Open-Ended AI 研究における自律型エージェントの限界と可能性

Kirgis ら（2026）の「シャドー評価」論文の精査と AI 研究自動化の現状分析

Gemini 3.7 Flash

1. 論文の概要と研究背景

人工知能（AI）による研究開発の自動化と再帰的自己改善（Recursive Self-Improvement: RSI）は、テクノロジー予測と AI 安全性の双方に関わる重要な論点である。Kirgis らは、主要 AI 研究機関がフロンティアモデルによるモデル構築・ポストトレーニング支援の事例を報告する一方、AI エージェントが未解決でオープンエンドな AI 研究を自律的に遂行できるかについては、厳密な実証がなお限定的であると指摘する。 [1]

プリンストン大学、英国 AI 安全研究所（UK AISI）、トロント大学、スタンフォード大学などの研究者 24 名（Peter Kirgis、Sayash Kapoor、Arvind Narayanan ら）は、論文「Can AI agents conduct open-ended AI research? Early evidence from two case studies」（arXiv:2607.27191、2026 年 7 月初稿、同年 8 月改訂）を公表した。同論文は、AI 研究自動化の新たな評価手法として「シャドー評価（Shadow Evaluations）」を提案し、2 件のケーススタディから初期的証拠を示した。 [1, 2]

従来の評価は、成功指標をプログラム的に定義できる「自動検証型評価」と、AI 生成論文を学会・ワークショップの匿名査読に付す「ブラインド人間査読型評価」に大きく分かれていた。前者は客観的・反復可能な採点に適するが、重要な問いの選択、仮説の構築、失敗時の方向転換などを直接測りにくい。後者はオープンエンドな研究を扱える一方、査読の確率的変動、査読者の専門性、試行総数が開示されない場合の選択的報告といった限界がある。 [1]

シャドー評価では、未公開の高品質な研究論文の中心的な問いをエージェントに与え、その論文の原著者が成果物を査読する。エージェントは正解を検索できず、評価者は当該問いを長期間検討してきた専門家である。本研究の主実験は Claude Opus 4.8 と OpenClaw で 2 課題について実施され、その後、TabPFN 課題について GPT-5.6 Sol Ultra と Codex を用いた頑健性確認が行われた。 [1, 2]

2. 実験設計とリソース・人手介入の条件

シャドー評価の設計上の核心は、「研究問いの汚染（Contamination）の低減」と「当該問いに深い専門知識を持つ査読者による評価」を両立させる点にある。未公開の問いを用いることで、訓練データの丸暗記やウェブ検索による直接的な解答を避け、原論文著者による詳細な評価を可能にする。もっとも、原論文著者は AI 生成物であることを知っており、自ら採用した方法を選好する可能性もあるため、非盲検性に由来するバイアスは残る。 [1]

表 1　AI 研究自動化の主な評価パラダイム

評価パラダイム	対象タスク	評価者・指標	主な利点	主な限界
自動検証型評価 (Verifiable Evaluations)	正解率、学習損失、実行速度など、事前に定義された指標の最適化	自動検証スクリプト、実行結果、ランキング指標	高い客観性と再現性。比較・反復・大規模化が容易	問いの設定、研究上の価値判断、方針転換を直接評価しにくい
ブラインド人間査読型評価 (Blind Peer Review)	論文全体の生成と学会・ワークショップへの投稿	匿名査読者、会議・ワークショップの採否	オープンエンドな研究成果を実際の査読過程で評価できる	査読の質と判断の変動、試行総数非開示時の選択的報告リスク
シャドー評価 (Shadow Evaluations)	未公開論文の中心的でオープンエンドな研究問い	当該問いを研究した原論文著者による詳細査読	汚染リスクを低減し、深い専門知識に基づく評価と過程分析が可能	小標本、客観的正解の不在、非盲検性、研究者の選択裁量

出所：Kirgis et al.（2026）の整理を基に作成。[1]

