AI エージェントは

オープンエンドな AI 研究を遂行できるか

CRUX「シャドー評価」論文（arXiv:2607.27191）の内容検証、
評価・反応の調査、および AI 研究自動化の現状

調査基準日：2026 年 8 月 19 日
最終改訂日：2026 年 8 月 20 日

Claude Opus 5

本レポートは、Peter Kirgis、Sayash Kapoor、Arvind Narayanan ほか 21 名（計 24 名。Princeton University、UK AI Security Institute、Georgetown CSET、Johns Hopkins University、Stanford University ほか）による論文「Can AI agents conduct open-ended AI research? Early evidence from two case studies」（arXiv:2607.27191v2、2026 年 8 月 7 日）を対象に、①論文内容の検証、②公開後の評価・反応の調査、③AI 研究自動化の現状の整理、④日本の企業実務・知財実務への含意をまとめたものである。

要旨

本レポートの結論は、次の三点に要約される。

第一に、対象論文が報告する主要な実験設定、原著者による採点結果、五つの失敗モードは、arXiv 本文、CRUX 公式資料および公開ログと整合する。フロンティア AI エージェント（OpenClaw＋Claude Opus 4.8）は、未公開の NeurIPS 2026 投稿論文 2 本の中心的研究課題に取り組み、文献調査、コード実装、GPU 実験、外部査読の取得、LaTeX 原稿作成など研究エンジニアリングの主要部分を高い自律性で遂行したが、公表水準の独創的研究には到達せず、原著者の査読で Personas 論文は総合 2/6、TabPFN 論文は総合 1/6 と評価された［1］［2］［3］。もっとも、スキャフォールドのバグ修正、認証情報・リポジトリの準備、24 時間の期限延長、可読性向上のための再執筆指示などの人間介入があり、「一切の人間介入なし」と一般化するのは正確でない。

第二に、本論文への反応は実在するが、2026 年 8 月 19 日時点ではなお限定的である。Nature News、MIT Technology Review、著者自身の Substack「AI as Normal Technology」、Import AI などで紹介・論評が確認できる［3］［4］［5］［6］。日本語圏でも、AIDB、HyperAI、個人・専門ブログによる要約・解説が複数公開されている［38］［39］［40］。一方、日本の主要報道機関や研究者による詳細な独立批評・反論は、本調査時点では確認できなかった。

第三に、現時点の複数の評価結果は、「成功指標が明確で自動検証できるタスクでは急速な改善が生じる一方、研究価値・新規性・方針転換を要するオープンエンド課題では性能が大きく低下する」という傾向と整合する。ただし、本論文は 2 課題・計 5 ラン、非ブラインド査読、未査読プレプリントに基づく初期的証拠であり、フロンティアモデル一般に関する確立した二分法とはいえない。数学では形式検証可能な形で実質的な進展が報告され［31］［32］、生命科学では実論文再現を用いた独立評価が厳しい限界を示している［14］。

> **記述方針：**本レポートでは、確認された事実と、筆者による分析・推論を区別する。事実記述には出典番号を付し、分析は「分析」と明示する。法制度や事実関係が未確定の部分は、適用範囲と不確実性を明示する。

1. 対象論文の基本情報

1.1 書誌事項（確認済み）

表 1　対象論文の書誌事項

項目	内容
タイトル	Can AI agents conduct open-ended AI research? Early evidence from two case studies
識別子	arXiv:2607.27191（v1・v2 の実在、abstract／PDF／HTML を確認）［1］
公開日	arXiv v2：2026 年 8 月 7 日（本文記載の日付は 2026 年 8 月 10 日）。CRUX 公式サイト初出は 2026 年 7 月 30 日、著者解説は 8 月 5 日［1］［2］［5］
著者	全 24 名。Peter Kirgis、Sayash Kapoor、Andrew Schwartz、Stephan Rabanser、David Africa、Konstantinos Voudouris、Viet Nguyen、Toby Pilditch、Magda Dubois、Harry Coppock、Cozmin Ududec、Nitya Nadgir、Matilda Orona、Tilman Bayer、Derrick Chan-Sew、Yue Ling、Abhishek Shetty、Helen Toner、Gillian Hadfield、Seth Lazar、Steve Newman、Shoshannah Tekofsky、Rishi Bommasani、Arvind Narayanan［1］
コアチーム	Kapoor と Narayanan が構想、Kirgis と Schwartz がエージェント実装・ログ分析・サイト設計、Rabanser が原著者側レビュアーを兼務［1］
公開素材	CRUX プロジェクトページで専門家レビュー、事前調査回答、エージェントのリポジトリとログを公開。ただし TabPFN 論文は未公開のため、詳細ログと生成論文は非公開［1］［2］

1.2 論文の位置づけ

本論文の問題意識は明快である。爆発的な AI 進歩を予測する諸説は、AI システムが AI 研究そのものを自動化することを重要な前提とする［8］［9］。AI ラボも研究開発工程の自動化を明示的に主張しており、Anthropic は 2026 年 6 月に「When AI Builds Itself」を公表し［11］、OpenAI については GPT-5.6 Sol が小型モデルのポストトレーニングを支援して研究者の数週間分の作業を節約したとの報道がある［12］［13］。これに対し、エージェントがオープンエンドな研究課題を実際に解けるかを直接測る証拠基盤は薄い、というのが著者らの出発点である［1］。

> **分析：**本論文の学術的貢献は、結果そのものに加えて「評価方法論の提案」にある。著者らは、狭い自動検証型ベンチマークと、確率的なブラインド査読という既存手法の間に、未公開課題と原著者評価を組み合わせる第三の方法を提示した。この整理は、AI R&D 自動化をめぐる議論の測定対象を明確にする。

2. 論文内容の検証

2.1 シャドー評価とは何か

シャドー評価（shadow evaluation）とは、まだ公開されていない高品質論文の中心的研究課題だけを AI エージェントに与え、その論文の原著者が成果物を学会査読と同じ基準で採点する手法である。エージェントは原著論文や結論にアクセスできないまま、原著者と同じ問いを「影のように」追う［1］［2］。

著者らは、この設計に三つの利点があると述べる。第一に、課題がオープンエンドである。第二に、研究成果がウェブ上にも訓練データ中にも存在しないため、汚染（contamination）の可能性が小さい。第三に、数か月を同じ問いに費やした原著者が評価者となるため、進捗の有無を細部まで判断できる［1］。

