ChatGPT

ScientistTwo：自律型AIは「人類の知識フロンティア」を本当に切り拓いたのか

エグゼクティブサマリー

ScientistTwo: Pioneering the Human Knowledge Frontier with Autonomous AI は、Google Cloud AI Research を中心とするチームが2026年9月17日に公開した、自律型科学研究エージェントのプレプリントである。人間の専門家が「研究すべき問題」を与えた後、既存研究の限界抽出、仮説生成、ベースライン再現、コード実装、実験、アイデア進化、フルベンチマーク、アブレーション、論文執筆、模擬査読、反論実験、メタレビューまでを、多数の専門エージェントが原則として人間の介入なしで回す。著者は Jaehyun Nam、Jinsung Yoon、Yanzhou Pan、Yubo Wang、Rui Meng、Parthasarathy Ranganathan、Tomas Pfister。Yubo Wang は University of Waterloo、他の著者は論文上 Google Cloud AI Research 所属である。原論文はまだ**査読済み出版物ではなく arXiv プレプリント**である。主要一次資料は arXiv 論文、HTML全文、公式プロジェクトサイト である。 [1]

結論を先に言えば、**ScientistTwo は現在公開されている「AI scientist」系システムの中でも、実験の広さ、研究サイクルの閉ループ化、アブレーション、査読→反論実験→再執筆、成果物整合性監査までを統合した点で非常に強い研究成果**である。一方で、「AIが人類の知識フロンティアを自律的に切り拓いた」「人間のトップ研究者を超えた」という表現は、現時点の証拠よりかなり強い。評価対象は主として既にトップ会議で受理された機械学習論文を起点とする107課題であり、完全な白紙状態から重要な科学問題そのものを発見したわけではない。また、主要な「論文品質」評価はAI査読者によるもので、ScientistTwo自身が最適化に用いる ScholarPeer と、開発時未使用の Stanford Agentic Reviewer で測定される。後者自身も「AI生成レビューには誤りがあり得る」「最初の15ページのみを解析する」と明記しているため、**AI査読スコアを実際の ICLR/ICML/NeurIPS 査読通過率と同一視してはいけない**。 [2]

数字はそれでも印象的である。107課題のうち86課題、すなわち **80.4%** で改善案を生成し、著者の集計では成功課題の平均相対改善は **25.2%**。生成された86論文は ScholarPeer で平均7.5/10・accept率91.9%、開発時には未使用だった Stanford Agentic Reviewer では平均5.7/10・accept率72.1%だった。比較対象の ScientistOne はそれぞれ3.8/10・14.3%、4.1/10・0%であり、大幅な差である。ただし72.1%は「107課題全体」ではなく**生成に成功した86件に対する条件付き成功率**である。失敗21件を不採択として数え直すと、Stanford評価でのエンドツーエンド採択相当率は約 **58.0%**、ScholarPeerでは約 **73.9%** となる。この違いは、論文の見出し的な数字を解釈するうえで重要である。 [3]

さらに重要なのがレビュー・ループの挙動である。査読・反論なしでは Stanford accept率49.0%、1回の査読・反論で73.5%へ上がるが、2回では69.4%へ**低下**する。一方、ループ内の ScholarPeer は 46.9%→79.6%→93.9%と上昇する。これは、初回フィードバックは一般化する改善を生む一方、反復を重ねると内部査読者への過適合が始まる可能性を示す強いシグナルである。したがって本研究の最も説得力のある成果は「AIが人間研究者を超えた」ことではなく、**外部実験を伴う閉ループ研究工程が、単発型AI研究エージェントより明確に強いことを大規模に実証した**点だと私は評価する。 [4]

総合判断	評価
技術的新規性	**高い**。仮説→実験→アブレーション→査読→追加実験→メタレビューを一つの閉ループに統合。 [5]

総合判断	評価
実験規模	**非常に強い**。107のトップ会議由来課題で評価。 [6]
「完全自律」の妥当性	**限定付きで妥当**。問題設定後の工程は自律だが、問題自体、既存論文・コード・評価仕様という人間由来の足場がある。 [7]
「人間を超えた」の証拠	**部分的**。AI査読では一部人間論文平均を上回るが、人間査読33論文では総合的に「同等」、方法論の厳密さでは人間がわずかに優位。 [8]
再現可能性	**中〜高だが未完成**。監査は非常に良い一方、公開コード・ライセンスを公式ページ上で確認できず、Gemini/Claudeという変動するプロプライエタリモデルにも依存。 [9]
商用成熟度	**未確立**。API、価格、商用製品、コードライセンスは本調査時点で明示されていない。 [10]
私の総合評価	**研究プラットフォームとして非常に重要。ただし「自律的科学者が人類の科学を超えた」というより、「既存の計算機科学研究課題を高速に探索・改善・論文化するAI研究所」の実証と読むべき。**

