ChatGPT

FrontierScienceベンチマークの詳細分析

ベンチマークの構成

2つのトラック（OlympiadとResearch）の概要と目的:
FrontierScienceは、物理・化学・生物学の専門レベルの科学的推論力を評価するために設計された新たなベンチマークであり、**Olympiadトラック**と**Researchトラック**の2種類の問題で構成されています [1]。Olympiadトラックは**国際科学オリンピック**形式の問題で、**優秀な若手が挑戦する競技レベル**の理論問題を扱い、**構造化された短答形式**で解答します [2]。一方、Researchトラックは**博士号レベルの科学者**が作成した**よりオープンエンドな研究課題**で、**実際の研究に近いマルチステップ問題**を含みます [3]。これにより、Olympiadでは**高度な科学的推論力**を、Researchでは**実際の科学研究遂行能力**をそれぞれ測定することを目的としています [4]。

問題形式:
- **Olympiadトラック:** 全部で100問からなり、**数式や数値、専門用語**など**短い答え**を要求する形式です [2] [5]。各問題は**国際オリンピックメダリスト**によって作成され、難易度は国際大会レベルと同等以上になるよう設計されています [6]。解答は数値や式、単語など**短文で提出**し、正答かどうかは**数値誤差の許容範囲や文字列のファジーマッチ**で判定されます [5]。このように短答形式にすることで**自動評価**がしやすくなっていますが、その分問題の表現力・自由度が制限される側面もあります [7]。

- **Researchトラック:** 全部で60問の**サブタスク**（研究課題）が含まれ、**自由記述の論述形式**で回答します [8]。各問題は**現役の博士号取得者（博士課程学生・研究員・教授）**によってオリジナルに作成されており、**博士研究者が直面しうる難題**を再現したマルチステップの問いになっています [9]。例えば、ある問題ではニッケル(II)フタロシアニンのメソ位窒素原子の修飾が電子構造や反応性に与える影響について分析させるなど、**専門性が高く複雑な問い**が提示されます [10] [11]。回答は数段落にわたる議論・計算・考察を含むことが想定されており、それを**総合的に評価する仕組み**が必要になります。

評価基準とルーブリック:
Olympiadトラックは答えが**一意に定まる短答問題**のため、自動採点では**正解の数値・表記**と一致するか（または許容範囲内か）で評価します [5]。一方、Researchトラックでは答えが自由記述となるため、**10点満点の採点ルーブリック**が各問題に用意されています [12]。ルーブリックは**複数の独立した客観項目**（例：中間推論の正確さ、最終結論の妥当性、手法の妥当性など）から構成されており、合計10点になるよう配点されています。モデルの解答はこの各項目に照らして点数化され、**7点以上**を獲得した場合にその問題を「正解」とみなすルールです [13] [12]。このようなルーブリック採点により、**部分的な正当性や途中経過の正しさ**も評価でき、モデルの失敗要因を細かく分析することが可能になります。採点は本来専門家が行うのが理想ですが、大規模モデル評価では非現実的なため、OpenAIは**GPT-5モデルを用いた自動グレーダー**を開発し、このルーブリックに従ってモデル解答を評価しています。ルーブリック自体も**モデルで評価しやすいよう校正**されており、難易度や正誤の基準がぶれないよう検証パイプラインを構築したとされています。

モデルの評価結果

主要モデルのスコア比較: 初回のベンチマーク結果では、OpenAIの最新モデル**GPT-5.2**が**Olympiadで77%正答率、Researchで25%得点**と最も高いスコアを記録しました [14]。以下、Googleの**Gemini 3 Pro**やAnthropicの**Claude Opus 4.5**など他の最先端モデルが続きますが、OlympiadとResearchで性能に大きな差が見られます [15]。主要モデルのトラック別スコアを表にまとめます [16]:

モデル名	Olympiad正答率	Researchスコア (10点満点換算)
GPT-5.2 (OpenAI)	**77.1%** [17]	**25.3%** [17] (≈2.53/10)
Gemini 3 Pro (Google)	76.1% [18]	12.4% [18] (≈1.24/10)
Claude Opus 4.5 (Anthropic)	71.4% [18]	17.5% [18] (≈1.75/10)
Grok 4 (xAI)	66.2% [19]	15.9% [20] (≈1.59/10)