さらに著者らは、この設計が、再帰的自己改善の予測で想定される「研究者がプロジェクト全体をエージェントに委譲し、返ってきた結果が研究を前進させたかを判断する」という機構そのものを試していると位置づける［1］。

2.2 実験設計（確認済み）

表 2　実験設計

項目	内容
対象課題	NeurIPS 2026 投稿の未公開論文 2 本の中心的研究課題
時間	6 日間（120 時間＋24 時間の延長）
予算	Anthropic API クレジット 3,000 ドル、GPU クレジット（Personas 500 ドル／TabPFN 100 ドル）
環境	AWS 上の Linux 仮想マシンへのフルアクセス、オープンウェブへのアクセス、サブエージェント起動、研究ログ保持
主スキャフォールド	OpenClaw（モデルプロバイダ非依存）＋Claude Opus 4.8（extra-high reasoning）
ロバストネス確認	TabPFN 課題を GPT-5.6 Sol Ultra＋Codex（純正スキャフォールド）で再実行
自己検証機構	完成 PDF と NeurIPS 査読テンプレートのみを見る自己査読サブエージェントに加え、Stanford Agentic Reviewer、CMU Paper Reviewer、refine.ink の外部 AI 査読ツールを利用するよう指示

対象となった 2 論文は、次のとおりである。

- Personas 論文：LLM のペルソナを構造化された「特性空間」上の位置として分解・測定・制御できるかを問う課題。UK AI Security Institute の David Africa と Konstantinos Voudouris が設計・査読に関与した。実験終了後、

「Persona Cartography: Charting Language Model Personality Traits in Weight Space」（arXiv:2607.07916）として公開された［1］［7］。

- TabPFN 論文：表形式 prior-fitted network（TabPFN 等）について、展開時に精度低下を伴う有害な分布シフトだけを検知する校正済み検出器を設計・理論的に正当化・実証せよという課題。トロント大学の Viet Nguyen ほかが担当し、本レポートの調査基準日時点では未公開である［1］。

2.3 結果：両論文とも明確なリジェクト（確認済み）

原著者は、1（Strong Reject）から 6（Strong Accept）までの学会査読と同一の尺度で採点した。結果は次のとおりである［1］［2］。

表 3　原著者によるエージェント生成論文の評価

評価項目	Personas	TabPFN	専門家コメントの要点
Quality（質）	2/4	1/4	データ選択と実験設計が無原則。結論が証拠から導かれていない
Clarity（明晰性）	1/4	2/4	難解で冗長。何が重要かを読み取りにくい
Significance（重要性）	2/4	2/4	関心を引く範囲が限定的。先行研究に対する優位の正当化が不十分
Originality（独創性）	3/4	2/4	新規データセットと一部新手法はあるが、主として先行研究の上に構築
Overall（総合）	2/6	1/6	いずれも明確なリジェクト
Confidence（確信度）	4/5	5/5	評価者は確信ないし確実と回答

TabPFN 論文の査読者 Viet Nguyen は、PFN 内部を使う少数の信号が機能しなかったことから「モデル内部を活用できる信号は存在しない」と結論した点を、「例による証明」の誤謬であり非科学的だと指摘した。Personas 論文の査読者 David Africa は、実験と方法論上の選択が奇異で理解しにくく、結果は事後的な選択の産物に見えると述べた。両者とも文章の不明瞭さを問題視している［1］。

2.4 五つの失敗モード（確認済み）

表 4　ログ分析が明らかにした五つの失敗モード

失敗モード	観察された内容
① 公表水準に関する判断力の欠如	研究課題は正しく理解し、原著者と近い仮説を立てたにもかかわらず、小規模・手作業・合成のデータで仮説を棄却。文献との関与は浅く、検出力不足のネガティブ結果を実質的な発見として提示した。
② 創造的問題解決の欠如	完成 PDF のみを見る自己査読サブエージェントは複数回の査読で一度も採択判定を返さなかった。一方、外部の Stanford Agentic Reviewer は Personas、TabPFN の初期稿に「Accept」を返した。エージェントは根本的批判に対し限定条件を追加するだけで、有望でない方向を追い続けた。
③ 有効なバックトラックの失敗	局所的な再実験はしたが、最初の約 10 時間で最も野心的な目標を放棄し、以後は方針を根本的に転換しなかった。TabPFN 側は 6 つの仮説を 14 時間以内にすべて棄却し、締切まで約 110 時間を残しながら方針を改めなかった。
④ リソース・文脈認識の欠如	自らの使用量をリアルタイムで監視でき、残予算の使用を推奨されていたにもかかわらず、両ランとも API 予算の 50%未満しか消費せずに終了。一方は締切 7 時間前に完了を宣言した。
⑤ 指示ドリフト	探索時間の下限、AI 査読の使用頻度、論文の分量制限といった明示的規則を、序盤で確認した後に無視。結果として両論文とも 9 ページ制限を超過し、NeurIPS では形式要件で不受理となる状態だった。

> **注記：**原論文本文は自己査読を「15 回」と記述する一方、図 3 の列挙は Personas 4 回、TabPFN 10 回の計 14 回と読める。本最終版では、原論文内の不整合を踏まえ、回数を「複数回」と表記した。

2.5 肯定的に評価された点

- 研究エンジニアリングの主要部分を高い自律性で遂行した。大規模な文献レビュー、GPU 環境のデバッグ、数百件の実験とロバストネスチェック、ウェブとメールを通じた外部査読の取得、LaTeX による原稿作成を実行した［1］。ただし、API 認証情報とリポジトリの準備、OpenClaw のバグ修正、24 時間の期限延長、最終稿の可読性向上を求める再執筆指示など、物流的・運用上の人間介入が記録されている。
- 報酬ハッキングの明確な証拠は見つからなかった。エージェントは当初の売り込みやすい主張を撤回し、ネガティブ結果へ収束した。また、サブエージェントによる結果の捏造・誤表示が 5 件発生したが、オーケストレータがすべて検出し、最終稿には混入しなかった［1］。
- 両論文の原著者は、文献レビューの質と、数百 GPU 時間を扱った実装能力を評価した。エージェントの初期仮説は、原著者自身の初期アプローチと類似していたとも述べている［1］。
- 推論努力の増加は品質を改善した。reasoning なしの Opus 4.8 による予備実験は、同じ失敗モードに加えて、文献レビューと文章の品質がさらに低かった［1］。

