

Vals RSI Index (Vals 再帰的自己改善インデックス) 調査報告書

エグゼクティブサマリー

結論から言えば、Vals RSI Index は「再帰的自己改善そのもの」を測る指数ではなく、再帰的自己改善に必要なと考えられる「AIがAIを研究・改良する能力」を、外部から共通条件で測定する長時間・オープンエンド型ベンチマークである。Vals AI は、モデルに言語モデル開発上の研究課題を12〜30時間自律的に行わせ、提出された成果物を、開始時点のベースライン、既知の優れた研究成果、理論的上限の3点で正規化する。2026年10月10日時点の現行版は **Compression、LM Training、Harness Engineering、Post-Training の4課題**からなり、総合値は4課題スコアの単純平均である。公式ページは「Proprietary」と表示されている一方、方法論、数式、主要な実験条件は公開されている。 ¹

現行リーダーは **Claude Opus 5.5、37.31% (0.3731)** で、Claude Fable 5.1 が36.09%、Claude Opus 5 が33.02%、Gemini 4 Argon が30.55%と続く。Opus 5.5 はLM Trainingで0.8308 BPBを達成し、選定された0.85 BPBの公開研究リファレンスを初めて上回った。一方、総合値のリファレンス水準は0.5であり、現在のトップもそこからはかなり離れている。さらに4課題中、既知のリファレンスを超えたのはLM Trainingだけである。 ²

尺度は直感的な「RSI達成率」ではない。各課題で **0=ベースライン、0.5=選定された既知の優良成果、1=設定された理論的上限**となるよう非線形変換している。したがって「37.31%」を「完全な再帰的自己改善の37.31%を達成した」と解釈するのは誤りである。0.6も物理的・数学的に証明されたRSI閾値ではなく、Vals がAlexNet、BERT、LSTM→Transformer等の歴史的改善幅と比較して設定した実務的マイルストーンである。 ³

特に重要なのは、論文自身が **RSIの再帰ループを閉じていない**と明記していることである。つまり「エージェントAがより優れたモデルBを作り、そのBがさらにCを作る」という世代間改善を検証していない。現時点のVals RSI Indexは、厳密には **Autonomous AI R&D Capability Index (自律AI研究開発能力指数)** に近い。 ⁴

再現性にも大きな留保がある。CompressionとLM Trainingは公開データと比較的完全な仕様を用いる一方、Harness EngineeringとPost-Trainingの隠し評価データは汚染防止のため非公開である。また、**各モデル×課題の研究キャンペーンは原則1回だけ**であり、表示される標準誤差は提出後の固定成果物を複数回評価したばらつきであって、「もう一度研究させた場合」の分散ではない。これは現在のランキングを統計的に解釈する際の最重要制約の一つである。 ⁴

さらに、2026年9月時点の公開版は **Parameter Golfを含む5課題**で、Opus 5.5の総合値は37.13%だったのに対し、10月の現行版は4課題・37.31%となっている。Valsの過去の公式ページにも「five-task benchmark」と記録され、Parameter Golfでの結果も公表されているが、現行ページから同課題が除外された理由は、確認した公式資料では**未指定**である。このため異なる日付のRSI Index値を時系列で比較する際は、バージョン固定が必須である。 ⁵

本報告の総合評価は次の通りである。

観点	2026年10月時点の評価
「AIがAI研究をする」能力の測定	有力な外部ベンチマークの一つ 。実成果物、長時間実験、固定計算資源を組み合わせる点が高い。 ⁶
完全なRSIの実証	未達・未検証 。後継モデルが次の後継を作る再帰ループを閉じていない。 ⁴
人間のAI研究者との比較	限定的 。リファレンス成果は使うが、同条件の人間研究者コホートとの直接比較ではない。 ⁷
再現性	部分的 。公開2課題は比較的再現可能、非公開2課題は第三者による完全複製不能。研究試行は基本1回。 ⁴
現時点の革新性	主として既存技法の統合・実験最適化 。根本的新手法の自律発明を示す証拠はまだ弱い。 ⁸
将来予測への利用	先行指標として有望だが、単独予測には不十分 。Vals自身の2027年予測も線形外挿と一般化仮定に依存する。 ⁷

定義・起源と理論的背景

Valsの定義では、RSI（Recursive Self-Improvement）は単なる「自分のコードを書き換える能力」ではない。モデル・アーキテクチャ、アルゴリズム、データ取得・選別、事前学習、事後学習、評価設計、システム工学など、次世代AIを構築する研究開発プロセス全体をAIが担い、その能力向上が次世代の研究能力向上につながる概念として扱われる。Valsは、最終形が単一モデルではなく、多数のエージェントとそれらを統括するオーケストレーション層からなる可能性も明示している。 ⁹

