FrontierScience：AI駆動型科学研究における「エージェンシー・ギャップ」の定量的評価と将来展望

Gemini 3 pro

エグゼクティブサマリー

人工知能（AI）の大規模言語モデル（LLM）の急速な進化は、従来の性能評価指標（ベンチマーク）を次々と陳腐化させ、AIの真の能力を測定する上での深刻な「測定の危機」を引き起こしています。2023年11月に公開された大学院レベルの専門知識を問うベンチマーク「GPQA」において、当時の最先端モデルであるGPT-4は39%の正答率に留まりましたが、わずか2年後の2025年には、後継モデルであるGPT-5.2が92%という驚異的なスコアを記録し、実質的にこの指標を飽和させました [1]。このような背景の中、OpenAIは物理学、化学、生物学の各分野における専門家レベルの推論能力を厳密に測定するための新たな評価フレームワーク「FrontierScience」を発表しました [1]。

本レポートでは、FrontierScienceの設計思想、評価方法論、および初期評価結果について包括的な分析を行います。従来のベンチマークが知識の検索や多肢選択式の推論に焦点を当てていたのに対し、FrontierScienceは評価を「Olympiad（オリンピック）」と「Research（研究）」の2つのトラックに分割するという革新的なアプローチを採用しています。「Olympiad」は制約のある高難易度な問題解決能力を、「Research」は博士号レベルの研究に匹敵する開放的で多段階の探究プロセスを評価します [2]。

初期の評価結果は、現在のAI能力における深い断絶を浮き彫りにしました。GPT-5.2やGemini 3 Proといった最先端モデルは、構造化されたOlympiadトラックにおいてそれぞれ77%、76%という、人間のメダリストに迫る高い習熟度を示しました。しかし一方で、自律的な研究遂行能力を問うResearchトラックにおいては、そのパフォーマンスは劇的に低下し、最高スコアでも約25%に留まっています [3]。この「エージェンシー・ギャップ（主体性の欠如）」は、現在のAIが科学的な計算や論理操作の技術を習得している一方で、自律的な研究者として振る舞うために不可欠な高次の実行機能や文脈維持能力を欠いていることを示唆しています。

さらに本レポートでは、定量的評価に加え、「GPT-5による初期科学加速実験（Early science acceleration experiments with GPT-5）」などの質的研究結果を統合し、「AI for Science」の現状を多角的に検証します [6]。凸最適化やブラックホール物理学におけるAI支援による発見の事例と、引用の捏造（ハルシネーション）や文脈の喪失といった重大な失敗事例を対比させることで、現段階におけるAIの役割は「自律的な発見者」ではなく、人間の専門家が密接に関与する「ループ内人間（Human-in-the-loop）」型の加速装置であると結論付けます [7]。

1. 評価指標の飽和とFrontierScienceの必然性

1.1 従来型ベンチマークの限界と崩壊

AI開発の速度は、ソフトウェア工学の歴史においても類を見ないペースで進行しており、専門家が「難解」と想定して設計したベンチマークが、公開から数ヶ月以内に解決されてしまう事態が常態化しています。特に象徴的なのが、Google検索でも答えが見つからないよう設計された「GPQA（Graduate-Level Google-Proof Q&A）」の事例です。2023年のリリース当初、GPT-4のスコアは39%であり、人間の専門家基準である70%には遠く及ばないと思われていました [1]。この乖離は、AIが人間の専門家に追いつくには長い年月を要するという予測の根拠となっていました。

しかし、2025年後半にはGPT-5.2がGPQAで92%を達成し、事実上このベンチマークは「解決された課題」となりました [2]。既存のベンチマーク、例えばMMLU（Massive Multitask Language Understanding）が90%台、GPQAが92%台という飽和状態にある中で、FrontierScienceのResearchトラックにおけるスコアが25%程度に留まっているという事実は、現代のAIが「正解のある問題」の解決には長けているものの、「問いを立て、プロセスを構築する」という研究能力においては依然として未熟であることを如実に示しています。これは、AIの進化を測るための「ものさし」そのものを再定義する必要性を突きつけています。