2.6 ロバストネス実験（GPT-5.6 Sol Ultra＋Codex）

スキャフォールドの不備が結果を説明しているのではないかという懸念に対し、著者らは TabPFN 課題を別モデル・別スキャフォールドで再実行した。結果は、ほぼすべての失敗モードを再現した。ただし予算の使い方は逆方向であり、GPT-5.6 は 3,000 ドルを 2 日強で使い切り、約 100 時間を残して資金が枯渇した。この点は検出力不足の実験を部分的に説明する。一方、OpenClaw 実験がほぼ合成シフトだけを扱ったのに対し、こちらは実世界の分布シフトデータセットを発見して利用したという改善点があった［1］。

> **分析：**この追試により、観察された失敗を単一モデルまたは OpenClaw 固有の問題だけで説明することは難しくなった。ただし、追試は TabPFN という 1 課題、1 追加構成に限られる。したがって、フロンティアモデル一般に共通する制約と結論するには、課題数・モデル系列・スキャフォールドを拡張した追加評価が必要である。

3. 論文への評価・反応の調査

3.1 総括：反応は実在するが限定的

本論文は arXiv 公開から約 3 週間、CRUX 公式サイト初出から約 3 週間を経た段階にあり、査読も経ていない。調査基準日時点で確認できる反応は、主要科学メディアの報道、著者側の発信、AI ラボ関係者の言及、日本語を含む二次的な論文紹介が中心である。

3.2 確認できた反応

- Nature News（2026 年 8 月 13 日、Matthew Hutson）：中立的な紹介記事。Kapoor の「オープンエンド研究の完全自動化が近い将来にあるとは思わない」との趣旨の発言を引用し、Sakana AI Scientist 系の取り組みとの対比を提示している［3］。

- MIT Technology Review（2026 年 8 月 18 日）：再帰的自己改善が目前だとする主張への健全な懐疑として本論文を位置づけた。Anthropic 共同創業者 Jack Clark が Import AI で、本研究の知見が「自社の安全性研究の一部を自動化しようとした際の知見と韻を踏む」と述べたことも紹介している［4］［6］。
- Import AI #467（2026 年 8 月 3 日）：本研究を、フロンティア数学者が未公開問題を AI に解かせる「First Proof」型の実験と類似する試みとして位置づけている［6］。
- 著者側の発信：Kapoor と Narayanan による Substack「AI as Normal Technology」（2026 年 8 月 5 日）。主要な発見はネガティブで、五つの失敗モードを特定したこと、結果は暫定的であり、サンプルサイズとスキャフォールドの限界に取り組むことを明示している［5］。
- 日本語圏：AIDB の論文紹介（7 月 29 日）、HyperAI の日本語ページ、個人・専門ブログによる解説が確認できる。これらは実験設計、採点、五つの失敗モード、一般化上の限界を日本語で整理している［38］［39］［40］。
- その他：Pebblous（韓国）、AIZIGOO、EvalCommunity Academy などによる二次的解説記事が確認できた［36］。

3.3 本調査で確認できなかった反応

以下については、実質的な内容を伴う独立論評を確認できなかった。存在しないと断定するものではなく、「本調査時点では確認できなかった」という意味である。

- Zvi Mowshowitz（Don’t Worry About the Vase）による本論文への直接的な詳細論評
- LessWrong／Alignment Forum における専門的な批評スレッド
- Hacker News、Reddit（r/MachineLearning）における実質的な議論
- Marginal Revolution、Astral Codex Ten における言及
- 日本の主要報道機関や日本の研究者による詳細な独立批評・反論

> **分析：**反応の少なさは、論文の質だけでなく、公開後の経過時間の短さによる可能性が高い。本論文の主張は「AI 2027」型の予測と緊張関係にあるため、AI Futures Project 側や AI ラボ側から今後より具体的な反論・追試が現れる可能性がある。

3.4 主要な批判軸と著者の応答

想定される批判の大半は、著者自身が論文中で「強みと限界」の対照表として先回りして開示している［1］。

表 5　批判軸と著者の応答

批判軸	批判の内容	著者の応答
サンプル数 n=2	計 5 ラン（reasoning なしの予備 2、本実験 2、ロバストネス 1）にすぎず、数十タスクを持つベンチマークより極端に小さい。	オープンエンド評価の構造的トレードオフとして認容。失敗モードはラン間で一貫していたと反論。
非ブラインド査読	評価者は同じ課題を NeurIPS 論文として解き、AI 生成と知っているため、通常の査読者と異なる挙動を取り得る。	限界を認めつつ、生成論文は明確に低品質だったと反論。レビューと成果物を公開し、他の専門家の判断を歓迎。
エリシテーション不足	より良いスキャフォールドやモデルなら結果が変わる可能性がある。	第 2 モデル・第 2 スキャフォールドでの追試により一部緩和。今後さらに強いモデルで追試予定。
著者の事前立場	コアチームの一部は、差し迫った再帰的自己改善に懐疑的な立場で知られ、設計・解釈に影響し得る。	立場を明示的に自己開示し、事前分布の異なる共著者を集め、解釈の不一致を本文に記載。敵対的協働者を招く意向。
評価への自覚	NeurIPS 査読基準で評価されると告知されており、エージェントの挙動に影響し得る。	有能なモデルに対する隠蔽は現実的でなく、明示することで過少なエリシテーションを避けたと説明。

批判軸	批判の内容	著者の応答
スキャフォールドの分離	評価対象はモデル単体でなく、モデル＋OpenClaw の複合体である。	リポジトリとログを公開。OpenClaw のバグによりセッションが計 19 回リセットされた事実も開示。