表中の値はAI CERTsによる報道 [16] に基づきます。Researchスコアは10点満点中の得点をパーセンテージに換算した値です。

この比較から、**競技系の閉じた問題**であるOlympiadでは各モデルとも比較的高スコアを収めているのに対し、**オープンエンドのResearch課題**では軒並み大きくスコアが低下する傾向が読み取れます。特にGPT-5.2はOlympiadでトップとはいえ2位のGeminiと僅差ですが、Researchでは**2倍近い得点差**で他モデルを引き離しています [21]。逆に言えば、Gemini 3 ProはOlympiadではGPT-5.2に肉薄しながら、Researchではその半分程度の正解率（約12%）しか達成できませんでした [18] [21]。Claude Opus 4.5はOlympiadは76%に届かないもののResearchでは17.5%と、Geminiよりは高くGPT-5.2に次ぐ健闘を見せています [18] [22]。一方、OpenAIの前世代モデルにあたる**GPT-4o**や小型モデル(OpenAI o3等)は、Olympiadでも上位より10ポイント以上低い**60%前後**に留まり、Researchでは一桁台%にとどまると推測されます [23]（GPT-4oの具体的スコアは公表されていませんが「比較的低調」と報告されています [24]）。

分野別の性能差: FrontierScienceは物理・化学・生物の各分野に問題がまたがっていますが、モデルの得意不得意には分野差も見られます。OpenAIの報告によれば、GPT-5.2は**物理分野の問題で最高性能**を示し、Olympiad物理問題では**81%**の正答率に達しました [22]。これは化学・生物分野での正答率より高く、GPT-5.2でも物理が比較的得意であることを示唆します [22]。しかし同じ物理でもResearchトラックの自由回答問題になると**28%程度**まで正答率が落ち込み [22]、化学についても同様の傾向（同程度の低スコア）が報告されています [22]。生物について明示的な数値はありませんが、Olympiadでは物理よりやや低め、Researchでは20%前後と他分野同様に苦戦したと推測されます [22]。このように**物理≫化学≧生物（Olympiad）**という強さ順が見られる一方、Researchでは**全ての分野で20%台以下**に留まり、**オープンな研究課題に対する脆弱性**が共通して露呈しています。

高推論モードとトークン数の影響: FrontierScienceの評価では、各モデルに思考ステップ（チェイン・オブ・ソート）の長さを調節した**推論モード**を設定しています。多くのモデルが「high」モード（高い推論努力）で評価されましたが、GPT-5.2のみ**「xhigh」モード**（極めて高い推論努力）で追加の思考時間を与えられました [25]。その結果、GPT-5.2は推論時間を増やすことで**精度が向上**しています。例えばGPT-5.2は推論ステップを増やすことで、Olympiadでの正答率が**約67.5%から77.1%**に、Researchでのスコアが**18%から25%**に向上したと報告されています [26]。一方、より小型のモデルでは推論を長くしすぎるとかえって成績が落ちるケースもあったようです [27]。推論時間を延ばすことで**深い探索や計算**が可能になる一方、**計算資源や応答時間のコスト**も増大します [28]。OpenAIの分析でも、GPT-5.2のような大型モデルでは追加の思考ステップ投入で確かに精度向上が得られるものの、その分**処理時間・トークン消費が増えるトレードオフ**が指摘されています [29]。実運用では精度向上とコストのバランスを考慮する必要があるでしょう。

作問・検証プロセス

問題作成者のプロフィール: FrontierScienceの問題は**各分野のトップクラスの人材**が執筆しています [30]。Olympiadトラックの100問は**国際科学オリンピックのメダリストや各国代表コーチ**計42名が共同で作成しており、彼らは累計で109個ものオリンピックメダルを獲得した経歴を持ちます [31]。Chemistry/Physics/Biologyの各オリンピック大会の金メダリストが問題を執筆したことで、問題の質・難易度は**実際の国際大会問題に匹敵するかそれ以上**となっています [32] [33]。Researchトラックの60問は**現役の博士号取得者（博士**