対象となったのは、NeurIPS 2026 に提出された未公開の 2 論文である。第 1 の「Personas」課題は、LLM のペルソナを重み空間上の構造化された特性空間として分解・測定・制御できるかを問うもので、重みベース介入、次元の独立性・合成可能性、未訓練の下流行動への波及が評価対象となった。第 2 の「TabPFN」課題は、表形式 Prior-Fitted Networks において、無害な共変量シフトと精度低下を招く有害な分布シフトを区別し、運用時に較正された警報を出す検知器を構築できるかを問うものであった。 [1]

主実験では、原論文著者が「選定された一つの研究問いを実質的に検討するために十分」と見積もった予算として、各ランに 6 日間の壁時計時間、3,000 ドルの API クレジット、AWS 上の Linux 仮想マシン、数百 GPU 時間相当の計算資源が与えられた。ただし、原論文著者が研究全体に費やした期間・GPU 時間と同等以上だったわけではない。 [1]

文献調査、GPU 環境のデバッグ、多数の実験、外部 AI 査読の取得、LaTeX 原稿の作成といった主要な研究エンジニアリング作業は、エージェントがほぼ自律的に遂行した。ただし完全無介入ではなく、OpenClaw の不具合に対する手動修正、24 時間の期限延長、可読性改善の指示などの人手介入があった。 [1, 2]

さらに、スキャフォールド固有の影響を検討するため、TabPFN 課題を GPT-5.6 Sol Ultra と Codex で再実施した。研究全体のデータは、予備実験 2 回、主実験 2 回、頑健性確認 1 回の計 5 回の実行に基づく。 [1]

3. 主要な実験結果と成果の分析

6 日間の実行後に提出された 2 本の論文について、原論文著者は NeurIPS 形式の査読票と、AI 会議の公式査読を模した 1〜6 の総合評価尺度（1: Strong Reject、6: Strong Accept）を用いて評価した。Personas 論文は 2/6、TabPFN 論文は 1/6 で、いずれも明確な不採択判定であった。 [1]

表 2　原論文著者による評価結果

評価項目	Personas 論文	TabPFN 論文	査読コメントの要旨
品質 (Quality)	2 / 4 (Fair)	1 / 4 (Poor)	データ・実験の選択に十分な原理的根拠がなく、結論が証拠から導かれていない。
明確性 (Clarity)	1 / 4 (Poor)	2 / 4 (Fair)	記述が稠密で不透明であり、重要な情報と周辺的な実装詳細の区別が難しい。
重要性 (Significance)	2 / 4 (Fair)	2 / 4 (Fair)	先行研究に対する位置付けと優位性の説明が弱く、学術的インパクトが限定的。
独創性 (Originality)	3 / 4 (Good)	2 / 4 (Fair)	新しいデータセットや設定はあるが、主要部分は既存手法の延長にとどまる。
総合評価 (Overall)	2 / 6 (Reject)	1 / 6 (Strong Reject)	両論文ともトップ会議水準に到達せず、明確な不採択。
査読者の確信度	4 / 5 (Confident)	5 / 5 (Absolutely Certain)	査読者はいずれも評価に高い確信を示した。

出所：Kirgis et al.（2026）, Table 1 および原論文著者の査読。 [1, 2]

もっとも、成果が皆無だったわけではない。原論文著者は、文献調査の広さ、数百 GPU 時間に及ぶ実験の遂行、初期仮説の妥当性を一定程度評価した。Personas 課題では、ミスアラインしたスタイルだけに狭く微調整しても、広範なミスジェネラリゼーション（望ましくない一般化）には波及しなかったという、潜在的に興味深い反直観的結果も得られた。問題は、こうした要素をトップ会議水準の研究成果へ発展させられなかった点にある。 [1]

中心的な観察は、「研究エンジニアリング能力」と「科学的発見・研究判断能力」の乖離である。エージェントは GPU 管理、実験実行、外部査読の取得、LaTeX コンパイル等を広範に実行した一方、研究の中核では、手作業で選んだ例や小規模な合成データに依存し、十分な証拠なしに否定的結論を一般化した。 [1]