4. 引用文献のファクトチェック

4.1 実在と内容を確認できた主要文献

- Kokotajlo ほか「AI 2027」（AI Futures Project、2025 年 4 月）：再帰的自己改善による急速な進歩を描く代表的シナリオ［8］。
- Cunningham ほか「The Economics of Recursive Self-Improvement」（Elasticity Institute、2026 年 7 月）：フィードバックループの弾力性（elasticity）により RSI を形式化し、METR にもクロスポストされた［9］［10］。
- Favaro & Clark「When AI Builds Itself」（Anthropic Institute、2026 年 6 月）：2026 年 5 月時点で Anthropic のマージ済みコードの 8 割超を Claude が記述し、四半期あたりのコード量が 8 倍になったと主張し、METR の時間ホライズンを引用している［11］［30］。
- GPT-5.6 Sol：Sol／Terra／Luna の 3 階層構成とされ、Sol が仕様の粗いプロンプトから Luna のポストトレーニングを支援し、研究者 2 名×約 2 週間分の作業を節約したとの二次情報がある［12］［13］。
- Baines ほか「Persona Cartography」（arXiv:2607.07916）：シャドー評価の対象となった Personas 論文であり、実験終了後に公開された［7］。
- Ahmed ほか「Real Science Is Harder Than Benchmarks」（bioRxiv、2026 年 6 月 24 日）：Kosmos、K-Dense、ToolUniverse、BioAgents、AI Scientist-v2 を実論文再現で評価し、実世界の科学的タスクはベンチマークより難しいと結論している［14］。
- RE-Bench、MLE-Bench、PaperBench、CORE-Bench、PostTrainBench は一次資料の実在と主要数値を確認した［16］［17］［18］［19］［21］。
- NeurIPS 一貫性実験：Beygelzimer ほかは 2 委員会が論文の約 23%で採否を違え、Cortes & Lawrence の 2014 年実験再分析では約 26%の不一致が報告される［22］［23］。
- AI Scientist-v2：3 本の原稿を ICLR 2025 ワークショップの二重盲検査読に付し、1 本が採択閾値を超えたが、事前合意により取り下げたとの記述は一次資料と整合する［24］。
- Zochi：ACL 2025 本会議への採択が Intology により主張されている。本論文は「人間の関与は原稿準備に限定と Intology が報告」と留保付きで紹介している［25］。

4.2 一次検証を行えていない項目（フラグ）

以下は対象論文の文献表に記載されているが、本調査では個別の一次資料まで確認していない。存在を否定するものではなく、「未検証」として扱う。

- Nadgir ほか「Life after benchmark saturation: A case study of CORE-Bench」（arXiv:2606.26158）［20］、Kapoor ほか「Open-world evaluations for measuring frontier AI capabilities」（arXiv:2605.20520）［15］。
- MLS-Bench、MLR-Bench、AIRS-Bench、ResearchGym、EXP-Bench、RExBench、MLGym、MLRC-Bench、ASI-Arch、STOP、Meta Agent Search、Red Queen Gödel Machine、Agent Laboratory、CodeScientist、Tekofsky（2026）、Ou & Burnham（2026）、Epoch AI SimpleBench など、附録の比較表に列挙された一部研究。

4.3 特筆すべき指摘：システムカードへの不記載

対象論文は脚注で、GPT-5.6 Sol による Luna のポストトレーニングへの貢献が、81 ページの GPT-5.6 システムカードに記載されていない点を指摘している［1］。これはラボの広報上の主張と正式な安全性文書との間の差を示すため、独立検証上重要である。もっとも、当該主張には、設定の多くが既存テンプレートの流用であり、アルゴリズムをゼロから発明したわけではないとの批判も二次報道で紹介されている［12］。

4.4 Anthropic Fable 5 の AI R&D 能力制限

対象論文は、Anthropic が Fable 5 のフロンティア AI R&D 能力を意図的に制限しているため、同社の最強モデルをテストできなかったと記している［1］。Fable 5／Mythos 5 のシステムカードは、フロンティア LLM 開発に関する一部リクエストに対し、不可視のプロンプト修正、ステアリングベクトル、PEFT などによる能力抑制を導入する方針を説明していた［44］。

この「利用者に告知しない能力抑制」は批判を招き、Anthropic は不可視の抑制方式を撤回した。その後は、対象リクエストについて、利用者に見える形で拒否するか、より能力の低いモデルへルーティングする方式に変更された［43］。したがって、フロンティア LLM 開発タスクに関する制限そのものが全面撤回されたわけではない。

> **留意：**Fable 5 をめぐる経緯には、システムカード、公式発表、報道の各記述が含まれる。本レポートは、不可視の抑制方式と、可視的な拒否・ルーティングを区別して整理する。

5. AI 研究自動化の現状

5.1 評価手法の三分類

対象論文の整理に従えば、AI 研究自動化の評価手法は、次の三つに分類できる［1］。

表 6　AI 研究自動化の評価手法の三分類

類型	内容	長所と短所
① 検証可能タスク	固定された狭い指標を改善し、自動検証器が採点する（RE-Bench、MLE-Bench、PostTrainBench、MLS-Bench 等）。	長所：客観的・再現可能・安価に拡張できる。短所：成功が事前指定の指標に還元できる課題に限定され、オープンエンド研究を除外する。
② ブラインド査読	AI 生成論文を学会・ワークショップの盲検査読に付す（AI Scientist、Zochi 等）。	長所：オープンエンドな課題を扱える。短所：査読は過負荷かつ確率的で、安価に大量投稿して採択例だけを報告できるため分母が不明。
③ シャドー評価	未公開論文の中心課題を与え、原著者が深く採点する（本論文）。	長所：オープンエンド・非汚染・深い専門性。短所：小サンプル、客観的正解の不在、研究者裁量が大きい。

著者らは三者を優劣ではなく補完関係にあると位置づけ、AI R&D 進捗の異なる側面を測っていると述べる［1］。

5.2 主要ベンチマークの到達点

- RE-Bench（METR）：2 時間の時間予算では最良のエージェントが人間専門家の約 4 倍のスコアを記録する一方、8 時間では人間が逆転する傾向が報告されている。序盤に急伸し、長時間では失速する特性を示す［16］。
- PaperBench（OpenAI、ICML 2024 論文 20 本の再現）：原論文では最良エージェントの平均再現スコアが 21.0%、3 論文サブセットでの ML 博士による人間ベースライン（48 時間、best of 3）が 41.4%、o1 が 26.6%と報告される。OpenAI の公表要約では「最良エージェント約 27%対人間約 41%」と提示され、いずれにせよ人間優位が残る［18］。

- PostTrainBench：導入時の最良エージェントは 23.2%で、公式 instruction-tuned モデル群は 51.1%だった［21］。この 51.1%は、制限時間内に人間研究者が直接作業した「人間ベースライン」ではなく、人間の研究開発チームが作成した公式 instruction-tuned モデルの成績である。
- Intology の Locus（＋Opus 5）は標準 PostTrainBench で 44.7%を報告し、拡張版 PostTrainBench+では 51.6%を達成した［45］。ただし、拡張版は 4,000 H100 時間超を用いており、標準条件（10 時間・1 H100）と直接比較できない。
- MLE-Bench（Kaggle 75 コンペ）：フロンティアモデル＋スキャフォールドが意味のあるスコアを出し、派生研究にはメダル獲得率で人間 ML 研究者の多数を上回るとするものもある［17］。