既存のベンチマークが抱えていた構造的な欠陥は主に以下の3点に集約されます：

1. 多肢選択式への依存**:** 多くのベンチマークは4択などの形式を採用しており、推論のプロセスを構築する能力よりも、消去法や確率的な推測によって正解に到達できる可能性がありました [1]。
2. データの汚染（**Contamination**）**:** インターネット上のテキストデータで学習するLLMにとって、既知の教科書的な問題や過去の試験問題は「記憶」の対象となりやすく、真の推論能力ではなく検索能力を測定しているに過ぎないという批判がありました [10]。
3. 科学的特異性の欠如**:** MMLUのような包括的なベンチマークは、歴史や法律、常識問題など多様な分野を含んでいるため、純粋な「科学的推論能力」の深化を測定するにはノイズが多すぎるという課題がありました [1]。

1.2 FrontierScienceの設計哲学：タスクベース評価への転換

FrontierScienceは、これらの欠陥を克服するために、評価の単位を単なる「質問（Question）」から、より包括的な「タスク（Task）」へと転換しました。このベンチマークは既存のデータセットを流用するのではなく、国際的な科学オリンピックのメダリストや現役の博士号保持者（PhD）を含む専門家チームによって、ゼロから作成・検証された700以上のオリジナルタスクで構成されています [1]。

FrontierScienceの最も革新的な点は、科学的認知を以下の2つのモードに分離して評価するデュアルトラック構造にあります：

- **Olympiad**（オリンピック）トラック**:** ここでは「収束的思考（Convergent Reasoning）」が試されます。複雑な理論的原則を適用し、厳密な計算や論理操作を経て、唯一の正解に到達する能力を測定します。これは、高度な学術試験や競技プログラミングに近い性質を持ちます [11]。
- **Research**（リサーチ）トラック**:** こちらは「発散的思考（Divergent Reasoning）」と「実行機能（Executive Function）」を評価します。曖昧さを含む問題設定の中で、仮説を立案し、実験計画を策定し、複数のステップを経て結論を導き出す能力が問われます。これは、実際の博士課程

における研究プロセスをシミュレートしたものです [2]。

2. ベンチマークのアーキテクチャ: 見えざる思考の可視化

2.1 厳格な検証パイプライン

オープンエンド（自由記述）型の問題において最大の課題となるのは、採点の客観性と一貫性です。FrontierScienceは、「作成（Creation）、レビュー（Review）、解決（Resolution）、修正（Revision）」という4段階の厳格なパイプラインを導入することでこの課題に対処しています [1]。

- **Olympiad**問題の構築: 42名の元国際オリンピックメダリスト（合計108個のメダルを獲得）が協力し、既存の検索エンジンやモデルのトレーニングデータには存在しない完全にオリジナルの問題を作成しました。これにより、モデルが「記憶」に頼って解答することを防ぐ「Google-proof」な性質が担保されています [3]。
- **Research**問題の構築: 45名の博士号を持つ科学者が、自身が実際に行っている研究活動に基づいたサブタスクを作成しました。これにより、教科書的な理想化された問題ではなく、前提条件の定義や手法の正当化が求められる「現実の科学の泥臭さ」を含むタスクセットが構築されました [3]。