リソース配分では「使わなさすぎ」と「早く使いすぎ」の双方が見られた。OpenClaw による 2 件の主実験は、3,000 ドルの API 予算の半分未満（Personas 1,130 ドル、TabPFN 1,235 ドル）しか使わず、時間を残して終了した。反対に、GPT-5.6 Sol の頑健性確認では約 2 日で 3,000 ドルを使い切り、約 100 時間を残してトークン予算が枯渇した。これは、時間・API・GPU を研究工程に合わせて配分する能力の不足を示す。 [1]

TabPFN ランは単一仮説だけを試したのではなく、最初の 14 時間に 6 つの異なるアプローチを検討して棄却した。しかし、その後も 110 時間以上が残っていたにもかかわらず、プロジェクト全体の方針を組み替えず、否定的結果の論文化へ進んだ。 [1]

4. 5 回の実行で反復して観察された 5 つの失敗モード

ログ解析では、主実験 2 回、予備実験 2 回、頑健性確認 1 回の計 5 回を通じて、次の 5 つの失敗モードが反復して観察された。これは限定された実行条件下の所見であり、一般的・普遍的な失敗パターンとして立証されたものではない。 [1]

第 1 は「研究水準（Bar for Publishable Research）に関する判断不足」である。エージェントは、トップ会議に必要な証拠の量と質を適切に見積もれず、小規模な合成データや手動選択例から仮説を早期に棄却し、十分に裏付けられていない否定的結果を実質的な貢献として扱った。 [1]

第 2 は「研究デザインの不備に対する非創造的な対応」である。初期には原論文著者も一定程度評価する仮説を提示できたが、実験の失敗や AI 査読からの批判に直面すると、概念枠組みや実験設計を再構成できなかった。批判への対応として限定条件や留保を追加し、結果的に研究の主張を狭めた。 [1]

第 3 は「プロジェクトレベルの効果的なバックトラックの欠如」である。バグ修正やハイパーパラメータ調整などの局所的な修正は行ったが、アプローチ全体が袋小路に入ったと判断し、過去の成果を捨てて探索段階から再始動することはできなかった。クリーンな文脈を持つサブエージェントを起動できても、再スタートにはほとんど用いなかった。 [1]

第 4 は「リソースおよびタイムラインの配分・認識の不備」である。トークン、GPU 予算、残り時間を確認する手段があっても、それらを工程別に配分できなかった。時間を残して拙速に終了する場合と、API 予算を早期に使い切る場合の双方が生じた。 [1]

第 5 は「指示漂流（Instruction Drift）とコンテキスト管理の不備」である。長時間の実行中に、探索時間、外部査読の実施頻度、論文のページ制限など、当初の明示的指示が次第に守られなくなった。最終論文はいずれも 9 ページ制限を超え、形式上もデスク・リジェクトの対象となり得る状態だった。 [1]

5. 専門家評価・AI 査読シグナル・事前アンケートおよびコミュニティの反応

原論文著者の定性評価では、Personas 論文について、実験・手法の選択が理解しにくく、結果が後付けの選択に見え、過度に防衛的な文章が実際の作業と発見を覆い隠しているとされた。TabPFN 論文については、内部機能を用いた少数の不成功なシグナルから「利用可能なシグナルが存在しない」と一般化することは、限定例からの過剰な推論であり科学的に不十分だとされた。 [1, 2]

エージェント内部の自己査読サブエージェントは、複数回の改訂・自己査読を通じて一度も採択判定を出さず、おおむね Weak Reject を提示した。AI 査読は、人間査読者が後に問題視したサンプル選択、根拠不足、記述の不透明さなどを相当程度捉えていたが、重大な問題と軽微な問題の優先順位付けが不十分だった。 [1]

表 3　AI 査読シグナルと人間査読の比較

査読主体・ツール	判定	評価の特徴とエージェントの反応
自己査読サブエージェント	概ね Weak Reject	人間査読と共通する懸念を多数提示したが、重大度の優先順位付けが弱く、エージェントは根本的な再設計より留保の追加を選んだ。
Stanford Agentic Reviewer	早期ドラフトで Accept	外部ツールの中で最も寛大な評価。エージェントはこの採択評価を過度に重視し、最終報告書で重要な文脈として引用した。
CMU Paper Reviewer	採択判定なし／批判的	内部文書を変換したような文章、少数セルへの依存など、主要な弱点を指摘した。
人間専門家（原論文著者）	Reject 2/6／Strong Reject 1/6	AI 査読と多くの懸念点は共通したが、手動選択例や合成例への依存などをより致命的な問題として優先した。