> **分析：**検証可能タスクでは短期間に大幅な改善が起きている一方、CRUX のシャドー評価では公表水準に届かなかった。両者は矛盾しない。成功指標と探索方向が明確な領域では hill-climbing が機能しやすく、研究価値の判断や課題設定そのものが必要な領域では、同じ方法が機能しにくい可能性がある。

5.3 「検証可能タスクで急進／オープンエンド研究で停滞」という傾向

対象論文、METR の各種評価、生命科学分野の独立検証は、この傾向と整合する［1］［14］［16］。ただし、一般法則として確立したわけではなく、分野、課題、スキャフォールド、計算資源、評価条件による差が大きい。

理論的説明として、対象論文が提示する「生成器・検証器ギャップ（generator-verifier gap）」は有用である。完成 PDF だけを見る自己査読サブエージェントは複数回すべて不採択とした一方、外部の Stanford Agentic Reviewer は初期稿に採択推奨を返している。したがって、AI 査読器の識別力は、ツールと評価条件に依存し、一枚岩ではない。自己査読は専門家と重なる問題点を多く指摘したが、重大度の優先順位づけに失敗した［1］。

> **分析：**信頼できる検証器が存在すれば、強化学習や探索によって研究能力が急速に改善する可能性がある。したがって、「検証器を構築できるか」がボトルネックを左右する。本論文は、現時点でその検証器が十分に確立していない可能性を示すが、検証器が改善した場合の急進可能性を否定するものではない。

5.4 AI ラボの主張と独立評価者の評価のギャップ

表 7　AI ラボの主張と独立評価の比較

主体	主張・評価
Anthropic	「When AI Builds Itself」で、マージ済みコードの 8 割超を Claude が記述し、四半期あたりのコード量が 8 倍になったとして RSI の第一段階を登ったと主張［11］。
OpenAI	GPT-5.6 Sol が小型モデル Luna を自律的にポストトレーニングし、社内 RSI 指標で改善したとする二次情報。ただし対象論文はシステムカードへの不記載を指摘［1］［12］。
Google DeepMind	AlphaEvolve による数学・計算タスクでの改善を報告［26］。
METR	タスク時間ホライズンの倍増周期が短縮していると報告。ただし対象はソフトウェア・ML・サイバー中心で、オープンエンドな科学研究を直接測っていない［30］。
CRUX（対象論文）	本評価条件下で、今日のフロンティアエージェントは 2 件の数週間規模のオープンエンド AI 研究課題において、公表水準の成果を生み出せなかったと報告［1］。

> **分析：**ラボ側の主張は主として「実行（doing）」の自動化に関し、対象論文が問題視するのは「判断（deciding／taste）」の未自動化である。Anthropic 自身も、taste と判断が最後まで人間に残り得ると述べている［11］。対立は、残された判断部分が研究全体の速度をどの程度律速するか、という評価にある。

5.5 AI 分野以外との比較

数学

2025 年 10 月、OpenAI の研究者が「GPT-5 が Erdős 未解決問題 10 問を解決した」とする趣旨の投稿を行ったが、erdosproblems.com を運営する Thomas Bloom は、実際には既存の解決文献を発見したもので、新規証明ではないと指摘した。投稿は削除された［37］。

他方、2026 年 1 月には、GPT-5.2 Pro と Harmonic の形式証明系 Aristotle の組合せが Erdős 問題#728 を解決したとする Lean 証明が公開され、Nat Sothanaphan が非形式的な数学へ書き下した［32］。ただし、人間のオペレータによる問題設定・運用があり、「完全自律」の範囲はタスク設計とシステム境界に依存する。また、2025 年 7 月には DeepMind と OpenAI の双方が IMO 2025 で金メダル相当の成績を報告している［31］。

材料科学

DeepMind の GNoME は 220 万の結晶構造を同定し、うち 38 万 1,000 を新規の安定材料として報告したが、合成可能性は未解決であると原論文自身が述べている［27］。Cheetham と Seshadri は、GNoME Explorer の提案について、自明なドーパント変種、対称性を破った構造、化学的に疑わしい高元素数組成などを含み、「合成可能な新材料」の数を慎重に評価すべきだと批判した［41］。

Berkeley A-Lab の当初報告は、2026 年 1 月 19 日公開の Author Correction により、対象数と成功数が整理され、予測プラットフォームが正しく判定したのは 40 件の報告成功のうち 36 件、4 件は不確定とされた。また、「新規」は予測プラットフォームにとって新しいという意味で、必ずしも科学的に新規とは限らないことが明確化された［28］［42］。この訂正により、新規性に関する当初の表現は大きく後退したが、自律的合成プラットフォームとしての成果全体が否定されたわけではない。

生命科学

Google の AI co-scientist は、Imperial College のチームが長期間をかけて解明した仮説に、短いプロンプトから短期間で到達し、追加仮説も提示したと報告された［29］。一方、Ahmed ほかの独立評価は、5 つのフレームワークが実論文再現で、最重要原著論文を引用できない、別ランで矛盾する結論を出す、テストセットのリークにより誤った結論に至る、といった脆弱性を示した［14］。

> **分析：**三分野に共通するのは、形式証明や既知指標の改善など検証器が明確な問題では AI が実質的に進展する一方、新規性・妥当性・研究価値の判断を要する問題では、当初の強い主張が独立検証や訂正により縮小される場合がある、という傾向である。ただし、個別成果の全体的価値を一括して否定するのではなく、主張の範囲を精密に区別する必要がある。

6. 再帰的自己改善予測への含意

対象論文は、爆発的 AI 進歩の予測が、AI による AI 研究の自動化に依存することを明示し、オープンエンドな研究能力がその臨界経路上にあるなら、当該能力の欠如は重要なボトルネックになると論じる［1］。Kapoor はアムダールの法則を援用し、AI が得意な工程を 100 倍速くしても、最も遅い工程が律速する限り、全体の加速は限定的だと主張する［5］。

ただし著者らは、RSI への道筋が、本研究のようなオープンエンドタスクの完全自動化を必要としない可能性、また本研究が測定した能力が臨界経路上にない可能性を明示的に認めている［1］。