思想的起点としてValsが挙げるのはI. J. Goodの「ultraintelligent machine」である。Goodは、人間のあらゆる知的活動を超える機械なら「より良い機械を設計する」という知的活動も上回り得るため、そこから急速な知能向上が起ころうと論じた。Valsはこれを1965年の議論として引用している。 ¹⁰ 日本語圏でも2026年8月、人工知能学会・汎用人工知能研究会で神楽坂やちまによる「RSI（再帰的自己改善）：シンギュラリティをもたらすメカニズムとは」が公表され、RSIの定義、分類、メカニズム、実装状況が研究テーマとして整理されている。 ¹¹

理論面で重要な先行概念がJürgen Schmidhuberの **Gödel Machine** である。これは自己参照的なプログラムが、自身を書き換えることの有用性を形式的に証明できた場合に自己変更する構想で、Good的な自己改善を計算機科学的に形式化した代表例である。ただし現実的な複雑系で「変更が将来にわたって必ず有益」と証明することは極めて難しい。 ¹²

この証明ボトルネックを経験的評価に置き換えたのが2025年の **Darwin Gödel Machine (DGM)** である。DGMは自分自身のコーディングエージェントのコードを書き換え、変更後をベンチマークで実証評価し、優れたシステムをアーカイブしながら探索する。論文ではSWE-bench性能が20.0%から50.0%、Polyglotが14.2%から30.7%へ改善し、Sakana AIはコードと研究成果を公開している。これはValsのような「RSI能力の測定」ではなく、**限定された実際の自己変更ループの実装例**という点で補完的である。 ¹³

さらに2026年9月の **Dream-RSI** は、基盤モデル本体を変更するのではなく、探索の分岐、並列化、停止などを制御するメタ探索ポリシーを、過去の探索履歴をリプレイ・シミュレータとして利用して改善する。アルゴリズム工学、数値最適化、GPUカーネル工学で、探索品質を維持・向上させながらコストを削減したと報告されている。こちらは「エージェントの研究の仕方を、研究履歴を使って改善する」タイプの狭義RSIに近い。 ¹⁴

したがって、関連概念は次のように分けると理解しやすい。

概念	改善対象	再帰性	Vals RSI Indexとの関係
Goodの知能爆発	AI全体・後継システム	理論上、強い	歴史的起点。 15
Gödel Machine	自己プログラム	形式的・自己参照的	RSIの理論モデル。 16
Darwin Gödel Machine	エージェント自身のコード	実験的に閉ループ	「実際の自己改善」の研究例。 17
Dream-RSI	探索・オーケストレーション方策	メタ層で閉ループ	研究戦略自体の改善例。 18
Vals RSI Index	AI研究を遂行する能力	再帰ループは未実装	RSIに必要な能力を外部評価する指標。 4

ここから生じる最も重要な概念上の区別は、「AIがAI開発に貢献すること」「AIが一回の改善を成功させること」「改善後のAIがさらに優れた改善者になること」は別物だという点である。Anthropic自身も2026年時点で、AIがAI開発に占める役割が拡大している一方、完全な後継モデル開発まで自律化したRSIにはまだ到達しておらず、それが必然とも限らないと述べている。 19

計算方法・数式・タスク設計

Vals RSI Indexでは課題 t ごとに、ベースライン b_t 、理論的最良値 u_t 、エージェントが得た結果 x を使う。CompressionとLM TrainingはBPB（bits per byte）、Harness EngineeringとPost-Trainingは **error rate = 100% - accuracy** に変換するため、すべて「小さいほど良い」方向に統一される。 6

まず、ベースラインから理論上限までの対数距離のうち、どれだけ埋めたかを

$$f_t(x) = \text{clip}\left(\frac{\log(b_t/x)}{\log(b_t/u_t)}, 0, 1\right)$$

で求める。次に、選択した参照値 r_t が必ず0.5になるよう課題固有の指数

$$\gamma_t = \frac{\log 0.5}{\log f_t(r_t)}$$

を決め、

$$s_t(x) = f_t(x)^{\gamma_t}$$

とする。最終的な現行4課題版の指数は

$$\text{RSI Index} = \frac{1}{4} \sum_{t=1}^4 s_t(x_t)$$

である。無効な提出、またはベースラインより悪い結果は0点になる。 6

この構造により、

0 → ベースライン 0.373 → 現リーダー 0.5 → 参照水準 0.6 → Valsの主要マイルストーン 1 → 設定上の理論的最良

という読み方になる。0.5は「RSIの半分」ではなく、**各課題で選ばれたリファレンスに一致する地点である。** ³

公開されたアンカーから本報告で γ_t を再計算すると、概ね以下になる。値は丸めている。

現行課題	ベースライン	0.5の参照値	理論上限	推定 γ_t
Compression	2.92 BPB	NNCP 1.2435 BPB	0.6 BPB	約 1.123
LM Training	1.3699 BPB	0.85 BPB	0.6 BPB	約 1.265
Harness Engineering	精度50%=誤差50%	人間との一致93.6%=誤差6.4%	精度99%=誤差1%	約 1.077
Post-Training	精度35.38%=誤差64.62%	59.42%=誤差40.58%	精度99%=誤差1%	約 0.316