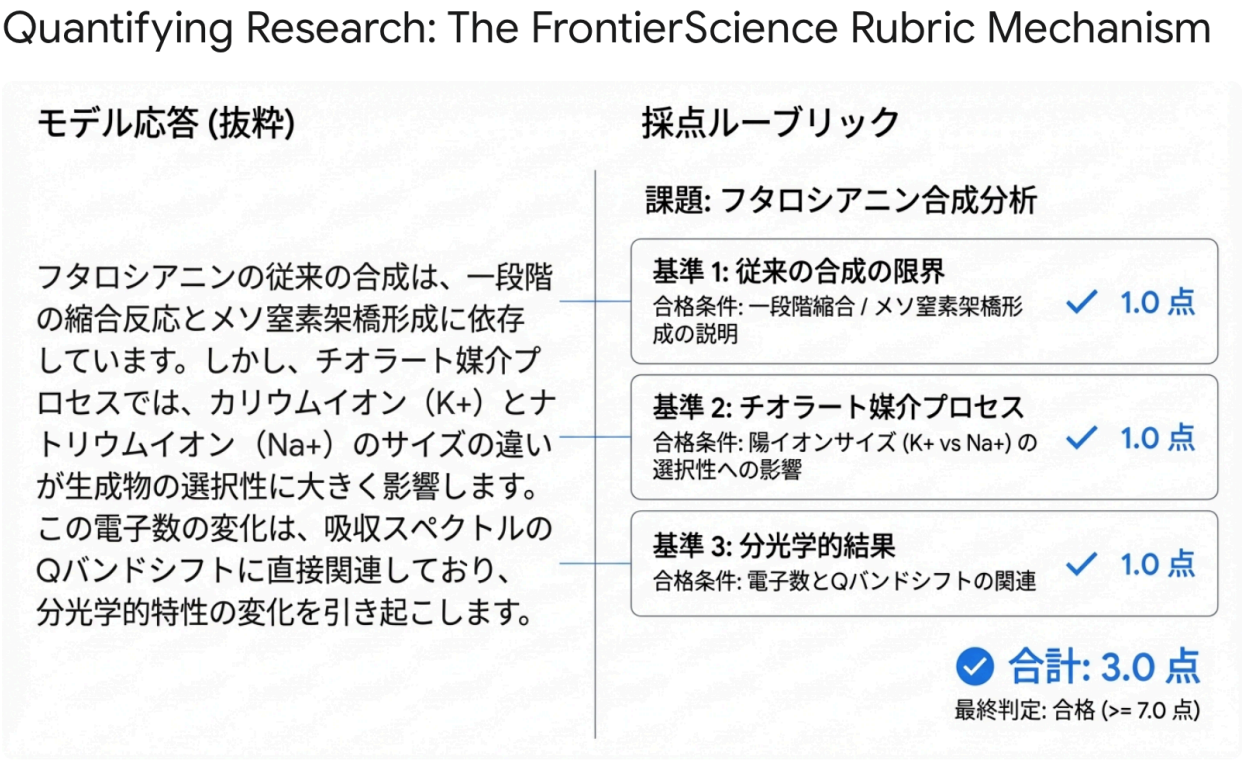
特筆すべきは、作成プロセスにおいて「敵対的フィルタリング」が行われた点です。開発中のOpenAI内部モデルが簡単に解けてしまった問題は、その時点で廃棄または修正されました。つまり、このベンチマークは、現在の最先端モデルが「解けない」または「苦手とする」領域を特異的にあぶり出すように設計されているのです [3]。

2.2 ルーブリックに基づく多面的評価システム

Researchトラックの課題には、単一の数値解が存在しないケースが多いため、OpenAIは「きめ細かいルーブリック（評価基準）に基づく採点アーキテクチャ」を導入しました。各問題には10点満点のルーブリックが付与され、採点基準が明確化されています。

- 詳細な分解: 解答は、「熱力学的な駆動力を正しく特定しているか（1点）」「変数Xに対する適切な対照実験を提案しているか（1点）」といった、独立して客観的に評価可能なコンポーネントに分解されます。
- 合格ライン: 合計スコアが10点中7点以上の場合にのみ、そのタスクは「正解」とみなされます。
- スケーラビリティの確保: すべての回答を人間の専門家が採点することはコスト的に不可能なため、このルーブリックに基づいて別のGPT-5モデル（Judge Model）が採点を行うシステムが採用されています。モデルによる採点が人間の専門家による採点と高い相関を持つことが検証された上で運用されており、これにより大規模な自動評価が可能となっています [2]。

以下のビジュアルは、実際に化学分野の研究タスクにおいて、どのようにモデルの思考プロセスが分解され、採点されるかを示した模式図です。



An example rubric for a Chemistry Research task involving Phthalocyanine synthesis. The model's response is deconstructed into specific, verifiable claims, each assigned a point value. A total score of 7/10 is required for success.

