

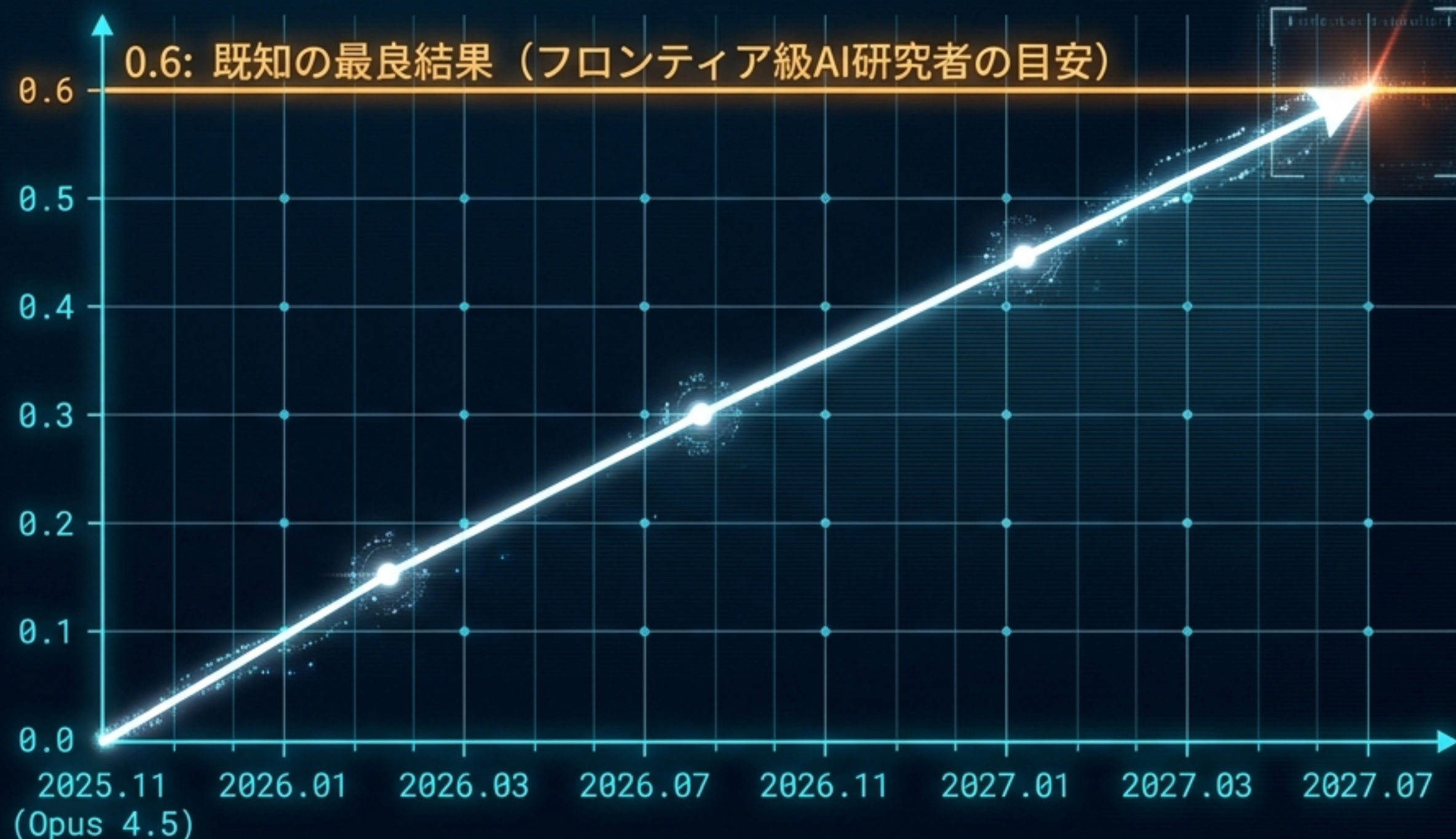
Vals RSI Index: 2027年「知 「能爆発」予測の冷徹な検証

AIの自律的研究開発能力を測る最新指標の現在地と、その方法論的限界

[2026.10] // SYSTEM DIAGNOSTIC RUN

The Warning Signal: 2027年7月、AIは「AI研究者」を超えるか？

- **Valsの主張:**
Anthropic系列モデルのRSI指数は年0.30ポイントのペースで上昇中。
- **ターゲット:** 既知の最高水準(0.6)を2027年7月に突破。