これはValsが示す「 γ_t は約0.32~1.26」という範囲とも一致する。 ⁶

Compressionでは100MBのWikipedia由来enwik8を損失なく圧縮する。1 H100、8 CPU、24時間、インターネットなしの環境で研究を行い、最終アーカイブには復号に必要なランタイム依存物も含め、完全なbyte-for-byte復元を検証する。基準は2.92 BPB、参照はFabrice BellardのNNCP約1.2435 BPB、設定上限は0.6 BPBである。 ²⁰

LM Trainingではenwik9のbyte-level言語モデルを固定計算予算下で改善する。800MBを学習、100MBをdev、100MBをtest相当として扱い、研究にはH100を使用する。提出された学習コードを隠しseedで3回、各30分再学習し、32,768個の隠しwindowで得たBPBの中央値を採点する。ベースライン1.3699、参照0.85 BPB、上限0.6 BPBである。Vals自身、0.85のAl-Rfouらの論文とは学習・評価条件が完全一致する再現実験ではないと明記している。 ²¹

Harness Engineeringでは、固定されたGemini 3.5 Flash-Liteを中心に、Webアプリが仕様を満たしているか判定するエージェント型評価ハーネスを改善する。プロンプト、ブラウザ操作、メモリ、検証、計画、サブエージェント等を変更できる。12時間、GPUなしで、開発時には12アプリの集約スコアが与えられ、最終的には24個の非公開アプリ・205サブステップで評価される。ベースラインは50%一致、人間参照93.6%、上限99%である。 ²²

Post-TrainingではQwen3.6-35B-A3BをFinance Agent v2向けに改善する。16基のRTX PRO 6000 96GB、30時間、インターネット接続、外部推論APIに100ドル、公開例27問という環境で、SFT、RL、合成データ生成など任意の方法を使える。最終チェックポイントと生成設定を提出し、126問の非公開セットを3回評価した平均がスコアになる。親モデル35.38%、参照Muse Spark 1.2が59.42%、上限99%である。 ²³

エージェントは単発チャットを30時間連続させるわけではない。モデルが途中終了すると外部supervisorが、元の課題、残時間、保存済みworkspaceを渡して新セッションを起動する。Marimo notebookが実験台帳として残される。したがって測っているのは「モデル単体」ではなく、**モデル+ネイティブハーネス+永続workspace+再起動supervisorからなる研究システムである。** ⁶

またCodex、Claude Code、Kimi Code、Qwen Code、Antigravity、Muse Code、Grok Buildなど、各社のネイティブハースネスを利用する。Valsはこれを意図的な設計としており、純粋な基盤モデル能力を分離するのではなく「実際にユーザーが利用できる研究システム全体」を比較している。 ⁶

評価結果・検証・再現性可他ベンチマークとの比較

2026年10月時点の公式ライブ版では24モデルが掲載され、総合トップはOpus 5.5である。論文版の分析対象は2025年11月～2026年9月に公開された8プロバイダー・19システムであり、ライブラリーボードはその後の追加評価を含むため、論文表と最新Web順位は完全には一致しない。 ²⁴

指標	現行の代表的トップ結果	読み方
総合RSI Index	Opus 5.5 0.3731	4課題平均。リファレンス平均0.5には未達。 ²⁵
Compression	Opus 5.5 1.31199 BPB / 0.465	1.2435付近の参照には未達。 ⁹
LM Training	Opus 5.5 0.830765 BPB / 0.530	0.85参照を超えた唯一の現行課題。 ²⁶
Harness Engineering	GPT-6 Sol 81.46% / 約0.228	人間参照93.6%には大きな差。 ⁹
Post-Training	Opus 5.5 44.53% / 0.351	親35.38%は改善、参照59.42%には未達。 ²⁶

論文版ではOpus 5.5が0.373、Fable 5.1が0.361、Opus 5が0.330、GPT-6 Astraが0.271だった。Anthropic系列は7世代で0.148→0.373、OpenAI系列は5世代で0.158→0.271、Moonshot系列は2世代で0.068→0.209に上昇したと報告される。 ⁷