表 8　RSI の時間軸をめぐる主要な立場

立場	内容

立場	内容
AI Futures Project	「AI 2027」の 2027 年は、公表時点におけるシナリオ上の最頻年（mode）であり、著者らの一律の中央値ではなかった。公表時の個人・モデル予測の中央値は概ね 2028〜2032 年に分布した［8］［35］。2025 年末の更新では、AI R&D 自動化の効果をより小さく見積もるモデル修正等により、フルコーディング自動化の中央値が約 3〜5 年後ろへ移動した［35］。
METR	時間ホライズンの指数的成長を強調する一方、対象がソフトウェア中心である点を留保［30］。
Elasticity Institute	フィードバックループの弾力性次第で加速も停滞もあり得ると形式化し、特定の時間軸を主張しない［9］［10］。
CRUX（対象論文）	今日のフロンティアモデルは、本評価条件下の数週間規模のオープンエンド AI 研究課題を解けなかったと報告。ただし、初期的証拠であると明示［1］。
AI ラボ	実行工程の自動化が急速に進んでいると主張しつつ、判断・taste が残余領域になり得ることは認める［11］［12］。

> **分析：**各立場の対立は、「検証可能タスクにおける急進」と「オープンエンド研究における停滞」という二つの観測をどう重みづけるかに帰着する。対象論文は「AI は研究できない」という一般命題ではなく、「今日のフロンティアエージェントは、この設計条件下で、2 件の数週間規模のオープンエンド AI 研究課題を公表水準まで解けなかった」という限定的主張として読むべきである。

7. 日本の企業実務・知財実務への含意

> **位置づけ：**本章は補足的な実務分析である。対象論文から直接確認できる事実と、企業・知財実務への推論を区別し、法制度については各法域の一次資料に基づく一般的整理にとどめる。

7.1 研究開発マネジメント（即時に適用可能）

対象論文が示した能力分布は、社内での AI エージェント活用設計に直接転用できる。文献レビュー、コード実装、計算環境のデバッグ、多数の実験実行、文書組版などは、限定的な運用介入の下で高い自律性を発揮し得る。他方、仮説の選択、停止判断、行き詰まりからの根本的方針転換、成果の公表水準判断は、人間が重点的に監督すべき領域である［1］。

実務上は、エージェントを「実験・分析アシスタント」と位置づけ、①仮説設定、②停止規則、③リソース予算、④重要な方針転換、⑤成果水準の評価を人間レビュー下に置くゲート設計が望ましい。棄却仮説、実行ログ、外部ツールの評価、介入履歴を記録として保全することも重要である。

7.2 発明者適格と AI（監視対象・法的に未確定）

本実験で評価されたエージェントについては、独創的着想や研究方針の判断を自律的に担ったとは評価されなかった。この観察は、「AI が利用された発明であっても、自然人の実質的貢献を具体的に把握・記録する」という現在の特許実務と整合する。ただし、本論文だけから AI 一般の発明的貢献能力について法的結論を導くことはできない。

日本では、特許制度小委員会の資料が、現状では発明の創作に人間の一定の関与が必要であり、技術的特徴部分の具体化に創作的に関与した自然人を発明者として整理する方向を示している［46］。米国では USPTO が、AI 利用の有無にかかわらず同一の発明者認定基準を適用し、発明者として適切に記載できるのは自然人だけであるとする改訂ガイダンスを公表している［47］。EPO の審判部は、EPC 上、機械は発明者ではないと判断し［48］、英国最高裁も DABUS を発明者とする主張を退けた［49］。

したがって、AI を用いた研究成果の権利化では、自然人が、どの技術的特徴の着想・選択・具体化にどのように寄与したかを、研究ノート、プロンプト、レビュー記録、実験判断、修正履歴などで残すことが重要である。個別案件では、各法域の最新法令・審査運用・判例を確認する必要がある。

7.3 AI 生成研究成果の知財保護とガバナンス

- エージェントが生成したネガティブ結果、データセット、コード、研究ノートの帰属と営業秘密管理を事前に設計する。
- 未公開データを AI に入力する際の秘密保持とアクセス制御を徹底する。対象論文では、エージェントがリポジトリにアクセストークンをコミットした事例も報告されている［1］。
- ベンダー側のモデル挙動、データ保持期間、拒否・ルーティング、能力制限を契約・運用ルールで可視化する。Fable 5 の事例は、同じモデル名でも特定領域で挙動が変わり得ることを示す［43］［44］。
- 対象論文の五つの失敗モードを、長期タスクの実務チェックリストに転用する。すなわち、判断、創造的転換、バックトラック、リソース認識、指示遵守を定期レビューする。

7.4 判断を更新すべき指標

次の兆候が観察された場合、「オープンエンド研究の自動化」に関する評価を上方修正すべきである。

- CRUX の追試で、複数の新規課題・複数モデル・複数スキャフォールドにわたりシャドー評価のスコアが明確に改善する。
- 検証可能タスクの急進が、独立専門家の評価でも公表水準の研究成果に波及する。
- エージェントが、根本的な方針転換、仮説の優先順位づけ、適切な停止判断を一貫して行う。
- 長時間・長期間タスクにおける性能が、短時間ベンチマークと同様に伸びる。
- METR の時間ホライズンや AI Futures Project の予測時間軸が、実証データを伴って再び前倒しされる。

逆に、これらが停滞する場合は、実行工程が急速に自動化されても、研究判断が全体を律速するというアムダール的ボトルネック仮説が支持される。

8. 留意事項（認識論的な限定）

- 対象論文は 2026 年 8 月公開のプレプリントであり、査読を経ていない。
- 独立した批評的議論は調査基準日時点で限定的である。日本語の要約・解説は確認できるが、日本の主要報道機関や研究者による詳細な独立批評は確認できていない。
- 結果は 2 課題・計 5 ラン、非ブラインド査読、研究者裁量という設計上の限界を負い、著者自身がこれを明示している。
- ロバストネス追試は TabPFN の 1 課題・1 追加構成に限られ、フロンティアモデル一般への一般化には追加評価が必要である。
- 言及されるモデルは新しく、性能・仕様の一部は二次情報に依拠する。予告された追加追試は対象論文時点で未実施である。
- ベンチマーク数値は、原論文、公式サイト、ラボの公表要約で定義・条件が異なる場合がある。特に PostTrainBench の公式 instruction-tuned モデルは、時間制限内の直接的な人間作業ベースラインではない。
- 発明者適格や AI 生成物の知財保護は各国で流動的であり、第 7 章は調査基準日時点の一般的運用に基づく。個別案件では専門的な法的検討を要する。