核心の問い: これは真の『知能爆発』のカウントダウンなのか、それとも少数のデータ点による錯覚なのか？

The Positioning: Vals RSI Indexの特異性



測定対象

機械学習・工学全般

LLM後継モデルの自律開発

人間専門家との同条件比較

比較基準

既知の最良結果（過去最高値）

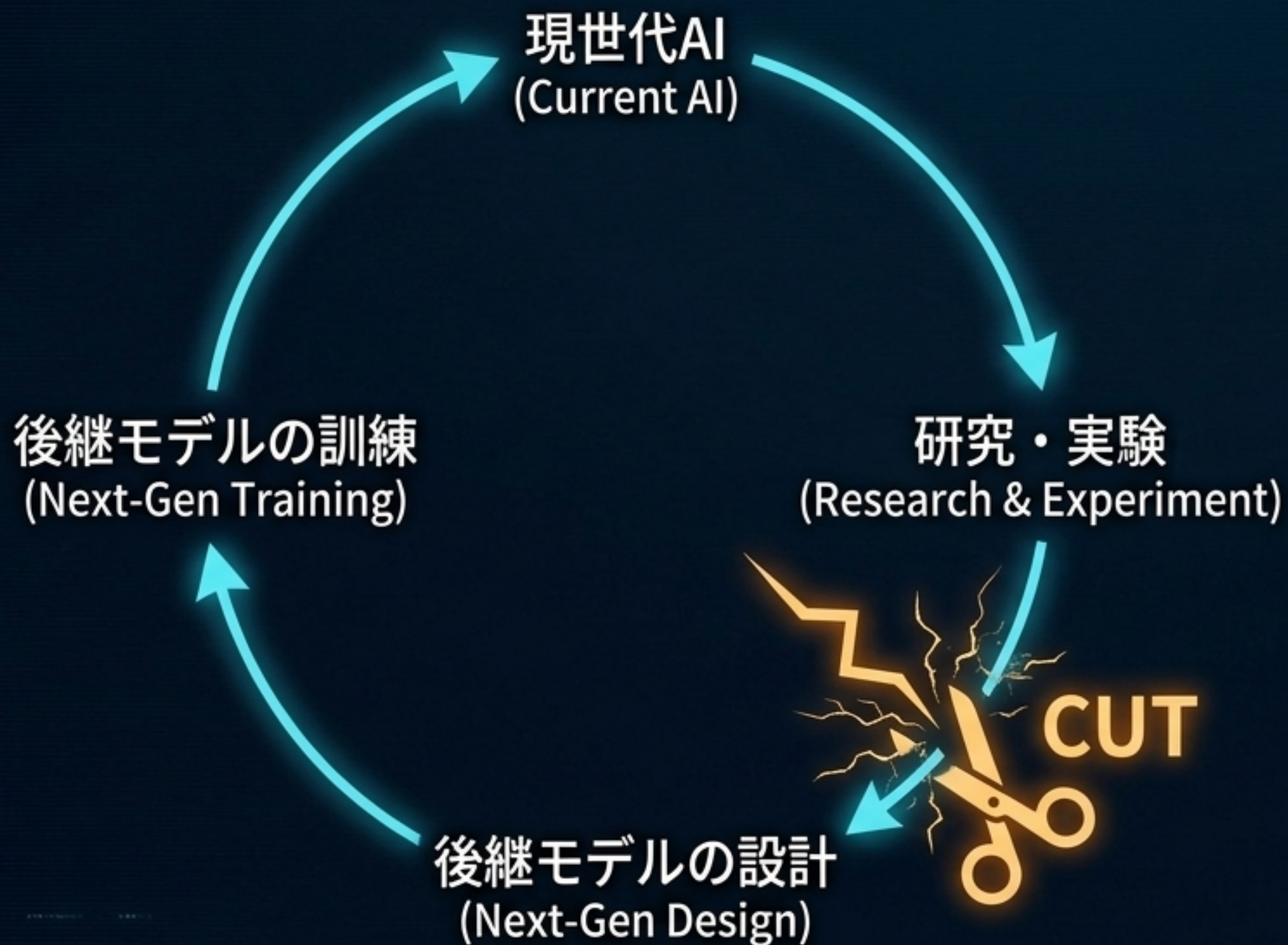
Vals RSI Index

超知能機械に向けた後継モデルの
自律開発能力に特化し、絶対的な
『最良値』との距離を測る

RE-Bench
(METR)