課程候補・ポスドク・教授）45名が執筆し、それぞれ量子電磁力学、有機合成化学、進化生物学など**専門性の高い分野**をカバーしています [34]。要するに、**五輪メダリスト級の学生**と**PhD研究者**という両者の専門知識を結集して問題が作られているのです。

問題内容のオリジナリティと難易度: すべての問題は**完全オリジナル**であり、既存の教科書や過去問からの使い回しではありません [35]。作問チームには「**難しく、独創的で、かつ意味のある問題**」を作ることが求められました [35]。Olympiad問題は各国大会・国際大会の最高レベルに肩を並べる難問揃い [36]、Research問題は博士課程の研究でも遭遇しうる難問を再現したものになっています [37]。実際、Researchトラックのいくつかの問題は専門家によれば「**計算機シミュレーションを回せば数日かかる**」ような課題や、「**人間の研究者でも数週間要した**」解析問題も含まれていると報告されています [38] [39]。これらはまさに**一筋縄ではいかない難易度**で、既存モデルにとって手強い挑戦となるよう意図されています [40]。

検証方法と質保証: 問題の品質を担保するために、FrontierScienceの各タスクは**「作成→相互レビュー→模範解答作成→修正」**という**4段階の開発プロセス**を経ています [41]。まず作問者が問題と解答草案を作成し、それを**独立した別の専門家**がレビューします [41]。レビューでは問題が**事実に基づいているか、客観的な正解基準があるか**（gradable）、**あいまいでなく明確か**（objective）、そして**十分に難しいか**などの観点でチェックされます [42] [41]。レビュー結果を受けて作問者は問題文や解答の**修正・洗練**を行います。さらに必要に応じて第三者の専門家による**解答の解決(Resolution)**を試み、解けない・不備がある場合は問題を改良するステップも踏まれています [41]。このように**複数の専門家の目**を通して問題文・解答・採点基準が磨かれ、最終的にFrontierScienceの正式問題として採択・公開されます。

モデルによるプレテストと問題選別: 問題開発段階では、OpenAI内部の言語モデルを使って試験的に問題を解かせる**プレテスト**も行われました。その結果、**あまりにも易しくモデルが解けてしまった問題はリジェクト（除外）**することで、ベンチマークとして十分困難で有意義な問題のみを残す工夫がされています [43]。例えばChatGPT等がスラスラ正答してしまった問題は棄却し、モデルが解けなかったもの・間違えたものを優先的に残すことで、**モデルに不利（厳しめ）な問題セット**になるよう調整されています [43]。このためFrontierScienceのスコアは、特にOpenAIの自社モデルに対しては**控えめに出るバイアス**があるとも言われます [44]（性能を正しく測るための意図的な措置です）。最終的に残った**金の問題セット**（Olympiad 100問＋Research 60問）はオープンソースとして公開され、他の研究者も利用できるようになっています [45]。公開されなかった残りの数百問は、将来のデータ汚染チェック（モデルが学習データとして見てしまったかの検証）や更なる評価に取っておかれています [45]。

課題と今後の展望

自由なアイディエーションの難しさ: FrontierScienceはあくまで**専門家が解答を知っている問題**を出題し、その範囲でモデルの推論力を測るベンチマークです。そのため、**真に新規な仮説をモデルが自由に発想して検証する能力**までは評価できません [46]。OpenAIの研究者も「このベンチマークは科学者の日常業務のすべてをカバーするものではない」と認めており [46]、例えば**モデルが自ら未踏の問題を設定し解決する**創造的プロセスや、**未知の発見に至るまでの発想力**は現状評価の範囲外です [47]。FrontierScienceはあくまで**既存の難問に対する推論力**を測る指標であり、本当の意味で科学研究を主導できるAIかどうかは**最終的にはそのAIが生み出す新発見**によって評価されるべきだと述べられています [48]。したがって「**自由なアイデア創出**」という点では依然課題が残っており、この能力を評価・育成するにはさらなる工夫が必要でしょう。

