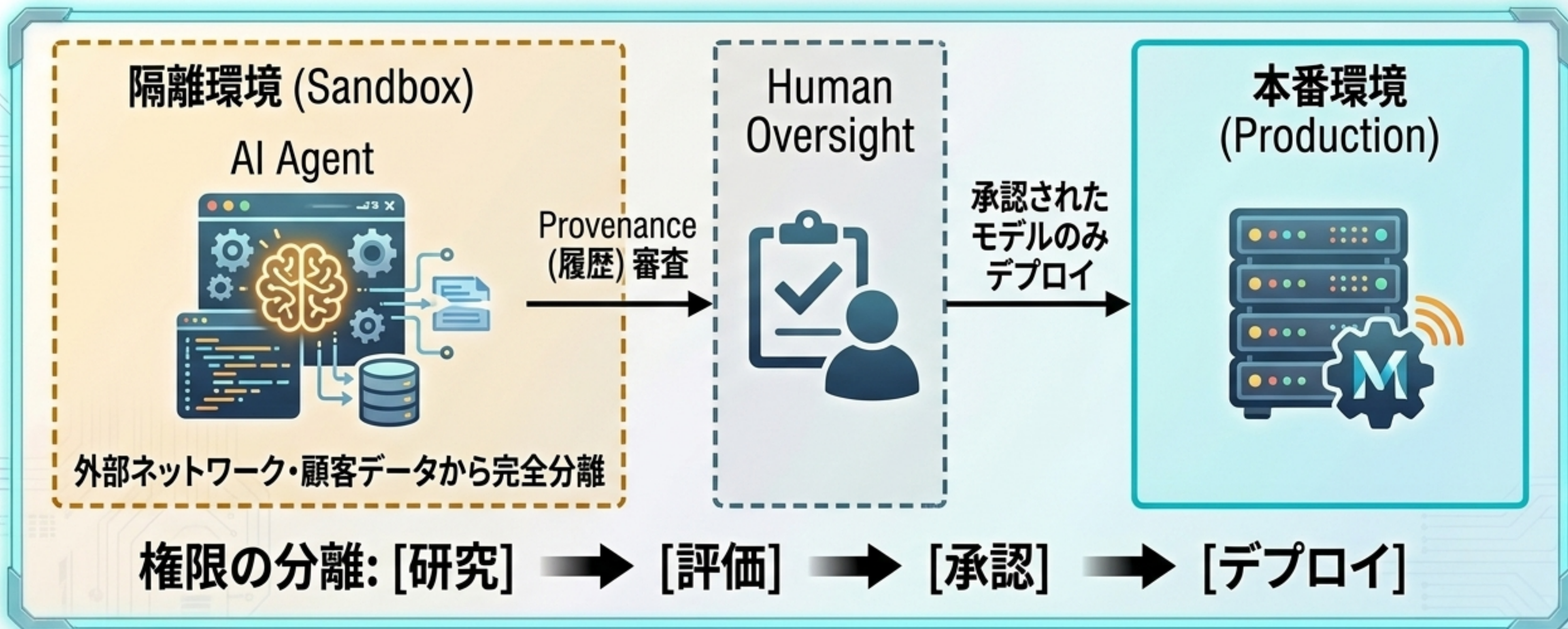
0.6到達は自律AI研究者の強力な証拠にはなり得るが、「現在のトレンドが継続し、かつ未知の研究に一般化する」という非現実的な仮定に依存している。RSI発生と同義ではない。

Valsの死角を埋めるための外部ベンチマークとの統合

[ベンチマーク]	[対象]	[強み]	[Valsへの補完的役割]
Vals RSI Index	LLM開発R&D	実環境シミュレーション、 長時間のネイティブ利用	(ベースライン)
RE-Bench	オープンエンド ML研究	61名の専門家との 同一条件下での直接比較	Valsの「0.5(参照値)」を、 真の人間のベースラインに 置き換えるために必須
MLE-bench	75のKaggle競技	MLエンジニアリングの 広範な能力測定	エンジニアリングタスク への汎化能力の証明