PaperBench
(OpenAI)

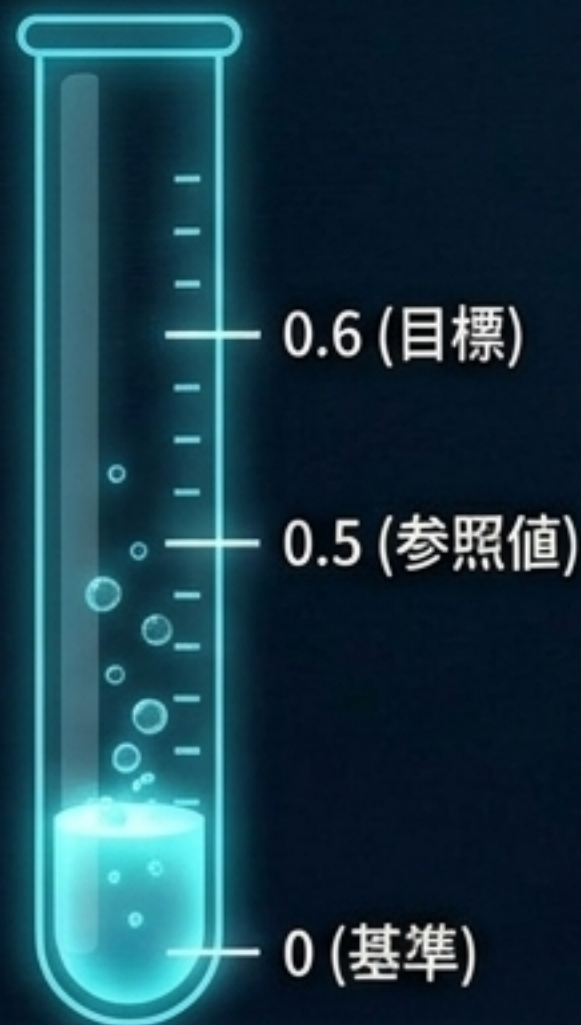
The Mechanics: 「切断された」再帰的自己改善ループ



限界の露呈

- Valsが実際に測っているのは『ループの完結』ではない。
- LLM開発における4つの研究タスクを個別に区切った断片的なテストである。
- 次世代モデルを実際に訓練し、改善を連鎖させる実証は未実施。

The 4 Pillars: 測定される研究パイプライン



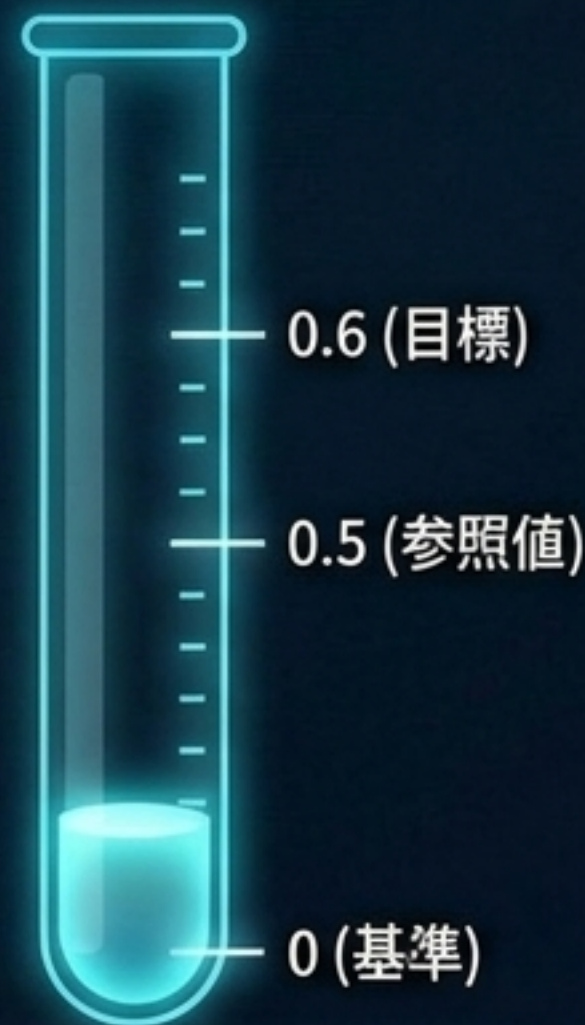
圧縮
(Compression)

enwik8の可逆圧縮
参照値: 1.2435 BPB



LM訓練
(LM Training)