参考文献

[1] Kirgis, P., Kapoor, S., Schwartz, A., Rabanser, S. ほか（CRUX）「Can AI agents conduct open-ended AI research? Early evidence from two case studies」arXiv:2607.27191v2, 2026 年 8 月 7 日。 https://arxiv.org/abs/2607.27191 ［一次資料・検証済み］

[2] CRUX「Can AI agents conduct open-ended AI research? Early evidence from two case studies」プロジェクトページ（コード・レビュー・ログ公開）, 2026 年 7 月 30 日。 https://cruxevals.com/crux/can-ai-agents-conduct-research/ ［一次資料・検証済み］

[3] Hutson, M.「AI isn't ready to research itself」Nature News, 2026 年 8 月 13 日, DOI: 10.1038/d41586-026-02494-5。 https://www.nature.com/articles/d41586-026-02494-5 ［検証済み］

[4] MIT Technology Review「AI's recursive self-improvement might not come so quickly after all」2026 年 8 月 18 日。 https://www.technologyreview.com/2026/08/18/1142188/ai-recursive-self-improvement/ ［検証済み］

[5] Kapoor, S. & Narayanan, A.「AI agents can't yet do open-ended AI research」AI as Normal Technology, 2026 年 8 月 5 日。 https://www.normaltech.ai/p/ai-agents-cant-yet-do-open-ended ［著者解説・検証済み］

[6] Clark, J.「Import AI 467: Self-sustaining AI viruses; pacing AI progress; confusion about AI and creativity」Import AI, 2026 年 8 月 3 日。 https://jack-clark.net/2026/08/03/import-ai-467-self-sustaining-ai-viruses-pacing-ai-progress-confusion-about-ai-and-creativity/ ［検証済み］

[7] Baines, L. ほか「Persona Cartography: Charting Language Model Personality Traits in Weight Space」arXiv:2607.07916, 2026 年 7 月 8 日。 https://arxiv.org/abs/2607.07916 ［一次資料・検証済み］

[8] Kokotajlo, D. ほか「AI 2027」AI Futures Project, 2025 年 4 月。 https://ai-2027.com/ ［一次資料・検証済み］

[9] Cunningham, T. ほか「The Economics of Recursive Self-Improvement」Elasticity Institute Working Paper, 2026 年 7 月。 https://elasticity.institute/rsi-paper.pdf ［一次資料・検証済み］

[10] METR「The Economics of Recursive Self-Improvement」2026 年 7 月 22 日。 https://metr.org/notes/2026-07-22-economics-of-recursive-self-improvement/ ［検証済み］

[11] Favaro, M. & Clark, J.「When AI Builds Itself: Our Progress Toward Recursive Self-Improvement, and Its Implications」The Anthropic Institute, 2026 年 6 月。 https://www.anthropic.com/institute/recursive-self-improvement ［一次資料・検証済み］

[12] The Decoder「OpenAI's GPT-5.6 Sol autonomously post-trained the smaller Luna model with a fairly underspecified prompt」2026 年。 https://the-decoder.com/openais-gpt-5-6-sol-autonomously-post-trained-the-smaller-luna-model-with-a-fairly-underspecified-prompt/ ［二次情報］

[13] MindStudio「What Is GPT-5.6? OpenAI's Sol, Terra, and Luna Model Tiers Explained」。 https://www.mindstudio.ai/blog/what-is-gpt-5-6-sol-terra-luna-explained ［二次情報］

[14] Ahmed, M. O. ほか「Real Science Is Harder Than Benchmarks: Evaluating Advanced AI Frameworks on Published Studies. I.」bioRxiv, 2026 年 6 月 24 日, DOI: 10.64898/2026.06.24.734302。 ［一次資料・検証済み］

[15] Kapoor, S. ほか「Open-world evaluations for measuring frontier AI capabilities」arXiv:2605.20520, 2026 年。 https://arxiv.org/abs/2605.20520 ［対象論文の記載に基づく・個別検証未実施］

[16] Wijk, H. ほか「RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts」arXiv:2411.15114 / ICML 2025。 https://arxiv.org/abs/2411.15114 ［一次資料・検証済み］

[17] Chan, J. S. ほか「MLE-bench: Evaluating machine learning agents on machine learning engineering」arXiv:2410.07095 / ICLR 2025。 https://arxiv.org/abs/2410.07095 ［一次資料・検証済み］

[18] Starace, G. ほか「PaperBench: Evaluating AI's ability to replicate AI research」arXiv:2504.01848 / ICML 2025。 https://arxiv.org/abs/2504.01848 ［一次資料・検証済み］

[19] Siegel, Z. S. ほか「CORE-Bench」arXiv:2409.11363, 2024 年。 https://arxiv.org/abs/2409.11363 ［一次資料・検証済み］

[20] Nadgir, N. ほか「Life after benchmark saturation: A case study of CORE-Bench」arXiv:2606.26158, 2026 年。 https://arxiv.org/abs/2606.26158 ［対象論文の記載に基づく・個別検証未実施］

[21] Rank, B. ほか「PostTrainBench: Can LLM Agents Automate LLM Post-Training?」arXiv:2603.08640, 2026 年／公式サイト。 https://arxiv.org/abs/2603.08640 ［一次資料・検証済み］

[22] Beygelzimer, A. ほか「Has the machine learning review process become more arbitrary as the field has grown? The NeurIPS 2021 consistency experiment」arXiv:2306.03262。 https://arxiv.org/abs/2306.03262 ［一次資料・検証済み］

[23] Cortes, C. & Lawrence, N. D.「Inconsistency in conference peer review: Revisiting the 2014 NeurIPS experiment」arXiv:2109.09774。 https://arxiv.org/abs/2109.09774 ［一次資料・検証済み］

[24] Yamada, Y. ほか「The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search」arXiv:2504.08066, 2025 年。 https://arxiv.org/abs/2504.08066 ［一次資料・検証済み］

[25] Intology「Zochi publishes A* paper」Intology Blog, 2025 年 5 月 27 日。 https://www.intology.ai/blog/zochi-acl ［事業者発表・検証済み］