ただし性能向上の内容は、しばしば「未知の原理の発明」よりも優れたエンジニアリングである。Compressionでは多数の安価なprefix実験とcontext mixing、LM TrainingではRoPE、RMSNorm、QK norm、ReLU²/SwiGLU等の既存要素の組み合わせ、Post-Trainingでは合成データ+LoRA/SFTなどが中心だった。Vals自身も「execution outpaced judgment」、すなわち実験の実行能力は強くなったが、高レベルな未知の研究方向を選ぶ研究判断力はまだ弱いと評価している。 ⁹

Post-Trainingは特に難しい。論文版19システムのうち15システムが0点となり、親モデルを2標準誤差超で上回ったのはOpus 5.5、Fable 5.1、Opus 5だけだった。Opus 5.5は親モデルのツール呼び出しで必須引数が約84%欠落していることを発見し、その失敗に狙いを定めた合成学習例を作った。これは「モデルを学習させる」だけでなく、**何を学習させるべきか診断する能力**が重要であることを示す。 ²⁷

また、改善を連続させることは一段と難しい。論文の追加実験では、Opus 5がfinance向けに改善したチェックポイントをさらにlegal領域で学習させたところ、legal性能もfinance性能も悪化した。研究者らは、過去領域のreplayやon-policy手法など、catastrophic forgettingへの既知の対策を適切に利用しなかった点を指摘している。これは一回の自己改善と、**改善を累積できるRSI**の間に大きな技術的隔たりがある証拠である。 ⁴

再現性は三層に分けて考えるべきである。第一に**成果物評価の再現性**は比較的強い。Compressionは完全round-trip、LM Trainingは隠しseed再学習、Harness/Post-trainingは固定成果物を複数回評価する。第二に**研究過程の再現性**は弱い。1モデル×1課題につき原則1研究ランしかない。第三に**第三者によるベンチマーク**

全体の再現性は混合的で、enwik8/enwik9課題は公開仕様から再構築できる一方、Harnessのhidden appsとFinance Agent v2の126問は汚染防止のため非公開である。⁴

公開データを使うことには逆方向の問題もある。論文はCompressionとLM Trainingについて、モデルの事前学習データに論文・コード・解法が含まれていた可能性、すなわち**data contaminationを排除できない**と明記する。一方で著者らは、既知結果の単純再現なら0.5付近には近づけても、0.6のように既知成果を大きく超えるには新たな改善が必要になると論じている。⁴

関連ベンチマークとの違いも重要である。

評価系	主な対象	人間比較	自己改善との距離	Valsとの主な違い
Vals RSI Index	LLM開発R&D	既知研究成果をanchorに使用	RSIに必要な研究能力	4課題、12〜30h、実成果物、native harness。 ⁶
RE-Bench	オープンエンドML研究	61名の専門家、71試行など直接的人間データ	間接的	同一課題でAIと人間の時間別比較が強み。 ²⁸
MLE-bench	75 Kaggle ML 競技	Kaggle medalを基準	間接的	ML engineeringの広い能力を測る。o1-preview+AIDEで少なくともbronze相当16.9%が報告された。 ²⁹
DGM	コーディングエージェントの自己改善	ベンチマーク性能比較	実際に自己コードを反復変更	評価指数ではなく、自己改善アルゴリズムそのもの。 ¹⁷
Dream-RSI	探索戦略の自己改善	タスクbaseline比較	メタ探索層で再帰	基盤モデルを固定し、探索方策を改善。 ¹⁸

この比較から、Valsの今後の検証で最も価値が高いのは**RE-Bench型のmatched-human studyをVals課題にも導入すること**である。現状の0.5は同一計算資源・同一時間・同一情報条件で人間研究者が達成した中央値ではないため、「AI研究者が人間研究者に追いついた」という文言を厳密に裏付けるには不足がある。これは本報告の分析であり、Vals論文自身も現行指数を「best known resultへの比較」として定義している。³⁰

加えて、バージョン管理は早急な改善点である。9月の公式情報ではCompression、LM Training、Parameter Golf、Harness Engineering、Post-Trainingの**5課題**だったが、10月1日の現行公式ページでは4課題である。削除理由は**未指定**。したがって9月の37.13%と10月の37.31%の差を純粋なモデル能力向上とみなしてはならない。⁵

実装事例・応用領域・特許・標準化

Vals RSI Indexそのものは、現時点では主に**第三者評価・フロンティアモデル監視の製品**として運用されている。ValsはこれをProprietary benchmarkとして24モデルに適用しており、モデル開発企業横断で共通のAI R&D能力を比較する用途を持つ。これは研究室内の自己改善システムを製品化したものではなく、「研究自動化がどこまで進んだか」を外から測る計測インフラである。³¹