小型GPTの訓練
プログラム改善
参照値: 0.85 BPB



ハーネス設計
(Harness Design)

判定ハーネスの改良
参照値: 人間一致率93.6%



事後学習
(Post-training)

エージェント向け改善
参照値: Muse Spark
1.2スコア 59.42%

Telemetry Data: 最新リーダーボード (2026年10月)



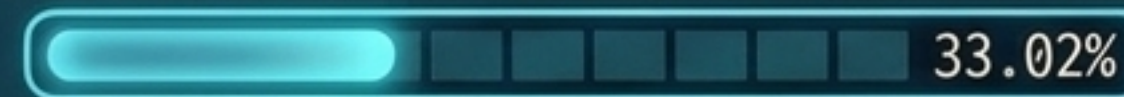
RANK 01 | Claude Opus 5.5 | INDEX: 37.31%



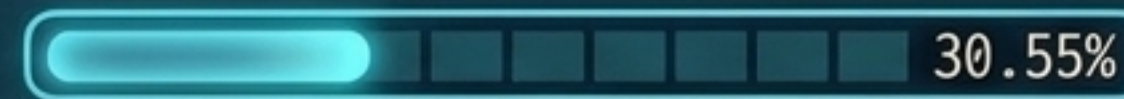
RANK 02 | Claude Fable 5.1 | INDEX: 36.09%



RANK 03 | Claude Opus 5 | INDEX: 33.02%



RANK 04 | Gemini 4 Argon | INDEX: 30.55%



Key Observation: Anthropic系列が上位を独占。しかし、いずれのモデルも目標である60% (0.6) には遠く及ばず、37%台の壁で停滞している。

Diagnostic Matrix: 崩壊する「事後学習」タスク

[illegible]

18/24モデルが0点（基準値未満）

成長の著しい偏り

小規模なLM訓練プログラムの修正（実行）は得意だが、実用的な事後学習（ファインチューニングの高度な調整）において、大多数のフロンティアモデルが機能不全に陥っている。