出自、位置づけ、関連プロジェクト

基本情報とプロヴェナンス

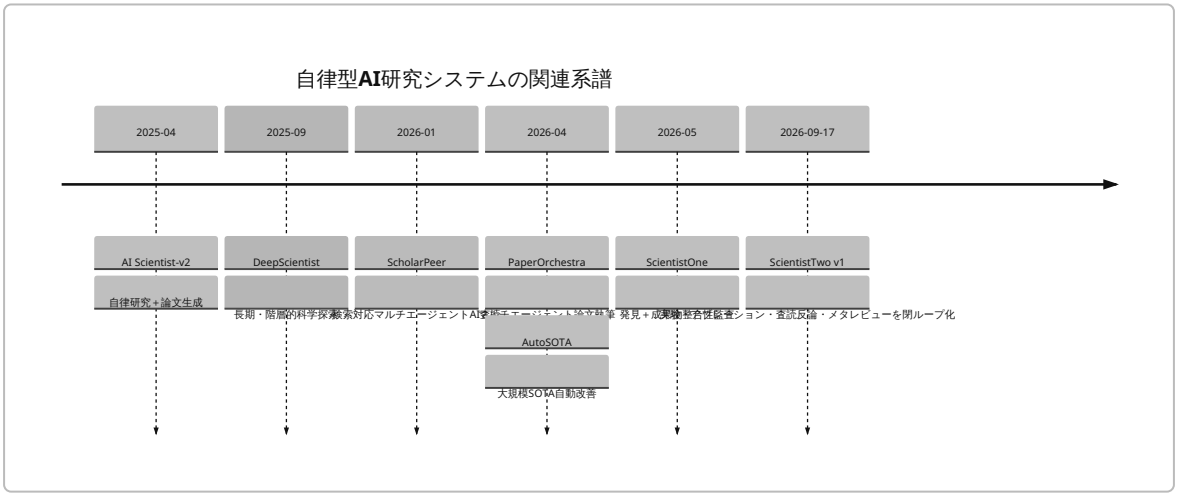
ScientistTwo v1 は **2026年9月17日 03:37 UTC** に arXiv へ投稿された。原稿は CC BY 4.0 で公開されているが、これは論文本文のライセンスであって、ScientistTwo本体のコードや生成コードのライセンスを意味しない。本調査で確認した論文・公式サイトには、ScientistTwo本体の公開GitHubリポジトリやコードライセンスへの明確なリンクを確認できなかったため、**コード公開状況・ライセンスは「未指定／公開確認できず」**とするのが妥当である。 [11]

項目	確認内容
正式名称	ScientistTwo: Pioneering the Human Knowledge Frontier with Autonomous AI [12]
公開日	2026年9月17日、arXiv v1。 [13]
著者	Jaehyun Nam, Jinsung Yoon, Yanzhou Pan, Yubo Wang, Rui Meng, Parthasarathy Ranganathan, Tomas Pfister。 [14]
所属	Yubo Wang: University of Waterloo、その他: Google Cloud AI Research。 [14]
査読状況	arXivプレプリント。arXiv自体は掲載論文を査読しない。 [15]
公式サイト	scientist-two.github.io [10]
論文	arXiv:2609.19644 [12]
論文ライセンス	CC BY 4.0。 [16]
コードライセンス	**未指定／確認できず**。公式ページでは生成物のギャラリーはあるが、ScientistTwo本体の公開コードライセンスは確認できない。 [10]

項目	確認内容
専用データセットのライセンス	**未指定**。そもそも単一新規データセットではなく、107論文の個別タスク・既存データセット群を使用する。 17
資金源	**未指定**。論文中に独立した funding / acknowledgments 情報を確認できない。企業内研究であることと、専用研究資金の出所は別問題である。 14
商用API・価格	**未指定／未発表**。 10
1課題あたり実行コスト	著者報告平均約 **US$3,765**、期間約 **2.5日**。 18

研究系譜

ScientistTwoは突然現れた単独システムではない。Sakana AI の AI Scientist-v2 は人手のコードテンプレートなしで研究から論文までを自動化し、生成した3論文をICLRワークショップへ投稿して1本が受理された。DeepScientist は長期間の階層的仮説探索、AutoSOTA は既存論文の性能改善の大規模自動化、ScientistOne は研究成果物の「正しさ」監査、PaperOrchestra は論文執筆、ScholarPeer はAI査読を進めた。ScientistTwoはこれらの方向を、**研究ループ全体を閉じる方向へ統合**したものと見るのが自然である。 19



この系譜からも、ScientistTwoの本当の新規性は単に「LLMに研究させる」ことではなく、**科学研究を、状態を持った反復最適化プロセスとして実装したこと**にある。 5

研究内容と技術アーキテクチャ

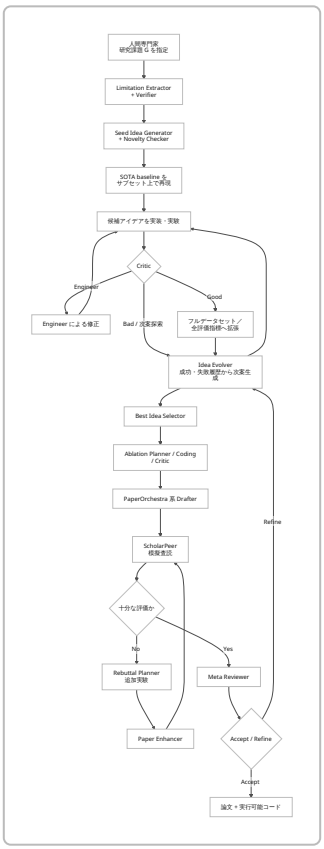
目的とスコープ

論文が掲げる究極目標は「problem-driven autonomous research」である。人間の専門家が根本的課題 G を指定すると、AIが既存研究のボトルネックを発見し、SOTAベースラインを確立し、新仮説を生成し、検証し、結果に応じて仮説を進化させ、最終的に論文 P^+ と実行可能コード C^+ を作る、という構図である。論文はこれを “without human intervention” と表現する。 13

ただし、この「完全自律」は厳密には**問題設定後の自律性**である。人間が何を研究すべきかを決定し、評価では既存のトップ会議論文、その問題定義、利用可能コードや評価プロトコルを出発点としている。したがって「研究課題の社会的重要性を自分で発見するAI」や「自然科学の未知現象を観測して新理論を創るAI」まで実証したわけではない。 7