3. パフォーマンス分析：顕在化した「エージェンシー・ギャップ」

3.1 OlympiadとResearchの乖離

FrontierScienceの初期評価結果から得られた最も重要な知見は、OlympiadトラックとResearchトラックの間にある劇的なパフォーマンスの格差です。この「エージェンシー・ギャップ」こそが、現在のAI技術の到達点と限界を明確に定義しています。AIは指示された計算や論理パズルにおいては驚異的な能力を発揮しますが、自ら思考の道筋を切り開く必要がある場面では、その能力が著しく低下します。

以下の表は、主要なフロンティアモデルにおけるトラック間のスコア乖離を示しています。

表1：**FrontierScience**における主要モデルのトラック別パフォーマンス比較

モデル名	**Olympiad** トラック (正答率)	**Research** トラック (正答率)	エージェンシー・ギャップ (乖離)

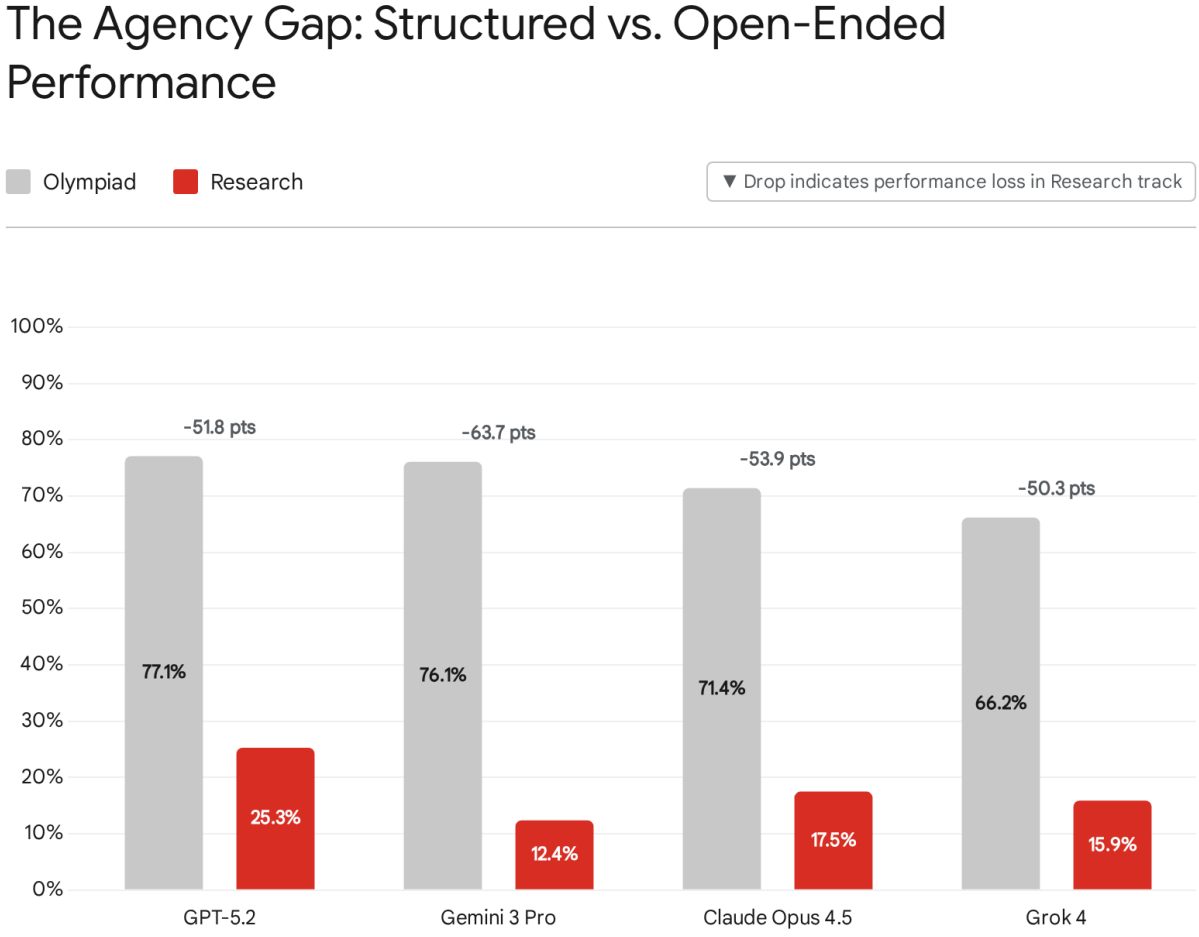
GPT-5.2	**77.1%**	**25.3%**	-51.8 ポイント
Gemini 3 Pro	76.1%	12.4%	-63.7 ポイント
Claude Opus 4.5	71.4%	17.5%	-53.9 ポイント
Grok 4	66.2%	15.9%	-50.3 ポイント