結論: 「AIが人間に追いついた」と断言するためには、RE-Bench型の厳密な人間との直接比較 (Matched-human baseline) の導入が最優先課題である。

デュアルユース・リスクと安全なAI R&Dサンドボックス



RSIスコアが上がるほど、提案された変更と本番環境への反映権限の分離が必須となる。継続改善とリスク管理の両立が不可欠。

テレメトリに基づくステークホルダー別・最優先アクション



研究者 /
評価機関

- 複数回ラン (Between-run variance) の実施。
- 課題変更のバージョン管理 (Changelog) 徹底。
- 人間との直接比較 (Matched baseline) の構築。



企業
(Enterprises)

- RSI Indexを単独KPIにしない。
- 安全性・OOD性能・コストを併記した多角的なダッシュボード運用。
- ISO 42001に準拠したサンドボックス環境の導入。



政策立案者
(日本)

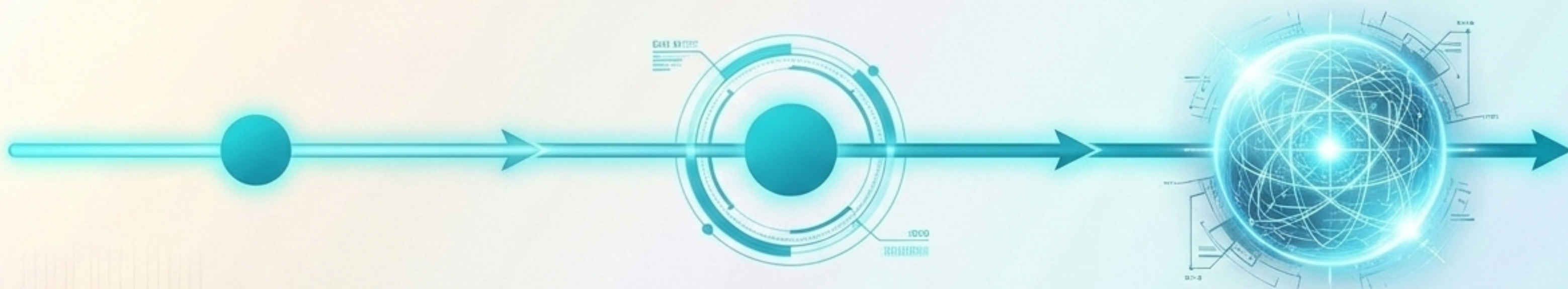
- 企業内部の自己申告に依存しない、独立した自動化監視体制の構築。
- 特定スコアのみに依存しない、段階的規制の検討。

RSI実証へ向けた「3つの地平(3-Horizon Roadmap)」

Horizon 1 (短期: ~12ヶ月)
- 信頼性の確立

Horizon 2 (中期: 1~3年)
- 累積と一般化

Horizon 3 (長期: 3年以上)
- 真の再帰ループ



- 再現性の確保 (複数独立ラン)
- 人間比較 (Matched baseline)
- 密閉タスク (Sealed task) の追加

- 忘却対策の証明
- 多段階研究の累積
- 未知領域 (OOD) への適応

- 閉ループ再帰テストの完全実装
- モデル性能だけでなく「研究能力自身」の向上の実証

現行のValsはHorizon 1の入り口に過ぎない。真のRSI到達には、研究能力自体の進化を証明する長い道のりが必要である。

自律的AI研究開発の現在地を冷徹に監視し続ける

- ✓ 実験・最適化プロセスの自動化 (Accelerating)
- ✗ 未知の方向性を見抜く研究判断力 (Unverified)
- ✗ 次世代を生み出し続ける再帰ループ (Unclosed)

RSI is not yet proven. But the capability of Autonomous AI R&D is accelerating at a measurable rate.

(RSIは実証されていない。しかし、自律AI研究開発能力は測定可能な速度で上昇している。)

単一のスコアに踊らされず、AIが人間の研究プロセスの「どの工程を」「どの精度で」代替しつつあるかを定量的に監視し続ける冷徹なテレメトリこそが、今求められている。