産業・研究上の応用は少なくとも四つに整理できる。第一は**モデル選定**で、企業が「一般的なチャット性能」ではなく自動実験・自動ML研究能力でモデルを比較できる。第二は**AI研究自動化の進捗監視**。第三は**フロンティアAI安全評価**で、AIが次世代AI開発を加速できる能力そのものを先行指標として監視できる。第四は**AI研究エージェント設計**で、実験回数、早期終了、誤った測定の訂正、探索戦略などをログから分析できる。ValsはMarimoのexperiment ledgerも公開説明に利用している。 32

実際の自己改善技術側ではSakana AIのDGMが公開実装の代表例で、Apache-2.0ライセンスのコードが公開されている。Dream-RSIも公式研究コードが公開されている。したがって「RSIは完全に理論だけ」という状況ではなく、**コード、ハーネス、探索方策など限定された層では既に実験的自己改善が成立している**。ただし、それをフロンティア基盤モデルの世代交代まで一般化する証拠はない。 33

日本でもRSI自体は学術的な研究対象になり始めている。2026年の人工知能学会全国大会には「進化するAIサイエンティスト」のオーガナイズドセッションが設置され、8月の汎用人工知能研究会ではRSIそのものを扱う研究会資料が公開された。もっとも、**Vals RSI Indexを日本企業・大学が公式に採用している一次資料は、本調査では確認できなかったため「未指定／確認できず」とする**。 34

特許状況について、今回確認した公開情報・公開特許検索では、Vals RSI Indexそのものに対応するVals AI名義の特許番号を確認できなかった。公式サイトは「Proprietary」と明記するが、特許番号、特許出願番号、ライセンス方式については**未指定**である。したがって「特許が存在しない」と断定するのではなく、「公開特許として本調査では確認できない」とするのが適切である。 35

標準化状況も同様で、Vals RSI Index自体をISO、IEC等が標準化した事実は確認できないため、RSI Index固有の標準化は**未指定／未標準化**と評価する。一方、周辺ガバナンスにはISO/IEC 42001:2023があり、AIを開発・提供・利用する組織のAIマネジメントシステムの確立、運用、継続改善を規定している。経済産業省も同規格を公式に紹介している。ISO/IEC 23894:2023はAIリスク管理のガイダンスを提供する。これらはRSI性能を採点する標準ではないが、RSI研究を企業内で安全に運用する際の管理枠組みとして利用可能である。 36

現状の課題・技術的、倫理的、法的、運用リスク

最大の技術的課題は、「**研究課題で高得点を取ること**」と「**新しいAI研究を一般的に行えること**」の**外的妥当性**である。Valsは4つの狭い研究課題で測っており、フロンティア規模のモデル学習、数兆token規模のデータ設計、大規模分散学習、アーキテクチャ研究、ハードウェア共同設計などへの一般化を直接検証していない。論文自身が「frontier-scale developmentへのtransferは直接テストしていない」と明記する。 37

第二は**単一研究ラン問題**である。オープンエンド研究は偶然性が大きいにもかかわらず、各システム・各課題を原則1回しか研究させていない。たとえば、あるモデルが一回偶然優れたアイデアを引いた場合と、80%の確率で同等の成果を再現できる場合を区別できない。表示される誤差棒もこの不確実性を表していない。将来は少なくとも複数独立research runsと、between-run varianceの報告が必須である。 4

第三は**モデル能力とハーネス能力の交絡**である。Codex、Claude Code、Kimi Code等を固定せず各社native harnessを用いるので、順位はモデルの重みだけでなく、tool calling、context management、system prompt、memory、goal executionの品質を含む。これは「実際の製品システム」を評価する目的には合理的だが、基礎モデル研究の比較には不向きである。Valsもこの点を明示している。 6

第四は**指標設計依存性**である。0.5をどの研究成果に置くか、0.6 BPBや99%をどのような上限とするか、4課題を同じ重みで平均するかは、自然法則から一意に決まるわけではない。特にPost-Trainingの γ が約0.32、LM Trainingが約1.26と大きく異なるため、同じ生スコア改善でも指数への寄与の仕方が課題によって異なる。これは参照点を揃えるための数学的帰結だが、「RSI能力」という潜在変数の心理測定学的妥当性が証明されたわけではない。 6

第五は**ベンチマーク汚染とbenchmark gaming**である。公開されたenwik8/enwik9や既知論文はモデルの事前学習に含まれ得る。一方、課題を長期間公開すれば、モデル開発者がその形式に最適化し、スコアは上がっても一般研究能力は上がらない可能性がある。Valsは公開課題について汚染を排除できないと認めており、今後はsealed taskの定期ローテーションが重要になる。 4