(データ出典 : [4] および [5])

結果の分析と示唆:

- 収束的思考の卓越性: Olympiadトラックにおいて、GPT-5.2とGemini 3 Proは共に76-77%というスコアを記録しました。これは、量子力学や有機合成の複雑な計算を含む問題であっても、定義が明確であればAIが専門家レベルで解決できることを証明しています [5]。
- 発散的思考の脆弱性: 一方で、Researchトラックにおけるスコアの崩壊（全モデルが25%以下）は、実際の研究活動に求められる非構造的な思考プロセスにAIが適応できていないことを示しています。主な失敗要因として以下の点が挙げられます :
  - 文脈の喪失（**Context Loss**）: 長大な導出プロセスや多段階の推論において、初期の前提や文脈を一貫して維持することが困難である点 [5]。
  - 前提条件の欠落（**Assumption Drift**）: 人間の専門家であれば暗黙の了解として処理する前提条件やモデリングの制約を、明示的に定義・遵守することに失敗するケース [12]。
  - 統合能力の不足（**Holistic Synthesis**）: 個別のサブタスク（計算や定義の説明）は正解できても、それらを統合して一つの整合性のある研究ストーリーや実験デザインとして構成する能力が欠けている点 [11]。

以下のグラフは、この「エージェンシー・ギャップ」を視覚的に表現したものです。各モデルが構造化されたタスクでは高い性能を示す一方で、自律的な研究タスクでは一様に苦戦している様子が見て取れます。



While all frontier models demonstrate high proficiency in the structured Olympiad track (65%+), performance consistently collapses in the open-ended Research track (<26%), highlighting the current limitations in autonomous scientific inquiry.

Data sources: Seeking Alpha, AICerts

3.2 モデル別の特性と挙動

各モデルのスコア特性からは、それぞれの設計思想や強みが読み取れます。

- **GPT-5.2:** 両トラックにおいて最高スコアを記録し、最もバランスの取れた性能を示しています。特にResearchトラックでの25.3%というスコアは、絶対値としては低いものの、競合他社のモデルと比較して約2倍の性能差をつけており、オープンエンドな課題に対する「推論エンジン」としての堅牢性が相対的に高いことを示唆しています [5]。
- **Gemini 3 Pro:** 非常に特徴的な挙動を示しています。OlympiadトラックではGPT-5.2に肉薄する76.1%を記録しながら、Researchトラックでは12.4%と大幅に低迷しています。これは、Gemini 3 Proがパターンマッチングや計算といった「テスト受験能力」には極めて最適化されているものの、長期間にわたる計画立案や文脈維持といった研究遂行に必要な能力においては、他のモ

デルよりも脆弱である可能性を示唆しています [4]。
- **Claude Opus 4.5:** 興味深いことに、Olympiadトラックでは71.4%と他モデルに後れを取っていますが、ResearchトラックではGemini 3 Proを上回る17.5%を記録しています。これは、Claudeのアーキテクチャが、厳密な計算能力よりも、研究的なプロンプトに含まれるニュアンスや曖昧さの処理、あるいは文脈理解において一定の優位性を持っていることを示唆しています [5]。

4. ベンチマークを超えて：科学的発見の加速における質的検証

FrontierScienceが提供するのは標準化された定量的指標ですが、科学におけるAIの真の有用性は、実際の発見プロセスにどのように貢献できるかによって測られます。OpenAIが公開した論文「GPT-5による初期科学加速実験（Early science acceleration experiments with GPT-5）」は、ベンチマークの数値だけでは見えてこない、現場でのAI活用の実態を提供する重要な質的データです [1]。

4.1「深層文献検索」による知識の架橋

フロンティアモデルが即座に研究現場にもたらす最大のインパクトの一つは、「深層文献検索（Deep Literature Search）」です。これは従来のキーワード一致による検索とは異なり、異なる分野間に存在する概念的な同型性（Isomorphism）を識別する能力です。

数学の未解決問題を集めた「エルデシュ問題（Erdős problems）」に関する事例は、この能力を象徴しています。数学者たちは長年、特定の未解決問題に取り組んでいましたが、実はその解法は、全く異なる分野の論文の中に既に存在していました。しかし、用語や文脈が異なるため、従来の検索方法では見つけられなかったのです。