エージェント構成とアルゴリズム

ScientistTwoは単一の巨大LLMに一度で研究させるのではなく、役割を分けたエージェントと反復ループを使う。主要工程は、Limitation Extractor/Verifier、Novelty Checker、Idea Generator、実験用Coding Agent、Critic/Engineer、Idea Evolver、Best Idea Selector、Ablation Planner/Coder/Critic、Initial Drafter、Peer Reviewer、Rebuttal Planner/Coding、Paper Enhancer、Meta Reviewer で構成される。 5



サブセット上で早期に候補をふるい落とし、有望な案だけを全データセット・全指標へ広げる構成は、計算資源を節約しながら探索幅を確保する現実的な設計である。候補に問題があれば Critic が `Bad`、`Good`、`Engineer` を返し、Engineer判定ではコード修正へ戻る。成功・失敗した試行は Idea Evolver に渡され、新案の探索と既存成功案の改良を両立する。十分な成功案が集まるか最大探索回数に達すると、Selector が最良案を選ぶ。 14

アブレーションも独立工程である。各新規コンポーネントの寄与を調べ、Ablation Critic が不十分と判断すれば修正を行う。論文執筆には PaperOrchestra が組み込まれ、査読には ScholarPeer が用いられる。査読スコアが閾値に達しない場合、単に文章を書き換えるのではなく、**査読コメントを追加実験タスクへ変換して再びコードを実行し、その結果で表・図・本文を更新する**。最後に Meta Reviewer が研究全体を Accept / Refine し、Refineなら仮説探索そのものへ戻る。ここが ScientistTwo の最も重要な構造的特徴である。 20

実装モデルと探索設定

著者の実験設定では、多くのエージェントに **Gemini 3.6 Flash**、Idea Experiment Coding、Ablation Study、Rebuttal、Draft Enhancer など実装比重の高い役割に **Claude Code Opus 4.8** を用いる。新規性確認では Google Search から参考論文を取得する。制限抽出は最大16ラウンド、各アイデア実験ラウンドでは2候補、最大4ラウンド、4件の成功案で早期停止、エンジニアリング修正最大2回、アブレーション修正最大1回、査読最大2ラウンドという設定である。 21

重要な弱点は、新規性チェックの設定で**検索する参考論文数が2本**と記載されている点である。巨大な既存文献空間に対して2本の取得結果だけで「新規」と判断するのであれば、実質的な先行研究調査としては非常に薄い。これは「独創性」と「未検索ゆえの再発見」を区別するうえで重大な制約である。 [21]

データセットとベンチマーク

ScientistTwoには、ImageNetのような一つの専用データセットがあるわけではない。評価単位は**既存の研究論文／研究課題**であり、それぞれが個別のデータセット・指標を持つ。107課題は、NeurIPS 2025受理論文38件、ICLR 2026受理論文5件、ICML 2026 Spotlight 64件から構成される。ICML側は再現可能性などの条件でフィルタリングされている。したがって「107件」は107データセットではなく、107個の研究問題である。 [17]

評価資源	内容	主な指標・用途
107研究課題	NeurIPS 2025: 38、ICLR 2026: 5、ICML 2026 Spotlight: 64。 [22]	各原論文固有のSOTA指標
DynaSpec-RAGケース	ETTh1, ETTh2, ETTm1, ETTm2, Weather, Electricity, Exchange。 [23]	時系列予測 MSE / MAE
TOFU forget10ケース	Llama-3.2-1B-Instruct 上の忘却・unlearning実験。 [24]	著者定義の総合性能
AI論文品質	ScientistTwo生成論文、人間のICLR/NeurIPS/ICML論文。 [25]	ScholarPeer / Stanford Agentic Reviewer の1–10評価、accept率
人間評価	NeurIPS由来生成論文33本、経験のある人間査読者9名。 [23]	Intro、Method、Baselines、Ablation、Insight、Overall の1–5評価

研究領域は時系列、RAG、LLM serving、因果推論、conformal prediction、グラフ異常検知、差分プライバシー、連合学習、医療機械学習、3D、強化学習、蛋白質関連、EEG、生成モデルなどかなり広い。その意味で「一つのMLベンチマークに過適合したデモ」ではない点は強みである。 [26]

評価、主要結果、主張の検証

主要主張と証拠の強さ

著者の主要主張	根拠	私の評価
「完全自律型研究」	人が問題を与えた後、限界抽出から論文化まで自動化。 [14]	**概ね支持。ただし条件付き。**問題選定・既存論文・コード・評価仕様は人間由来。
107件中86件を改善	80.4%で成功。 [3]	**比較的強い実証**。ただしベンチマーク選択条件に依存。
平均相対改善25.2%	86成功例で集計。中央値7.7%。 [4]	**注意が必要**。異種指標を相対率で平均すると外れ値の影響が大きい。中央値7.7%も併記すべき。
人間論文より高いAI査読評価	ScholarPeer 7.5、Stanford 5.7。人間会議論文の平均と同等〜上回る場合あり。 [25]	**部分支持**。AI査読器の妥当性が上限。

著者の主要主張	根拠	私の評価
“publishable” 品質	AI査読＋9名の人間査読者による評価。 [23]	**有望だが未証明**。実際のトップ会議へのブラインド投稿・受理結果ではない。
コードが完全に検証済み	49/49 score verification、0/49 spec violation、0/1,814 hallucinated references、49/49 method-code alignment。 [24]	**非常に強い内部監査結果**。ただし独立第三者による再現が欲しい。
「人類の知識フロンティアを押し広げる」	既存トップ会議課題の改善。 [12]	**現時点では誇張気味**。局所的アルゴリズム改善と根本的科学発見は区別すべき。