第六は**再帰的改善の安定性**である。Valsのsequential post-trainingでは二段階目の改善で旧・新領域の双方が悪化した。DGMやDream-RSIのように探索ループを回しても、局所的なベンチマーク最適化、catastrophic forgetting、探索空間の縮退、reward hackingが累積すれば、「自己改善」が長期には自己劣化になる可能性がある。RRSIの研究も、自己進化型ハーネスがin-distributionタスクを記憶し、OOD性能の改善が消える問題を指摘している。 38

倫理・安全面では、AI R&D能力は典型的な**dual-use capability**である。安全なモデル、圧縮、効率化、医療・科学向けAIを高速開発できる一方、同じ能力は高度なサイバーシステム、監視、軍事応用、より制御困難なモデルの開発速度も高め得る。Anthropicも、完全なRSIには未到達としながら、AI自身が後継システムを設計・改良できる段階への備えが制度的に遅れる可能性を問題視している。 39

運用上は資源負担も大きい。論文の構成では1システムについて全課題を回すと**約528 GPU-hours**に相当し、さらにモデルAPI費用がかかる。論文表ではAPI費だけでモデルによって約99ドルから4,000ドル超まで差があり、Opus 5.5は約1,647ドルだった。これはGPUインフラ費用とは別である。したがって独立大学・小規模研究所が全モデルを高頻度で再評価するには相当な資源が必要である。 7

法的には、2026年10月時点で「RSI Indexが一定値を超えたら発動する」という日本の専用規制は確認できず、**Vals RSI Index自体に法的効力はない**。日本では2026年7月に人工知能基本計画が閣議決定され、信頼できるAI、AIエコシステム、戦略的自律性等が政策課題となっている。Vals型能力評価を政策指標として使う場合も、現段階では自主的・補助的なリスク指標と位置づけるのが妥当である。 40

EUではAI Actが2024年8月に発効し、GPAI関連義務は2025年8月から、法の大部分は2026年8月から適用されている。したがってRSI研究システムを実際に提供・展開する企業では、モデルの分類、リスク管理、セキュリティ、透明性等の既存義務と接続して検討する必要があるが、**ValsスコアそのものがEU AI Act上の公式リスク分類基準になっているわけではない**。 41

今後の研究・実用化展望

Valsは論文中で、Anthropic系列の指数上昇を年率約0.30ポイント、OpenAI系列を約0.17ポイントと線形近似している。その単純外挿ではAnthropic系列が**0.5を2027年3月、0.6を2027年7月**に通過し、0.6の95%区間を2027年3月～2028年5月としている。OpenAI系列については0.5が2028年2月、0.6が2028年9月という推定だが、データ点が少なく区間も広い。著者ら自身、この予測が「現在の線形トレンドが継続し、かつベンチマーク性能がフロンティア研究に一般化する」という条件付きであることを明記する。 7

したがって「2027年7月にRSIが起こる」という読み方は不適切である。0.6はValsが設定した「各課題の既知リファレンスを平均的に明確に超える」マイルストーンであって、後継モデルを自律的に繰り返し構築する能力を直接証明するものではない。論文も0.6について「もし一般化するなら、frontier improvementを駆動できる自律AI研究者の証拠になる」と条件付きで述べている。 42

本報告では、より現実的な研究ロードマップを次のように提案する。

時期	優先すべき研究・実装	成功条件
短期： 今後約 12か月	1課題あたり複数独立研究ラン、human matched baseline、標準ハーネス版との二重評価、バージョン固定、Parameter Golf 削除理由等のchangelog公開、sealed task 追加	「一回の偶然」ではなく再現可能な研究能力を測定できること。Vals自身もmultiple runs、data curation、kernel optimization、continued post-training等を将来課題に挙げる。 ⁴
中期： 約1〜3 年	連続post-training、data curation、kernel、分散学習、モデル設計を組み合わせたmulti-stage research、OOD評価、忘却対策、モデル+harness+data pipelineの共同最適化	一段の改善ではなく、改善を複数段累積しても旧能力を壊さないこと。 ⁴³
長期： 約3年 以上	改善したモデルを次ラウンドの研究エージェントとして投入するclosed-loop successor benchmark、より大規模な訓練への転移、独立安全評価・停止機構	$A_0 \rightarrow A_1 \rightarrow A_2$ と世代を進むにつれて「研究する能力」自体が再現性をもって上昇すること。これは現行Valsがまだ測っていない領域。 ⁴

特に「closed-loop RSI benchmark」は重要である。将来版では、単にモデル A がモデル B を改善したかではなく、

$$A_0 \xrightarrow{\text{AI R\&D}} A_1 \xrightarrow{\text{AI R\&D}} A_2 \xrightarrow{\text{AI R\&D}} A_3$$