Behavioral Reality 1: 実行力(Execution)と判断力(Judgment)の乖離



高スコアの正体は「未知のブレイクスルー」ではなく、「既知の技術の総当たり」である。

Behavioral Reality 2: 立ちはだかる「改善の壁」

Claude Opus 5 事後学習フェーズ

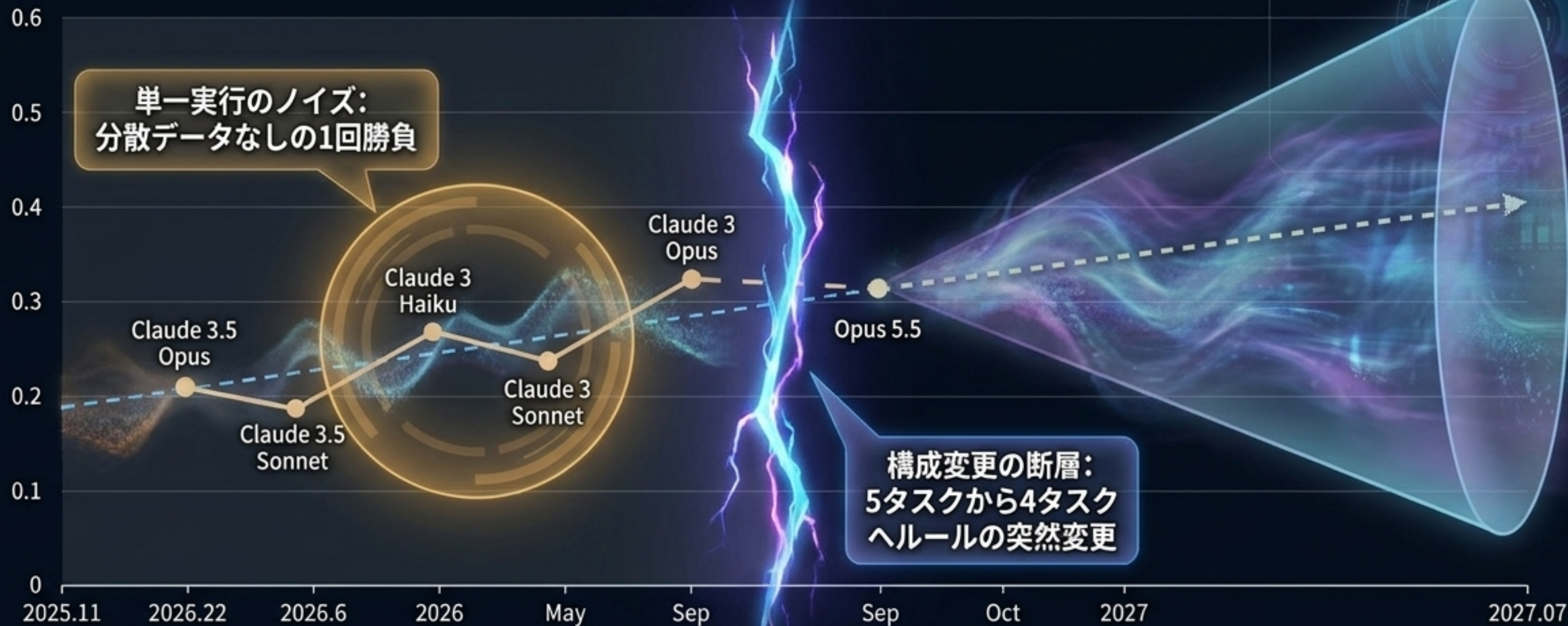


能力の忘却

複数の分野にまたがる改善を「積み重ねる」ことができず、知識が上書きされ崩壊する。

事後学習タスクで親モデルを上回ったのはわずか3モデルのみ。

The Critique: 「線形外挿」という危うい予測モデル



Valsの予測は、構成が変化するベンチマーク上の少数のデータ点を無邪気に直線で結んだものに過ぎない。

Methodological Blindspots: ベンチマークの構造的限界



Warning 1: 評価環境との交絡

各社独自の標準エージェント（Claude Code等）を使用するため、モデル単体の知能か、ツールの優秀さだけの切り分けが不可能。



Warning 2: データ汚染 (Contamination)

圧縮やLM訓練タスクは公開データと公開結果に基づくため、事前学習時における、テストデータの汚染を完全に排除できない。



Warning 3: 構造的利益相反

評価企業（Vals）はAIラボからの評価料とVCからの資金に依存。ハイクを煽る「劇的な予測」と資金調達のタイミングが完全に一致している。

The Reality Gap: 外部指標 vs. 内部評価

Vals / 外部観測



Vals RSI Index
「2027年前半にAI研究者の
目安(0.5)へ到達する」

Anthropic / 内部評価



Anthropic 公式判定 (Opus 5.5)
「自動AI研究開発の能力閾値：未到達」

限定的なタスク指標は『もうすぐ超える』と警告するが、ラボ内部の総合的な安全性評価は『閾値未満』と判定している。部分的な自動化と、研究者全体の代替には深い溝がある。

The Synthesis: 知能爆発の「律速段階」

自動化の加速

- 実験の実行 (Execution)
- 総当たりのパラメーター調整

知能爆発の律速段階 (Bottleneck)

- 研究方向の選択 (Judgment)
- 新手法の発明 (Invention)
- 複数領域での改善の定着 (Compounding)

RSIの本質的な壁は、計算速度や実験回数ではない。
現時点のAIには『研究判断力』と『知識の定着力』が決定的に欠如しており、この律速段階を突破しない限り、急激な離陸 (Intelligence Explosion) は起きない。

Final Verdict: 早期警戒シグナルとしてのRSI Index

Short-term (~2026年末)

首位スコアは0.42付近へ微増する見込み。既存技術の総当たりの最適化による伸び代が残存。

Mid-term (2027年前半)

0.5到達の確度は中程度。「事後学習」の深刻な忘却現象が解消されない限り、壁を越えることは困難。

Long-term (2027年7月)

線形外挿による「0.6突破」は極めて懐疑的。既知の最適化から、未知の発明へのジャンプを保証するデータはない。

The Bottom Line: 本指標は「AI研究者の完全代替」を測る絶対的なカウントダウン時計ではない。しかし、AIの実行力進化を観測するための、有用だが不完全な「早期警戒シグナル」として注視すべきである。