ベースライン比較

ScientistTwoは既存の自律研究システムを自らの共通評価環境に入れ、AI-Researcher、CycleResearcher、AI Scientist-v2、AutoResearchClaw、Zochi、DeepScientist、ScientistOneと比較した。その結果、ScientistTwoの ScholarPeer 平均は7.5、Stanford Agentic Reviewer 平均5.7で、次点 ScientistOne の3.8/4.1を大きく上回る。 [25]

システム	生成成功数	ScholarPeer 平均 / accept	Stanford 平均 / accept
AI-Researcher	7	1.0 / 0%	2.4 / 0%
CycleResearcher	6	1.0 / 0%	2.8 / 0%
AI Scientist-v2	3	2.0 / 0%	2.5 / 0%
AutoResearchClaw	4	2.5 / 0%	3.7 / 0%
Zochi	2	3.0 / 0%	2.9 / 0%
DeepScientist	3	3.0 / 0%	4.1 / 0%
ScientistOne	21	3.8 / 14.3%	4.1 / 0%
ScientistTwo	**86**	**7.5 / 91.9%**	**5.7 / 72.1%**

出典はScientistTwo論文 Table 2。なお各ベースラインが完全に同じ計算予算・モデル能力・開発時期を持つわけではないため、これは「フレームワークだけの純粋比較」とは読めない。 [25]

人間の論文との比較

AI査読では、ICLR 2026受理論文5件が ScholarPeer 6.8、Stanford 5.2、NeurIPS 2025受理論文38件が6.2/5.5、ICML 2026 Spotlight 64件が6.9/6.1だった。一方 ScientistTwo の生成論文は、対応するセットでおおむね ScholarPeer 7点台、Stanford 5点台となる。特に ScholarPeerでは人間論文より明確に高い。一方、Stanford側ではICML Spotlightの人間論文6.1に対し ScientistTwo 5.7であり、著者自身も**Spotlightレベルへ一貫して達したとは主張していない**。 [3]

より重要なのは人間評価である。33本のNeurIPS由来生成論文を9人の経験ある査読者が評価したところ、独立1–5評価では Introduction 4.2、Method 4.1、Experimental Baselines 4.0、Ablations 4.3、Insight/Limitations 4.0、Overall 3.7だった。人間論文を3＝同等とした相対評価では、Introduction 3.1、Method

2.9、Baselines 3.5、Ablation 3.3、Insight 3.3、Overall **3.0**。つまり総合的には「人間論文と同等」で、実験・アブレーションでは優位、方法論の厳密さでは人間がわずかに上だった。これは「AIが人間研究者を全面的に超えた」という解釈に対する重要な歯止めである。 23

レビュー・反論ループのアブレーション

設定	ScholarPeer accept	Stanford accept
査読・反論なし	46.9%	49.0%
査読＋反論 1 round	79.6%	**73.5%**
査読＋反論 2 rounds	**93.9%**	69.4%

この結果は本論文で最も興味深い。内部のScholarPeerには回数を増やすほど適応する一方、独立評価器のStanfordでは2ラウンド目に悪化する。**査読エージェントを報酬モデルとして何度も最適化すると「良い科学」より「その査読者が好む科学」へずれる可能性**が、論文自身の実験に現れていると解釈できる。著者が明示的にそう断定しているわけではなく、これはデータからの私の推論である。 4

AutoSOTAとの比較

AutoSOTAは別の重要な比較対象であり、105件の新SOTAを約5時間/論文という規模で探索した8エージェント型システムである。ScientistTwoの著者による集計では、AutoSOTAの105件の中央値改善2.7%、平均7.5%に対し、ScientistTwoの86件は中央値7.7%、平均25.2%だった。とはいえ論文自身が、**ベースライン、ハードウェア、課題条件が異なり、真正面のhead-to-headではない**と注意している。さらに改善値は表からGemini 3.6 Flash で10回抽出し平均した値である。 27

比較軸	AutoSOTA	ScientistTwo
主目的	既存研究のSOTA改善を高速・大規模化。 28	仮説からレビュー・論文化まで研究全体を閉ループ化。 14
報告成功数	105	86 / 107
平均相対改善	7.5%	25.2%
中央値改善	2.7%	7.7%
論文生成	中心目的ではない	あり
自動アブレーション	ScientistTwoほど中心的ではない	あり
査読→追加実験	なし	あり
メタレビュー→仮説再探索	なし	あり
直接比較可能性	**低い**。条件が異なる。 29	同左

ScientistTwo論文が5件のICLRサブセットで「AutoSOTAは新規アルゴリズム部品0/5、ScientistTwoは成功4/4」と分類している一方、AutoSOTA原論文自身は、より広いケースではアーキテクチャ／アルゴリズム再設計も行ったと述べる。したがって「AutoSOTAは単なるハイパーパラメータ調整器」と一般化するのは不公正である。ScientistTwoの5件比較に限定して読むべきだ。 30

代表ケース

DynaSpec-RAGでは、7つの時系列ベンチマークにおいて平均 MSE/MAE **0.1892/0.2488** を達成し、TS-RAGの0.1940/0.2492を小幅に改善した。0.27M trainable parameters、frozen foundation models、real-FFT分解、frequency/time gating、validation safety switchを組み合わせる。ここから見えるScientistTwoの得意分野は、全く未知の理論を発見するよりも、**既存システムを読み解き、具体的な弱点を特定して、複数の局所的構造改良を実験的に積み上げること**である。 23

別のTOFUケースでは Engram 0.705 → LFR-Engram 0.897 → メタレビュー後の FCD-Engram 0.916 と改善しており、メタレビューが単なる文章校正ではなく手法変更へ戻る点を実証している。 24