出所：Kirgis et al.（2026）, Sections 4.4-4.5 および Table 4。 [1]

この結果は、AI 研究における Generator-Verifier Gap（生成能力と検証能力の非対称性）の可能性を示唆する。ただし、評価対象の 2 論文はいずれも不採択だったため、AI 査読が研究品質を正確に識別したのか、それとも一様に厳しい判定を出しただけなのかは区別できず、Verifier の正確性は確立されていない。 [1]

12 名の共同研究者を対象とした事前アンケートでは、Weak Accept 以上を得る確率の予測中央値は 30%だったが、予測は 10%以下から 95%まで大きく分かれ、回答者の確信度も低かった。エラーの無限ループ、結果の捏造・誤報、p ハッキングや選択的報告なども懸念されていたが、実験では重大なデータ改ざんは確認されず、エージェントは否定的結果を報告した。 [1]

公開後、一部の業界メディアやオンライン投稿では、6 日間という期間、汎用スキャフォールドの制約、研究判断を AI に全面委任することの難しさなどが論点となった。ただし、これらは体系的な反応調査ではなく、オンラインコミュニティ全体の支配的見解を示すものではない。 [11-13]

6. AI 研究自動化（Autonomous AI R&D）の現状と実務的示唆

2026 年 8 月 20 日時点の AI 研究自動化の評価・実証は、大きく「自動検証型」「ブラインド人間査読型」「専門家評価型」の 3 系統に整理できる。専門家評価型のうち、Kirgis らの厳密なシャドー評価と、公開済み研究を別の専門家が比較評価する類似手法は、対象の公開状態と評価者の関係が異なるため区別する必要がある。 [1, 3-10]

表 4　AI 研究自動化の主要な評価・実証系統

系統	代表例	課題設定・評価方法	現状と留意点
自動検証型評価	CORE-Bench [4] MLE-Bench [5] RE-Bench [6] Autoresearch [8] AlphaEvolve [7]	再現、Kaggle 型競技、研究エンジニアリング、学習損失やアルゴリズムの改善を、自動検証可能な指標で採点する。	明確な指標の下で高い成果を示す例がある。一方、重要な問いの選択や抜本的な方針転換を直接評価するものではない。
ブラインド人間査読型評価	AI Scientist-v2 [9] Zochi [10]	エンドツーエンドで研究を生成し、学会・ワークショップの二重盲検査読に提出する。AI Scientist-v2 は 3 本を提出し、1 本が採択基準を上回ったが事前合意により撤回。Zochi は ACL 2025 採択が報告されたが、原稿準備・内部レビュー・rebuttal に人間が関与した。	実際の査読過程を利用できるが、査読の変動、試行総数、選択的報告、人手関与の範囲を確認する必要がある。
シャドー評価型	Kirgis et al.（2026）／CRUX シャドー評価 （使用スキャフォールド：CRUX 2）[1, 2]	未公開論文の中心的な問いをエージェントが解き、当該論文の原著者が詳細に評価する。	汚染を抑え、深い専門評価が可能。2 課題・5 回の実行では、研究エンジニアリングは広範に遂行したが、研究判断と方向転換に課題が残った。
類似する専門家評価型	Ahmed et al.（2026）[3]	公開済みの自然科学研究を対象に、高度な AI フレームワークの成果を専門家が原研究と比較する。評価者は元研究の著者ではない。	専門家によるオープンエンド評価という点で類似するが、未公開課題・原著者評価というシャドー評価の条件とは異なる。

注：Kirgis et al.（2026）の Table 2 および Appendix D と各一次資料を基に再整理。[1, 3-10]

本研究から直接比較実証された結論ではないが、実務上は Human-in-the-Loop による役割分担が有力な運用形態の一つと考えられる。人間研究者が、高次の問いの設定、仮説の選択、証拠水準の判断、失敗時の抜本的な方向転換を担い、AI エージェントが、文献調査、実験環境の構築・実行、データ処理、反復的な下書き作成を担う構成である。 [1]