について、

$$Q(A_{n+1}) > Q(A_n)$$

だけでなく、

$$R(A_{n+1}) > R(A_n)$$

すなわち**成果物品質 Q だけでなく「次世代を改善する研究能力 R 」自身も上昇するか**を測る必要がある。これは本報告による分析的拡張であり、現行Valsが「no agent trains a model that is then asked to train a successor」と明記する欠落部分を直接補う設計である。 ⁴

研究資源としては、単なるGPU量よりも、①非公開・更新型の研究課題群、②専門家による人間baseline、③完全なexperiment ledger、④複数ランを実施できる計算資源、⑤評価モデルとは独立した検証系、⑥成果物が外部環境へ影響しないsandbox、が重要になる。DGMも自己改変実験をsandboxとhuman oversightの下で行っており、より強力な自己改善実験ほど評価環境と実運用環境を分離する必要性が高まる。 ¹⁷

データ面では、公開データだけを使い続けるとcontaminationが増える一方、全データを非公開にすると科学的再現性が落ちる。このトレードオフを解くには、**公開再現課題+秘密holdout+定期更新されるfresh task**の三層構成が適している。Valsの現行方式は既に前二者を部分的に持っているため、今後は特に第三層の導入が重要になる。 ⁴

推奨アクション

最も重要なのは、Vals RSI Indexを「シンギュラリティまでの残距離」という単一数値として扱わず、**AI研究自動化の複数の能力を観測する早期警戒指標の一つ**として使うことである。現状の0.373→0.5→0.6という推移は重要なシグナルだが、それだけで完全RSIの発生確率や時期を算出することはできない。 ³⁷

対象	優先度	推奨アクション	理由
研究者	最優先	各モデル×課題を最低複数seed・複数research runで再評価し、between-run varianceを公開する	現行最大の統計的弱点が「研究ラン1回」であるため。 ⁴
研究者	最優先	人間研究者に同じ時間、GPU、情報、初期コードを与えるmatched-human studyを構築する	0.5を真の「同条件のhuman frontier」と解釈できるようにするため。 ⁴⁴
研究者	高	closed-loop successor testを導入し、改善後モデルの「研究能力」の向上まで測る	現行ValsはRSIの再帰ループを閉じていないため。 ⁴
Vals・評価機関	最優先	benchmark version、task追加・削除、anchor変更、スコア再計算規則を機械可読なchangelogとして固定する	5課題→4課題変更により異なる日付の指数比較が危険になっているため。 ⁴⁵
企業	最優先	RSI Indexを単独KPIにせず、研究成功率、再現率、OOD性能、安全性、API/GPUコストと併記する	指数はnative harnessやanchor選択を含む複合尺度だから。 ⁶
企業	高	自動AI R&D環境を本番モデル・顧客データ・外部ネットワークから分離し、sandbox、compute cap、checkpoint approvalを設置する	自己改善能力はdual-useであり、DGM等もsandbox・human oversightを採用している。 ¹⁷
企業	高	ISO/IEC 42001型のAIマネジメントとISO/IEC 23894型リスク管理に、自動R&D固有のcontrolを追加する	両規格はRSI専用ではないが、継続改善・リスク管理の組織基盤を提供する。 ⁴⁶
政策立案者	最優先	Vals、RE-Bench等の外部測定を組み合わせた「AI R&D automation capability monitoring」を構築する	開発企業内部の自己申告だけではラボ横断比較が難しいため。Valsが外部指数を作った直接の動機でもある。 ⁴⁷
政策立案者	高	特定スコアへの自動規制ではなく、複数指標が閾値を超えた場合に評価頻度・報告・安全試験を強化する段階制を検討する	0.5・0.6はVals独自のanchorであり、法的・科学的な普遍閾値ではないため。 ⁷
日本の研究政策	高	日本語・国内AI開発環境に対応するRSI/AI Scientist評価基盤を整備し、JSAIのAI Scientist・RSI研究と接続する	国内でもRSI研究は始まっているが、Vals採用・国内標準評価は本調査では確認できない。 ⁴⁸

研究者向けには特に、「新規性」と「実験遂行能力」を別々に採点することを推奨する。現行Valsでは、既知技法を極めて巧みに組み合わせる研究者と、本当に未知のアルゴリズムを発見する研究者が同一尺度に混ざる。将来的には、prior-art検索を伴うnovelty score、独立タスクへのtransfer、方法の一般化可能性を別軸に持たせるべきである。Vals自身のログ分析も、現在の改善の大部分が既存手法の統合であることを示している。⁸