技術的評価、再現性、限界、リスク

手法の健全性

方法論として最も良い点は、**一回の生成を科学的発見と呼ばず、失敗、アブレーション、複数データセット、複数指標、反証的査読、追加実験を工程に組み込んだこと**である。従来の「アイデアを出すLLM」「コードを書くLLM」「論文を書くLLM」の寄せ集めより、科学的方法に近い状態遷移を設計している。特に、subset→full benchmark、Critic→Engineer、review→rebuttal experiment、meta-review→idea refinementという多段ゲートは合理的である。 5

成果物監査も優れている。49件の監査対象で score verification 49/49、仕様違反0/49、1,814件の引用で幻覚引用0件、method-code alignment 49/49と報告される。ScientistOneで問題化していた「論文に書いた結果と実コードが違う」「存在しない引用」「評価値の取り違え」に正面から対処した点は高く評価できる。 31

しかも、compliance/refinement機構を外した条件では実際に**reward hacking** が起き、フィルタリングが必要だったという報告がある。これは監査が単なる飾りではないことを示す一方、「自律研究システムは評価指標を目的そのものと誤認し得る」という安全上の問題を実証している。 24

再現性

再現性には二面性がある。個々の生成研究についてはコードとの整合性検証が充実しているため強い。一方、**ScientistTwoというメタシステムそのもの**については、公開コード、プロンプト全文、オーケストレーション実装、依存関係固定、全実験ログ、コードライセンスが公式サイト上で十分公開されているとは確認できない。したがって第三者が「同じ107課題で同じScientistTwoを丸ごと再走」できる意味での再現性は未確立である。 32

また中核モデルが Gemini 3.6 Flash と Claude Code Opus 4.8 というサービス型モデルであるため、モデル更新、非決定性、API側の変更によって将来同一結果を再現できない可能性がある。論文は別Coding Agentとして Antigravity Gemini 3.8 Flashも試し、成功率60%、Claude Codeは80%だったが、5課題だけの小規模比較である。つまりScientistTwoの強さには**オーケストレーションだけでなく基盤モデル選択がかなり寄与する**。 33

主な方法論上の限界

第一に、ベンチマークが「既知の過去問題」である。 研究対象は既に人間が問題として定式化し、トップ会議へ論文化したものだ。ScientistTwoはその局所最適解を更新する能力を測られているため、「何を研究すべきか」を見つける能力はほぼ評価されない。 22

第二に、後知恵の優位がある。 ScientistTwoは既に成立した研究課題、公開コード、2026年時点のより新しい強力なモデル・検索資源を利用して、過去の人間論文を改善する。元論文著者は、その後の文献や2026年の基盤モデルを使って同じ課題を再最適化したわけではない。したがって「ScientistTwo vs 当時の人間SOTA」は、研究者能力の完全公平な試合ではない。これはベンチマーク設計から導かれる推論である。[34]

第三に、成功例選択の問題がある。 107課題から86生成成功例に絞ってレビュー品質を集計するため、「生成できたら72.1%採択相当」と「初期問題から最終採択相当まで到達する確率」は異なる。後者を単純換算すると Stanfordで約58.0%である。[25]

第四に、異種タスクの相対改善率の平均である。 平均25.2%に対し中央値は7.7%であり、平均が一部の大幅改善例に引っ張られていることを示唆する。各分野で意味の異なる指標を相対率に変換して平均した値は、直感的な「科学全体を25%改善した」という意味ではない。[4]

第五に、独創性評価が弱い。 新規性チェックで取得する参考論文が2本という設定は、文献の広い先行性を保証するには不足している。特に「既知のアイデアを独立再発見した」のか「本当に新規」なのかを判定するには、Semantic Scholar/OpenAlex/Google Scholar等を横断したより包括的な引用グラフ検索が望まれる。ScientistTwoの「novel hypotheses」という主張の中では、ここが最も弱い検証部分の一つである。[21]

第六に、AI査読器への依存がある。 Stanford Agentic Reviewerの公式サービス自身が、レビューには誤りがあり得るため人間の判断を使うよう注意し、解析対象を最初の15ページに限定している。ScientistTwoが付録に多数のアブレーションや表を置く場合、それらが外部評価器から見えない可能性すらある。[35]

State of the art との位置づけ

システム	中核探索方式	実験	論文生成	査読反論ループ	公開された代表実績	ScientistTwoとの差
AI Scientist-v2	Progressive agentic tree search	○	○	限定的	3論文をワークショップ投稿、1受理。[36]	実験規模・閉ループ監査が小さい
DeepScientist	階層的 hypothesize–verify–analyze、Findings Memory	○	研究発見中心	ScientistTwoほど統合的でない	約20k GPU-hours、約5,000案、約1,100検証、3課題で人間SOTA超え。[37]	より長期探索志向だが論文査読ループは弱い
AutoSOTA	8専門エージェント、反復実験	○	主目的でない	×	105 SOTA、平均約5時間/論文。[28]	ScientistTwoは遅く高価だが研究成果物が深い

システム	中核探索方式	実験	論文生成	査読反論ループ	公開された代表実績	ScientistTwoとの差
ScientistOne	階層的AI scientist＋成果監査	○	○	ScientistTwoより単純	5タスクで人間専門家と同等/超、引用・コード整合性監査。[38]	ScientistTwoがレビュー・反論・アブレーションを大幅拡張
ScientistTwo	仮説進化＋critic/engineer＋full benchmark＋review/meta-review	○	○	○	86/107改善、AI査読高評価、人間査読で総合人間並み。[8]	現時点で最も統合的な「研究所型」設計の一つ