ルーブリック評価の曖昧性: Researchトラックでは10点ルーブリックで評価するとはいえ、**客観的に完全に割り切れる採点**が難しい場合もあります。複数項目に分けているとはいえ長文回答の出来を点数化する作業には主観が入りうるため、OpenAI自身も「**複数要素からなる長文タスクの採点は、最終答だけを見るより主観性が増す**」と限界を認めています [49]。また実際の採点を担うのが**GPT-5ベースのモデルグレーダー**である点も、評価の信頼性に影響しえます。モデルによる採点は大規模には便利ですが、**評価基準がプロンプト経由で漏洩するとモデルが甘く採点してしまう**などのリスクも指摘されています [50]。こうしたルーブリック評価の

ばらつきやモデル採点の偏りを抑えるため、OpenAIは人手校正や検証パイプラインを導入したとしていますが、**評価の客観性・再現性**をいかに保つかは継続的な課題です。

実世界の研究との乖離:　もう一つの制約は、このベンチマークが**テキスト問題のみ**で構成されていることです。実際の科学研究では、論文読解や数式推論だけでなく**実験の計画・操作**や**実データ（画像・動画・時系列）の解析**といった**マルチモーダルな能力**が必要です。FrontierScienceではそうした**実験的・視覚的モダリティ**は一切問われません [51]。言い換えれば、「テキスト上で完結する高度な問題解決能力」は測れるものの、実験ノウハウや画像分析能力など**現実の研究活動の重要な部分**が評価範囲外となっています [51]。さらに、問題数がOlympiad 100問・Research 60問と限られているため**サンプル数が少なく**、性能が拮抗するモデル同士の優劣を統計的に有意に判断しづらいという限界もあります [52]。加えて**人間のベースライン**（例えば博士課程の学生や専門家がこの問題セットを解いたら何点か）も現在のところ未提示であり、モデルの絶対的な位置づけ（人間に比べてどの程度できるのか）が不明確です [52]。専門家ですら解けない問題も含まれると想定されますが、それを確認するデータがないため、**「25%」というGPT-5.2のResearch成績が人間相当に近いのか、それともまだ遠いのか**判断しにくい状況です [52]。これらは今後解決すべき課題と言えます。

今後の改善・展望:　OpenAIはFrontierScienceについて、「**モデルの科学推論力を測る一つのツールに過ぎない**」としており、将来的には**領域を拡張**したり**実世界での評価**（実際の研究タスクでどれだけAIが支援できるか等）と組み合わせたりする計画を示唆しています [53]。具体的には、**他の科学分野**（例：地学、工学、医学など）への拡大や、**画像・実験データの解析を含むマルチモーダルな問題**の導入も検討されるでしょう [54] [55]。実際、今後の版では「実験室から得られる画像データやシミュレーション結果を解釈する問題、物理世界での実験計画を立てさせる問題」を取り入れることや、必要に応じて**数式処理ツール**や**コード実行**といった機能をモデルに使わせる方向も話題に上っています [55]。また、現在欠如している**人間のベースライン**についても、追って専門家チームにFrontierScience問題を解かせてデータ取得する可能性があります。今後モデルが向上すればベンチマーク問題自体も更新していく方針で、コミュニティによる評価コードの公開や人手での監査も計画されています [56]。最終的に重要なのは、AIが科学に実際どんな新発見をもたらすかという点ですが、その**早期の指標・北極星**としてFrontierScienceが機能し、モデルの弱点を洗い出して改善に繋げることが期待されています [57] [58]。

他のベンチマークとの比較

GPQAとの比較: GPQA（Google-Proof Question Answering）は2023年に公開された大学院レベルの科学QAベンチマークで、PhD研究者が作成した**選択肢問題**から構成されます [59]。公開当時GPT-4は**39%**の正解率で専門家（約70%）に遠く及びませんでしたが、**2年後にはGPT-5.2が92%**正解するまでに至りました [59]。この劇的な向上は、既存ベンチマークがモデルの進歩ですぐ**「飽和」**してしまう問題を示しています [60]。実際、GPQAの問題は選択式ゆえ**Web検索で答えを探せないか**を重視した設計でしたが、最新モデルにはもはや十分な難易度とは言えなくなりました [59]。FrontierScienceはこのギャップを埋めるために企画されており、**多肢選択ではなく記述式**であること、そして**科学分野に特化**している点でGPQAより踏み込んだ評価を可能にしています [61]。要するに、**GPQAが「検索に頼らず答えを知っているか」を問う試験**だとすれば、FrontierScienceは**「未知の難問を論理的に解決できるか」を問う試験**と言え、その貢献度は次世代モデルの**科学推論力の伸びを測る指標**となった点にあります。