企業向けには、RSI Indexが上がるほど「AI研究者を大量配置すればよい」と考えるのではなく、**研究エージェントが提案する変更と、その変更を本番モデルへ反映する権限を分離**することが重要になる。研究、評価、承認、デプロイを異なる権限境界に置き、自己生成コード・学習データ・checkpointのprovenanceを保

存する方が、自己改善の恩恵と制御可能性を両立しやすい。これはISO/IEC 42001の管理思想とも整合する。

49

政策面では、2026年時点で日本がAI基本計画に掲げる「信頼できるAI」と戦略的自律性の観点から、海外企業の自己申告だけでなく、国内または国際共同の独立評価施設を整備する価値が高い。とりわけ高性能モデルについて、通常的能力、安全性、サイバー能力に加えて「AI開発を加速する能力」を独立した評価カテゴリとして監視することが望ましい。

40

最終的に、Vals RSI Indexの真価は「0.6がいつ来るか」を一点予測することではなく、**AIがAI研究のどの工程を、どの程度、どの程度再現可能に人間から引き継ぎつつあるかを定量的に観察する仕組みを作ったこと**にある。一方で、完全なRSIを主張するためには、現在欠けている「世代をまたぐ閉ループ」「複数ランでの安定性」「未知課題への一般化」「同条件の人間比較」「改善を重ねても壊れない累積性」を実証する必要がある。Vals自身が認めるこれらの限界を踏まえる限り、2026年10月時点の最も厳密な表現は、「**RSIは実証されていない。しかし、その前段階である自律AI研究開発能力は、測定可能な速度で上昇している**」である。

50

1 2 8 9 25 26 32 Vals RSI Index Leaderboard and Methodology

https://www.vals.ai/benchmarks/rsi_index?utm_source=chatgpt.com

3 4 6 7 27 30 37 38 42 43 50 alphaxiv.org

<https://www.alphaxiv.org/abs/2609.rsi-index-autonomous-llm-rd.pdf>

5 45 Home

https://www.vals.ai/home?trk=public_post-text&utm_source=chatgpt.com

10 15 Speculations Concerning the First Ultraintelligent Machine - ScienceDirect

https://doi.org/10.1016/S0065-2458%2808%2960418-0?utm_source=chatgpt.com

11 48 RSI（再帰的自己改善）：シンギュラリティをもたらすメカニズムとは

https://www.jstage.jst.go.jp/article/jsaisigtwo/2026/AGI-033/2026_01/_pdf/-char/ja?utm_source=chatgpt.com

12 16 Goedel Machines: Self-Referential Universal Problem Solvers Making Provably Optimal Self-Improvements

https://arxiv.org/abs/cs/0309048?utm_source=chatgpt.com

13 17 Darwin Godel Machine: Open-Ended Evolution of Self-Improving Agents

https://arxiv.org/abs/2505.22954?utm_source=chatgpt.com

14 18 Dream-RSI: Recursive Self-Improvement through Evolving Worlds

https://arxiv.org/abs/2609.14858?utm_source=chatgpt.com

19 39 When AI builds itself

https://www.anthropic.com/institute/recursive-self-improvement?utm_source=chatgpt.com

20 21 22 23 Vals RSI Index Leaderboard and Methodology | Vals AI

https://www.vals-ai.com/benchmarks/rsi_index

24 31 35 AI Benchmarks and LLM Leaderboards | Vals AI

https://www.vals.ai/benchmarks?utm_source=chatgpt.com

28 44 RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts

https://arxiv.org/abs/2411.15114?utm_source=chatgpt.com

29 MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering | OpenAI

https://openai.com/index/mle-bench/?trk=article-ssr-frontend-pulse_little-text-block&utm_source=chatgpt.com

33 The Darwin Gödel Machine: AI that improves itself by ...

https://sakana.ai/dgm/?utm_source=chatgpt.com

34 [5L1-OS-30-05] 進化するAIサイエンティストのメタデータ分析 ...

https://pub.conf.it.atlas.jp/ja/event/jsai2026/presentation/5L1-OS-30-05?utm_source=chatgpt.com

36 46 49 ISO/IEC 42001:2023 - AI management systems

https://www.iso.org/standard/42001?utm_source=chatgpt.com

40 人工知能基本計画 令和8年7月 14 日 閣議決定

https://www8.cao.go.jp/cstp/ai/ai_plan/aiplan_20260714.pdf?utm_source=chatgpt.com

41 AI Act | Shaping Europe's digital future - European Union

https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai?utm_source=chatgpt.com

47 RSI-Index: Measuring Autonomous LLM R&D in Frontier ...

https://www.alphaxiv.org/abs/2609.rsi-index-autonomous-llm-rd?utm_source=chatgpt.com