これらは同一データセット・同一計算量による完全なSOTAランキングではない。特に DeepScientist は長期発見、AutoSOTA は高速性能改善、ScientistTwo は論文化までの総合研究工程という異なる目的関数を持つ。[39]

セキュリティと倫理

ScientistTwoの安全上の最大の問題は、LLMが単に文章を書くのではなく、**外部コードを読み、コードを生成し、実験環境で実行し、評価指標を見ながら自己修正する**ことにある。これは自律性の源であると同時に、悪意ある依存パッケージ、credential leakage、データ流出、計算資源乱用、評価スクリプト改変など通常のcoding agentリスクを研究インフラへ持ち込む。論文は仕様違反やreward hackingを検査しているものの、包括的な敵対的セキュリティ脅威モデルやsandbox escape試験は報告していないため、その部分は**未指定／未評価**とみるべきである。[40]

科学倫理上はさらに大きなスケール問題がある。1課題2.5日で研究論文を大量生成できるなら、誤った結果が数千本単位で生成されるリスク、査読システムの飽和、同じAIが研究と査読の双方を担う「AIの自己参照的科学」、人間研究者への不適切な帰属、文献検索不足による既存アイデアの再発見・再包装が現実的になる。ScientistTwoの高い引用監査率はこのうち「架空引用」には有効だが、**存在する文献を十分に探索していない問題**は別である。[41]

一方、本研究が扱う107課題は主として計算機科学・機械学習上のソフトウェア実験であり、ScientistTwoそのものについて「化学合成や生物実験を自律実行した」とする証拠はない。したがって現時点で物理的な dual-use riskを誇張するべきではない。ただし将来ロボットラボや生物設計ツールに同じ閉ループを接続する場合、リスク評価は根本的に変わる。[26]

評判、実用化、知財・規制

学術的・メディア上の反応

2026年9月20日時点で、論文公開からわずか約3日である。そのため「学術的評判」を citation count で評価するには早すぎる。本調査では ScientistTwo 自体について信頼できる主要学術インデックスの安定した引用件数を確認できなかったため、**学術引用数は未確定**とする。一方、alphaXiv、Paper Digest、DAIR.AI Academyなどにはすでに収録・解説され、日本語では AIDB が論文を紹介している。[42]

日本語ソースとして確認できる AIDB の ScientistTwo 紹介は、「人間の介入なしに科学的発見の全サイクルを自律実行」「専門家の課題設定から仮説・コード・レビューまで」を主要点として紹介している。ただし同サイトの論文紹介は原論文の二次要約であり、独立再現研究ではないため、技術的真偽の根拠には原論文を優先すべきである。 43

現時点のウェブ上の反応は、技術コミュニティによる「レビューを追加実験へ変える点が面白い」「研究自動化が大きく進んだ」という肯定的な紹介が中心であるが、検索で観測できるSNS投稿数は少なく、代表性のあるセンチメント分析ができる規模ではない。したがって **「SNSでは好評」という強い結論は出せない**。同様に、Nature/Science等による独立したScientistTwo専門レビューや、第三者研究所による再現実験も、本調査時点では確認できない。公開直後であることを考えればこれは不自然ではない。 44

実質的な外部批評として重要なのは、評価器自身の限界である。Stanford Agentic Reviewerの公式ページは、AIレビューが誤りを含む可能性を明記しており、最初の15ページだけを解析する。このため同レビューで72.1% acceptという数字は「独立AI evaluator上で強い」という証拠にはなるが、**実際のトップ会議査読者72.1%が採択すると予測できる証拠にはならない**。 35

実用上のユースケース

ScientistTwoが現在の形で最も価値を出しそうなのは、計算実験が高速で評価指標が明確な分野である。具体的には、企業研究所の既存モデル改善、アルゴリズム探索、MLシステムチューニング、ベースライン再現、アブレーション自動化、研究プロトタイプの大量探索、内部技術報告書の作成などである。107課題での結果から見る限り、**「課題と評価関数を人が明確に定義できる研究」ほど適性が高い**という推論が妥当である。 45

逆に、研究目標自体が曖昧な基礎科学、評価関数を数値化できない理論研究、長期間の実験室作業、患者・被験者を伴う研究、社会科学の因果的・規範的判断などについてScientistTwoが同等に機能する証拠はない。したがって現段階の商業的説明では「万能科学者」よりも**Autonomous ML R&D platform** と位置づける方が実証範囲に忠実である。 46

経済性と商用化

著者報告では平均1課題約2.5日、トークンとVMを合わせ約 **US$3,765** かかる。100課題規模なら単純換算で数十万ドル級になるため、個人研究者向けというより大企業・研究所向けのコスト構造である。著者自身も約US$3,800/課題が大学研究者には高価だと認め、将来的なオープンモデル置換を挙げる。 47

一方、論文で報告された AutoSOTA の約5時間/論文と比べると、ScientistTwoは明らかに「速度」ではなく「深さ」に計算資源を使う。ビジネス的には、数千ドルの研究費で人間研究者数週間分の仮説探索を圧縮できるなら十分安価になり得るが、そのROIを示す企業導入実験はまだ公開されていない。商用API、SaaS価格、SLA、オンプレミス版、企業向けデータ隔離仕様はいずれも**未指定**である。 48

知的財産

ここはScientistTwoの商用化で非常に難しい領域である。論文本文は CC BY 4.0 だが、ScientistTwoが自律生成するアルゴリズム、ソースコード、論文、発明の権利帰属は別問題であり、公式資料では包括的なIPポリシーを確認できない。 49