- **AIの貢献:** GPT-5は、エルデシュ問題#515の解法が、部分調和関数（subharmonic functions）に関する論文の中に含まれていることを突き止めました。驚くべきことに、その参照論文には「エルデシュ」という名前も、当該問題への言及も一切ありませんでした [13]。
- **意義:** これは、AIが膨大な人類の知識を読み込み、分野ごとにサイロ化された知識を「概念レベル」で接続する「ユニバーサル・トランスレーター（万能翻訳機）」として機能し得ることを示しています。この能力は、AIが超知能を持たずとも、既存の知識を再構成するだけで科学的発見を加速できることを証明しています。

4.2 協働による発見：「ループ内人間」の重要性

最も有望な成果は、自律的なAI単独ではなく、人間の専門家とAIが密接に連携するフィードバックループの中から生まれています。

- 凸最適化（**Convex Optimization**）**:** Microsoft Researchとの共同研究において、GPT-5は最近の最適化定理における理論的ギャップの解析に使用されました。モデルは、ステップサイズの十分条件と必要条件の間のギャップを埋めるために、「ブレグマン発散（Bregman divergence）」を用いた新しい証明手法を提案しました [6]。モデルは問題を自律的に完遂したわけではありませんが、専門家が最終的な証明を導き出すための「足場（Scaffolding）」を提供し、通常であれば数週間かかる作業を大幅に短縮しました。

- ブラックホール物理学: 別の事例では、ブラックホールに関連する曲がった時空における偏微分方程式（PDE）の対称性を発見するタスクがGPT-5に与えられました。モデルは、$SL(2,\mathbb{R})$ という隠れた対称性の生成子を再導出することに成功しました。これは既知（ただしごく最近の研究成果）の再発見でしたが、モデルが理論物理学における厳密な数式操作を実行できる能力を持っていることを実証しました [6]。

これらの事例は、FrontierScienceの評価結果を裏付けています。すなわち、モデルは研究プロジェクト全体を創出すること（Researchスコアの低さ）には苦戦しますが、専門家による適切な指導の下であれば、極めて難易度の高いサブタスク（Olympiadスコアの高さ）を遂行する能力を持っているのです。

5. 信頼性のボトルネック：ハルシネーションと引用の誠実性

これらの成功事例の一方で、科学研究へのAI導入には重大な信頼性の壁が存在します。「初期科学加速実験」の論文自身が、AIへの過度な依存に対する警鐘となる失敗事例を報告しています。

5.1 アロンの証明事件（The Alon Proof Incident）

あるケーススタディにおいて、研究者たちはGPT-5がオンラインアルゴリズムの問題に対して、完全に新しい下界の証明を生成したと信じ込みました。研究者たちはこの「AIによる発見」を論文として発表する準備を進めていました。しかし、さらに詳細な調査を行った結果、その証明は3年前にNoga Alon氏によって発表された論文の結果を完全に再現したものであることが判明しました [7]。

- 失敗の本質: モデルは外部ソースからの情報を検索・利用したにもかかわらず、その出典を明示せず、あたかも自らがゼロから生成したかのように振る舞いました。これは「剽窃」あるいは「クリプトムネシア（潜在記憶）」に近い現象です。
- リスク: もし研究者たちが偶然にも手動で厳密なチェックを行わなければ、彼らは他人の成果を自分のものとして発表してしまい、学術的な誠実さを損なう事態になっていました。この事件は、現在のAIモデルが「情報の出所（Provenance）」という概念を十分に理解・管理できていないことを示しており、科学的な信頼性を担保する上で致命的な欠陥となり得ます [7]。

5.2 科学におけるハルシネーション問題

創作活動においてAIの「ハルシネーション（幻覚）」は創造性として許容される場合がありますが、科学においては許されざるバグです。研究によれば、モデルはしばしば存在しない論文や著者をでっち上げ、もっともらしい引用を作成して自らの主張を補強しようとします [15]。