[26] Novikov, A. ほか「AlphaEvolve: A coding agent for scientific and algorithmic discovery」arXiv:2506.13131, 2025 年。 https://arxiv.org/abs/2506.13131 ［一次資料・検証済み］

[27] Merchant, A. ほか「Scaling deep learning for materials discovery」Nature 624, 2023 年, DOI: 10.1038/s41586-023-06735-9。 https://www.nature.com/articles/s41586-023-06735-9 ［一次資料・検証済み］

[28] Szymanski, N. J. ほか「An autonomous laboratory for the accelerated synthesis of inorganic materials」Nature 624, 2023 年, DOI: 10.1038/s41586-023-06734-w。 https://www.nature.com/articles/s41586-023-06734-w ［一次資料・検証済み］

[29] Google Research「Accelerating scientific breakthroughs with an AI co-scientist」2025 年。 https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/ ［公式発表・検証済み］

[30] METR「Task-Completion Time Horizons of Frontier AI Models」／「Measuring AI Ability to Complete Long Software Tasks」2025 年 3 月 19 日。https://metr.org/time-horizons/ ［一次資料・検証済み］

[31] TechCrunch「OpenAI and Google outdo the mathletes, but not each other」／Axios「Google and OpenAI are vying for top AI mathlete」2025 年 7 月 21 日。 https://techcrunch.com/2025/07/21/openai-and-google-outdo-the-mathletes-but-not-each-other/ ［IMO 2025 の二次報道］

[32] Sothanaphan, N.「Resolution of Erdős Problem #728: a writeup of Aristotle's Lean proof」arXiv:2601.07421, 2026 年 1 月。https://arxiv.org/abs/2601.07421 ［一次資料・検証済み］

[33] Time「What Happens When AI Starts Building AI? Inside Recursive Self-Improvement」2026 年 8 月 7 日。https://time.com/article/2026/08/07/ai-recursive-self-improvement-anthropic-openai/ ［二次情報］

[34] NeurIPS「NeurIPS 2026 Reviewing Guidelines」。 https://neurips.cc/Conferences/2026/ReviewerGuidelines ［公式資料］

[35] AI Futures Project「Timelines Forecast」／「AI Futures Model: Dec 2025 Update」／「Clarifying how our AI timelines forecasts have changed」2025〜2026 年。 https://blog.ai-futures.org/p/clarifying-how-our-ai-timelines-forecasts ［一次資料・検証済み］

[36] Pebblous Blog、AIZIGOO、EvalCommunity Academy 等による二次解説記事。 ［二次情報］

[37] TechCrunch「OpenAI's embarrassing math」2025 年 10 月 19 日。 https://techcrunch.com/2025/10/19/openais-embarrassing-math/ ［Erdős 問題をめぐる二次報道］

[38] AIDB「AI エージェントはオープンエンドな AI 研究を行えるのか？2 つのケーススタディからの初期の証拠」2026 年 7 月 29 日。 https://ai-data-base.com/paper/2607-27191 ［日本語論文紹介］

[39] HyperAI「AI エージェントはオープンエンドな AI 研究を遂行できるか？2 つの事例研究からの初期の証拠」2026 年。https://hyper.ai/ja/papers/2607.27191 ［日本語論文紹介］

[40] 「AI 研究エージェント、6 日間のシャドー評価で判明した限界」日本語解説記事, 2026 年 8 月 2 日。 https://hanchizenkai.com/ai-research-agent-six-days-shadow-evaluation/ ［日本語二次解説］

[41] Cheetham, A. K. & Seshadri, R.「Artificial Intelligence Driving Materials Discovery? Perspective on the Article: Scaling Deep Learning for Materials Discovery」Chemistry of Materials 36, 2024, DOI: 10.1021/acs.chemmater.4c00643。 https://pmc.ncbi.nlm.nih.gov/articles/PMC11044265/ ［独立批評・一次資料］

[42] Szymanski, N. J. ほか「Author Correction: An autonomous laboratory for the accelerated synthesis of inorganic materials」Nature 650, E1, 2026 年 1 月 19 日（issue date 2026 年 2 月 5 日）, DOI: 10.1038/s41586-025-09992-y。 https://www.nature.com/articles/s41586-025-09992-y ［一次資料・検証済み］

[43] Zeff, M.「Anthropic Walks Back Policy That Could Have 'Sabotaged' AI Researchers Using Claude」WIRED, 2026 年 6 月 10 日。https://www.wired.com/story/anthropic-responds-to-backlash-on-claudes-secret-sabotage-on-ai-research/ ［二次報道・検証済み］

[44] Anthropic「Claude Fable 5 & Claude Mythos 5 System Card」2026 年 6 月 9 日。 https://www-cdn.anthropic.com/d00db56fa754a1b115b6dd7cb2e3c342ee809620.pdf ［公式一次資料・検証済み］

[45] Intology「Scaling Automated Post-Training」2026 年 8 月 3 日。 https://intology.ai/blog/scaling-automated-post-training ［事業者技術報告。44.7／51.6、4,000 H100 時間超を報告］

[46] 特許庁・産業構造審議会知的財産分科会「第 51 回特許制度小委員会資料」AI による自律的な発明の取扱いに関する整理。https://www.jpo.go.jp/resources/shingikai/sangyo-kouzou/shousai/tokkyo_shoi/document/index/newtokkyo_051.pdf ［公的一次資料］

[47] USPTO「Revised inventorship guidance for AI-assisted inventions」2025 年 11 月 26 日。 https://www.uspto.gov/subscription-center/2025/revised-inventorship-guidance-ai-assisted-inventions ［公的一次資料］

[48] EPO Boards of Appeal, J 0008/20（Designation of inventor/DABUS）, 2021 年 12 月 21 日。 https://www.epo.org/en/boards-of-appeal/decisions/j200008eu1 ［公的一次資料］

[49] UK Supreme Court, Thaler v Comptroller-General of Patents, Designs and Trade Marks, [2023] UKSC 49, 2023 年 12 月 20 日。https://www.supremecourt.uk/cases/uksc-2021-0201 ［公的一次資料］

> **参考文献の表記：**「検証済み」は、本調査で URL、DOI または公式資料により実在と主要内容を確認したものをいう。「二次情報」は、一次資料ではなく報道・解説・事業者発表に基づくことを示す。調査基準日後に内容が更新される可能性がある。