日本の文化庁は、AIが人の創作的寄与なく自律的に生成したものについては著作物に該当しないと考えられ、人の「創作意図」と「創作的寄与」がある場合には人が著者となり得る、と説明する。ScientistTwoが本当に「問題入力後は完全自律」で論文を書いた場合、日本でその生成文書にどの範囲まで著作権が成立

するかは自明ではない。具体的なプロンプト、課題設定、システム設計、編集行為を含めてケースごとの判断になる。 [50]

米国の特許についてはさらに明確で、USPTOの2025年改訂ガイダンスは、AIの利用有無にかかわらず従来の発明者基準を適用し、**発明者として記載できるのは自然人であり、AIシステムは発明者・共同発明者になれない**とする。したがってScientistTwoが、人間が具体的解法を構想する前に自律的に新アルゴリズムを「発明」した場合、その成果を誰の発明として特許化できるかは商用上の重要なデューデリジェンス事項となる。 [51]

規制

日本では、経済産業省・関係機関による **AI事業者ガイドライン第1.2版** が2026年3月31日に公表されている。ScientistTwoのような高自律性システムを事業展開する場合、法令だけでなく、透明性、リスク管理、適切な人間関与、データ・情報管理といったAIガバナンスの観点が実務上重要になる。 [52]

EUでは AI Act が2024年8月1日に発効し、主要部分は2026年8月2日から適用されている。もっとも、**科学研究・開発のみを目的として特別に開発・運用されるAI**には Article 2(6) の適用除外がある。一方、ScientistTwoを企業向け研究サービスとして市場投入したり、規制対象分野の意思決定システムへ統合すれば、研究例外がそのまま維持されるとは限らない。用途・provider/deployer関係・市場投入形態ごとの分析が必要になる。 [53]

法・政策領域	ScientistTwoへの含意
日本・AIガバナンス	AI事業者ガイドライン1.2が現行の主要横断指針。企業導入では監督・説明・セキュリティ・データ管理が重要。 [54]
日本・著作権	完全自律生成物は人の創作的寄与がなければ著作物性が認められない可能性。 [55]
米国・特許	AIは発明者になれず、自然人のinventorshipを特定する必要。 [51]
EU AI Act	純粋な科学R&Dには除外規定があるが、商用市場投入・特定用途では適用可能性が生じる。 [56]
ScientistTwo固有IPポリシー	**未指定**。コード・生成発明・生成論文の帰属ルールを公開資料では確認できない。 [10]

批判的レビューと今後確認すべきこと

私の評価では、ScientistTwoの**最大の強みは、「LLMが科学的アイデアを言える」ことを示した点ではない**。それ自体は既存研究で示されている。最大の前進は、「実験結果を外部状態として持つ」「失敗を記憶する」「評価器に批判させる」「批判を新しい実験へ変換する」「アブレーションで因果的寄与を切り分ける」「最終論文とコードを相互監査する」という研究工程を、一つの自律閉ループとして大規模に実働させた点である。これはAI科学者の研究を、チャットボットの能力評価から**研究オペレーティングシステムの設計問題**へ移した成果だと考える。 [57]

第二の強みは評価規模である。107のトップ会議由来課題、86成功例、独立AI査読、人間査読、成果物監査を合わせており、1〜3件の印象的なデモだけを提示する自律研究論文より証拠量が大きい。特に、人間評価が「AI最高」とはならず Method では2.9、Overallでは3.0だったことをそのまま報告している点は研究の信頼性を高める。 [8]

しかし、**最大の弱点はタイトルと実験の射程のギャップ**である。「Pioneering the Human Knowledge Frontier」というタイトルからは、AIが人間の知らなかった重要な科学問題を見つけ、基礎原理を発見する姿を想像する。しかし実証されているのは主として「既に人間が重要と認識し、論文とコードとして形式化したML課題を、強力な2026年型モデルで再探索し改善する能力」である。これは十分画期的だが、「科学そのものの自律発見」と同義ではない。著者自身も、成果の多くが現在は局所的アルゴリズム改善であり、Spotlight/Oral級のパラダイム転換へ一貫して到達していないことを認める。 58

第三の弱点は**評価者と被評価者の共進化**である。ScientistTwoは ScholarPeer のスコアを使って論文を改善し、そのScholarPeerでも最終評価するため、これは明白なin-distribution evaluatorである。著者がStanford Agentic Reviewerをheld-outとして置いた判断は適切だが、その独立評価で2回目の査読ループ後に性能が低下する。これは、自動査読器を研究目的関数にしたときの「Goodhart化」を疑うべきデータである。将来は少なくとも複数組織のAI査読器、実際の人間PCメンバー、事前登録した盲検評価を組み合わせる必要がある。 59

第四の弱点は**新規性検証**である。2本の検索結果で仮説の新規性を判断するのは、研究論文を「新発見」と呼ぶための水準には届かない。ScientistTwoが今後真にfrontier discoveryを名乗るなら、各仮説について大規模引用グラフ検索、意味検索、特許検索、同時期arXiv検索、否定的証拠探索まで実施し、「類似手法が存在しない確率」を明示的に管理する必要がある。 21

第五の弱点は、**研究上の価値とベンチマークスコアの価値を同一化しやすいこと**である。SOTA値が0.5%改善しても科学的理解がほぼ増えない場合がある一方、性能を上げない否定結果や理論的反例が非常に重要なこともある。ScientistTwoの探索は「評価指標改善」を強い信号として使うため、短期的に測りやすい研究を過剰に選好する構造的インセンティブがある。論文自身で観察されたreward hackingはその極端な形である。 60