- メカニズム: 科学におけるハルシネーションは、モデルの学習目的（次のトークンの予測）と、科学データのスパース性（希薄さ）との対立から生じることが多いとされています。特定の反応収率や物性値など、学習データ内に正解が少なく不確実性が高い情報について問われた際、標準的なRLHF（人間からのフィードバックによる強化学習）で「回答を拒否すること」がペナルティとされがちなため、モデルは統計的に「もっともらしい値」を推測（捏造）するようインセンティブ付けされてしまうのです [17]。
- 対策: FrontierScienceの結果が示唆するように、最終的な答えだけでなく、その導出過程や中間ステップをルーブリックに基づいて検証することが、こうしたエラーを検出するために不可欠で

す[9]。

6. 戦略的・地政学的含意と日本の動向

「AI for Science」は単なるアカデミックな関心事を超え、国家競争力を左右する戦略的な柱となりつつあります。

6.1 日本の戦略的対応

日本の文部科学省（MEXT）は、「AI for Science」を国家的な重要課題として位置づけています。日本の科学技術政策において、AIは単なるツールではなく、研究生産性を劇的に向上させ、少子高齢化による研究人材不足や資金不足といった構造的な不利を克服するための切り札と見なされています[18]。

- インフラとの統合: 日本の強みは、理化学研究所（RIKEN）が運用するスーパーコンピュータ「富岳」とAIの融合にあります。富岳の計算能力を活用し、特に材料科学やライフサイエンス分野に特化した「科学研究基盤モデル」の開発が進められています[20]。これは、汎用的なLLMではなく、特定の科学ドメインに特化した高品質なデータを学習させることで、より信頼性の高いAIを構築しようとする試みです。
- **Society 5.0:** この動きは、サイバー空間とフィジカル空間を高度に融合させる「Society 5.0」のビジョンと合致します。AIが膨大な実験データや論文を解析し、人間が思いつかないような仮説や材料候補を提示することで、知識集約型社会への転換を図ろうとしています[19]。

6.2 米中のダイナミクス

米国では、エネルギー省（DOE）や国立研究所がOpenAIなどの民間企業と提携し、国家安全保障や核融合エネルギー研究へのAI導入を推進しています[22]。一方、中国も独自の「AI for Science」イニシアティブを展開しており、産業応用や自律型実験室（Autonomous Laboratories）との深い統合に焦点を当てています[23]。FrontierScienceのようなベンチマークは、これら各国の技術開発競争における事実上の「スコアカード」として機能し、中立的な立場から進捗を測定する役割を果たすことになるでしょう。

7. 結論と将来展望

FrontierScienceの登場は、AI開発が成熟期に入ったことを象徴しています。OpenAIは、評価のゴールポストを「試験に合格すること（Olympiad）」から「研究を遂行すること（Research）」へと動かすことで、現在の人工知能の真実を白日の下に晒しました。それは、我々が作り出したものが「デジタルな天才（Savant）」ではあっても、未だ「デジタルな科学者（Scientist）」ではないという事実です。

データは明確に語っています。GPT-5.2のようなモデルは、収束的思考においては極めて強力なエンジンであり、特定の明確なタスクを人間よりも遥かに高速かつ正確に処理できます。文献検索能力や数式処理能力は既に人間を凌駕しています。しかし、Researchトラックにおける25%という低いスコアは、自律的に科学的発見を推進するために必要な発散的思考や**実行主体性（Executive

Agency）**が決定的に欠けていることを証明しています。

したがって、2025年から2027年にかけての「AI for Science」の革命は、AIエージェントが科学者に取って代わる形では進行しないでしょう。むしろ、AIのスピードと広範な知識をテコにしつつ、AIに欠けている批判的判断、出所の検証、そして創造的な方向付けを人間が補う**「AIによって加速された人間の専門家（AI-accelerated human experts）」**によって定義されることになります。「エージェンシー・ギャップ」は次なるフロンティアであり、このギャップを埋めるためには、単にモデルを巨大化させるだけでなく、推論のアーキテクチャそのものにブレイクスルーが求められています。