したがって、AI 研究自動化の導入評価では、「コードが書けるか」「実験を回せるか」だけでなく、研究課題の価値判断、反証可能な実験設計、否定的結果からの再探索、残予算と時間の再配分、長期指示の保持を独立した評価項目として設ける必要がある。 [1]

7. 結論と研究の限界

Kirgis らの研究は、2 件の未公開 NeurIPS 2026 研究課題について、2 回の予備実験、2 回の主実験、1 回の頑健性確認からなるシャドー評価を行った。その結果、現行のフロンティア AI エージェントは、文献調査、実験環境の構築・デバッグ、実験実行、外部査読の取得、論文作成などの研究エンジニアリングを広範に遂

行できる一方、トップ会議水準のオープンエンド研究に必要な研究水準の判断、研究設計の再構成、行き詰まりからの抜本的な方向転換、リソース管理、長期的な指示保持に課題を残すことを示した。 [1, 2]

もっとも、これは 2 つの研究課題と計 5 回の実行に基づく初期的証拠である。非盲検評価によるバイアスの可能性、客観的な正解の不在、研究問いの選択、モデル・スキャフォールドの制約、解釈上の裁量がある。したがって、AI が一般にオープンエンド研究を遂行できないことを決定的に立証したものではなく、特定のモデルと運用条件の下で、全面的な研究自動化に重要な課題が残ることを示した初期検証と位置付けるのが適切である。 [1]

今後は、対象分野・研究課題・モデル・スキャフォールドを拡張し、自動検証、ブラインド査読、シャドー評価、類似する専門家評価を相互補完的に用いることが望まれる。同時に、研究の全面委任ではなく、AI の高速なエンジニアリング能力を人間の研究判断と結合する運用を検証していくことが、短期的には最も重要な課題である。 [1, 3]

引用文献

ウェブ資料の最終確認日：2026 年 8 月 20 日。タイトル部分はクリック可能なハイパーリンク。

[1] Kirgis, P., Kapoor, S., Schwartz, A., et al. (2026). Can AI agents conduct open-ended AI research? Early evidence from two case studies. arXiv:2607.27191v2.

[2] CRUX (2026). Can AI agents conduct open-ended AI research? Early evidence from two case studies (project materials). Expert reviews, survey responses, repositories and logs.

[3] Ahmed, M. O., Amale, S. A., Bhavsar, R. D., et al. (2026). Real science is harder than benchmarks: Evaluating advanced AI frameworks on published studies. I. Uncertainty quantification, ML on Therapeutic Data Commons, and agent-based modeling. bioRxiv:2026.06.24.734302.

[4] Siegel, Z. S., Kapoor, S., Nadgir, N., Stroebl, B., & Narayanan, A. (2024; rev. 2026). CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark. arXiv:2409.11363v2.

[5] Chan, J. S., Chowdhury, N., Jaffe, O., et al. (2025). MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. ICLR 2025.

[6] Wijk, H., Lin, T. R., Becker, J., et al. (2025). RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents against Human Experts. Proceedings of ICML 2025, PMLR 267, 66772-66832.

[7] Novikov, A., Vũ, N., Eisenberger, M., et al. (2025). AlphaEvolve: A coding agent for scientific and algorithmic discovery. arXiv:2506.13131.

[8] Karpathy, A. (2026). autoresearch: AI agents running research on single-GPU nanochat training automatically. GitHub repository.

[9] Yamada, Y., Lange, R. T., Lu, C., et al. (2025). The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. arXiv:2504.08066.

[10] Intology (2025). Zochi Publishes A* Paper. Intology Blog, 27 May 2025.

[11] AI Weekly (2026). AI agents finish the engineering of AI research, miss the science.

[12] StartupHub.ai (2026). AI Agents Stall on Core AI Research.

[13] Reddit, r/PracticalAgenticDev (2026). AI agents did the engineering, but failed the research. Online discussion; not a representative survey of the wider community.