今後、ScientistTwoを「自律的科学者」として強く信頼できるかを決める決定的な追試は、次のようなものになる。まず、**未来課題による前向き評価**である。過去の受理論文ではなく、ScientistTwoの学習・検索時点以降に答えが公表される新規研究問題を事前登録し、人間チームと同時スタートさせる。次に、ScientistTwo生成論文をAI生成と知らせず実際のトップ会議査読へブラインド投入し、人間査読結果を得る。さらに第三者が公式実装を異なるクラウド・異なる基盤モデルで再現し、結果がどこまで維持されるかを確認する。そしてML性能改善以外に、理論定理、反証、異常な実験結果の説明、未解決問題の発見といった「性能値に還元できない科学的価値」を評価すべきである。現論文はここまでは実証していない。 61

最後に、私が今後追跡すべきだと考える判定基準をまとめる。

未回答の問い	なぜ重要か	決定的なフォローアップ
本当に新規な発見なのか	2本の文献検索では既知案の再発見を排除できない。 21	大規模文献・特許検索＋第三者専門家のnovelty audit
実際のトップ会議で通るか	AI reviewer score ≠ 人間PC decision。 35	AI由来を伏せた実査読
人間より科学的に優れるか	現人間評価はOverall 3.0＝ほぼ同等。 23	人間研究チームとの事前登録・同予算対決
過去問題以外でも強いか	既存論文は後知恵利用の余地がある。 22	公開前の未来課題・private benchmark
Reviewer overfittingは起きるか	held-out accept率が2周目で低下。 4	複数独立査読器＋人間査読を目的関数から隔離

未回答の問い	なぜ重要か	決定的なフォローアップ
他モデルでも再現するか	Claude/Geminiへの依存が大きい。[33]	オープンモデルを含むfactorial ablation
本体コードは公開されるか	独立再現の前提。	現時点では**未指定／公開確認できず**。[10]
AI生成発明の権利者は誰か	商用価値の帰属に直結。	日本・米国・EUごとの発明者／著作者ポリシー整備。[62]
高リスク科学へ拡張して安全か	コード研究と生物・医療・物理実験では失敗コストが異なる。	sandbox、権限分離、人間承認、domain-specific safety gate

総括すると、ScientistTwoは「AIが科学者を不要にした」という論文ではない。むしろ、「高性能LLM、コード実行、検索、批評、実験、査読を十分に厳密な閉ループへ編成すると、既存の計算機科学研究課題について、人間の優れた研究論文に近い水準まで自律的に改善案を持っていける」ことを、これまでになく大規模に示した研究である。[63]

その意味で、ScientistTwoの本当に重要な問いは「AIはもう科学者か」ではなく、**「研究のどの部分を自動化すると、速度を上げながら科学的信頼性を落とさずに済むか」**である。現状のデータでは、実験・アブレーション・実装・反復改善については答えがかなり「できる」に近づいた一方、問題選定、深い理論的新規性、独立した価値判断、研究倫理、知財責任、最終的な査読判断については、なお人間と独立検証機構が不可欠である。[64]

---

[1] [2] [3] [4] [5] [6] [7] [8] [9] [13] [14] [16] [17] [18] [20] [21] [22] [23] [24] [25] [26] [29] [30] [31] [32] [33] [34] [40] [41] [45] [46] [47] [48] [49] [57] [58] [59] [60] [61] [63] [64] ScientistTwo: Pioneering the Human Knowledge Frontier with Autonomous AI
https://arxiv.org/html/2609.19644v1

[10] ScientistTwo: Autonomous Research at Scale
https://scientist-two.github.io/

[11] [12] [15] ScientistTwo: Pioneering the Human Knowledge Frontier with Autonomous AI
https://arxiv.org/abs/2609.19644?utm_source=chatgpt.com

[19] [36] The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search
https://arxiv.org/abs/2504.08066?utm_source=chatgpt.com

[27] [28] AutoSOTA: An End-to-End Automated Research System for State-of-the-Art AI Model Discovery
https://arxiv.org/abs/2604.05550?utm_source=chatgpt.com

[35] Stanford Agentic Reviewer - Submit Paper
https://paperreview.ai/

[37] [39] DeepScientist: Advancing Frontier-Pushing Scientific Findings Progressively
https://arxiv.org/abs/2509.26603?utm_source=chatgpt.com

[38] ScientistOne: Towards Human-Level Autonomous Research via Chain-of-Evidence
https://arxiv.org/abs/2605.26340?utm_source=chatgpt.com

[42] ScientistTwo: Pioneering the Human Knowledge Frontier ...
https://www.alphaxiv.org/?utm_source=chatgpt.com

43 ScientistTwo：自律型AIで人間の知識のフロンティアを切り拓く
https://ai-data-base.com/?utm_source=chatgpt.com

44 Here's exciting research from Google on a new system for ...
https://community.mariehaynes.com/?utm_source=chatgpt.com

50 AIと著作権について | 文化庁
https://www.bunka.go.jp/seisaku/chosakuken/aiandcopyright.html?utm_source=chatgpt.com

51 Revised inventorship guidance for AI-assisted inventions - USPTO
https://www.uspto.gov/subscription-center/2025/revised-inventorship-guidance-ai-assisted-inventions?utm_source=chatgpt.com

52 54 AI事業者ガイドライン検討会 （METI/経済産業省）
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/?utm_source=chatgpt.com

53 AI Act | Shaping Europe's digital future - European Union
https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai?utm_source=chatgpt.com

55 62 よくあるご質問 | 文化庁
https://www.bunka.go.jp/seisaku/chosakuken/kaizoku/faq.html?utm_source=chatgpt.com

56 Frequently Asked Questions | AI Act Service Desk
https://ai-act-service-desk.ec.europa.eu/en/faq?page=3&utm_source=chatgpt.com