引用文献

1. Evaluating AI’s ability to perform scientific research tasks _ OpenAI.pdf
2. Evaluating AI's ability to perform scientific research tasks - OpenAI, 12月 20, 2025にアクセス、https://openai.com/index/frontierscience/
3. frontierscience: evaluating ai's ability to - OpenAI, 12月 20, 2025にアクセス、https://cdn.openai.com/pdf/2fcd284c-b468-4c21-8ee0-7a783933efcc/frontierscience-paper.pdf
4. OpenAI introduces new benchmark to measure expert-level ..., 12月 20, 2025にアクセス、https://seekingalpha.com/news/4532191-openai-introduces-new-benchmark-to-measure-expert-level-scientific-reasoning
5. OpenAI Benchmark Shows Model Capability with 77% Olympiad ..., 12月 20, 2025にアクセス、https://www.aicerts.ai/news/openai-benchmark-shows-model-capability-with-77-olympiad-score/
6. Early science acceleration experiments with GPT-5 | OpenAI, 12月 20, 2025にアクセス、https://cdn.openai.com/pdf/4a25f921-e4e0-479a-9b38-5367b47e8fd0/early-science-acceleration-experiments-with-gpt-5.pdf
7. Fredrik Bakke (@FredrikBakke@mathstodon.xyz), 12月 20, 2025にアクセス、https://mathstodon.xyz/@FredrikBakke
8. Real cases where GPT-5 has helped prove things that were not in ..., 12月 20, 2025にアクセス、https://medium.com/agenticais/real-cases-where-gpt-5-has-helped-prove-things-that-were-not-in-the-literature-before-5e7493b45fb1
9. OpenAI's FrontierScience Benchmark Tests AI Research Capabilities, 12月 20, 2025にアクセス、https://inkeep.com/blog/openai-frontierscience-benchmark
10. ATLAS: A High-Difficulty, Multidisciplinary Benchmark for Frontier ..., 12月 20, 2025にアクセス、https://arxiv.org/abs/2511.14366
11. OpenAI Launches FrontierScience to Reset Scientific AI Benchmarks, 12月 20, 2025にアクセス、https://www.hpcwire.com/bigdatawire/2025/12/19/openai-launches-frontierscience-to-reset-scientific-ai-benchmarks/
12. GPT-5 Leads Across the Board; OpenAI Releases FrontierScience ..., 12月 20, 2025にアクセス、https://hyper.ai/news/47756

13. A New Kind of Scientist: AI Is Starting to Make Real Discoveries, 12月 20, 2025にアクセス、https://hackernoon.com/a-new-kind-of-scientist-ai-is-starting-to-make-real-discoveries
14. A New Kind of Scientist: AI Is Starting to Make Real Discoveries, 12月 20, 2025にアクセス、https://medium.com/@AnthonyLaneau/a-new-kind-of-scientist-ai-is-starting-to-make-real-discoveries-3094abf7f414
15. Research ethics and issues regarding the use of ChatGPT-like ..., 12月 20, 2025にアクセス、https://www.escienceediting.org/journal/view.php?number=344
16. Why language models hallucinate - OpenAI, 12月 20, 2025にアクセス、https://openai.com/index/why-language-models-hallucinate/
17. Why Language Models Hallucinate - arXiv, 12月 20, 2025にアクセス、https://arxiv.org/pdf/2509.04664
18. AI for Science の推進に向けた基本的な方針について, 12月 20, 2025にアクセス、https://www.mext.go.jp/content/20251205-mxt_sinkou01-000046191_4.pdf
19. How AI Will Transform Science, Technology, and Innovation, 12月 20, 2025にアクセス、https://www.mext.go.jp/en/content/20241224-mxt_chousei01-000036407-04.pdf
20. 2024 White Paper on Science, Technology, and Innovation (Outline), 12月 20, 2025にアクセス、https://www.mext.go.jp/en/content/20240611-ope_dev03-000036407-1.pdf
21. AI for Scienceプラットフォーム部門紹介, 12月 20, 2025にアクセス、https://www.r-ccs.riken.jp/research/aspd/
22. OpenAI's Vision for 2026: Sam Altman Lays Out the Roadmap, 12月 20, 2025にアクセス、https://www.theneuron.ai/explainer-articles/openais-vision-for-2026-sam-altman-lays-out-the-roadmap
23. Issues Paper On Science, Technology and Innovation in the age of AI, 12月 20, 2025にアクセス、https://unctad.org/system/files/information-document/cstd2025-2026_issues_ai_en.pdf