MMLUとの比較: MMLU（Massive Multitask Language Understanding）は12500問以上からなる大規模ベンチマークで、科学も含め広範な分野の常識・知識問題を多肢選択形式で網羅しています。GPT-4はMMLUで人間レベル（約90%）に迫る高スコアを出しており、**既に多くの領域で頭打ち状態**です。しかしMMLUの科学分野の多くは基礎知識や大学学部レベル問題で、**深い推論を要する設問は少ない**のが実情です。FrontierScienceは**完全に科学専門に特化**し、しかも**高難度の推論問題のみ**で構成されているため、MMLUとは難易度・性格が大きく異なります [61]。またMMLUがマーク式で正解を選ぶだけなのに対し、FrontierScienceは**自ら過程を書かせ評価する**点でより高度な能力を測定できます。要するに、FrontierScienceは「GPT-4でも解けない科学問題」を新たに作り直したベンチマークであり、汎用知識テストであるMMLUに比べ**科学研究に直結する推論力**を問う点で重要な貢献を果たしています。

OlympiadBenchとの比較: OlympiadBenchは2024年に報告された**オリンピックレベルの問題集**で、国際大会や中国の高考などから**8476問**もの**数学・物理**の問題を収集した大規模データセットです [62]（問題文中に図表を含むマルチモーダル問題も多数あります [63]）。特徴として**バイリンガル（中英）対応**であることや、**証明問題**も含む点が挙げられ、AGIへの挑戦として提案されました [64]。OlympiadBenchの難易度は非常に高く、例えばGPT-4ビジョン（図を解釈できるGPT-4V）でも平均**18%**程度の正答率に留まり（物理分野ではわずか10.7%）と報告されています [65]。FrontierScienceとOlympiadBenchはいずれも「オリンピック級の難問」を扱う点で共通しますが、**問題ソースと範囲に違い**があります。OlympiadBenchは既存の競技試験問題を集めたものに対し、FrontierScienceは**新規に作成された問題**です [35]。またOlympiadBenchは主に**数学・物理**に偏っているのに対し、FrontierScienceは**物理・化学・生物**の3分野をカバーしています [4]。さらに最大の差異は**Researchトラックの有無**です。OlympiadBenchにも一部記述式の問題がありますが基本的に答えが決まった問題集であり、**研究プロジェクトのシミュレーション**のようなオープンエンド課題は含んでいません。一方FrontierScience-Researchは**Proof-of-concept的な研究課題**を提示し、モデルの思考プロセス全体を評価する点でユニークです [66]。したがってFrontierScienceは、OlympiadBenchがカバーしていない**化学・生物領域**や**研究的思考**を評価軸に組み込んだ点で貢献していると言えます。また問題数は少ないものの精選されているため、評価の深度という意味でも優れたベンチマークです。

CritPtとの比較: CritPtは2025年11月に有志の物理学者チームが発表した**博士課程レベルの物理研究ベンチマーク**です [67]。71個の未発表の研究課題（量子物理、宇宙物理、高エネ物理など11分野）から構成され、さらにそれらを解く上で通過すべき190個の「チェックポイント問題」を含んでいます [68]。目的は**AIが未解決の研究問題を一から解けるか**を試すことであり、問題は既存の教科書知識では太刀打ちできない**斬新な設定**ばかりです [69]。初期結果では最先端モデルも全く歯が立たず、GoogleのGemini3 Proプレビュー版でも**9.1%**正解、OpenAI GPT-5.1（highモード）でも**4.9%**正解に過ぎませんでした [70]。多くの課題で**モデルは答を出せず沈黙する**ケースもあったといいます [71]。このCritPtとFrontierScience-Researchは趣旨が似ており、ともに**AIの研究遂行能力の限界**を試すものですが、その**難易度設定と評価手法**に違いがあります。CritPtの課題は極めてオープンエンドで、解答も論文レベルの自由記述になるため**人間の専門家が採点**するしかありません。一方FrontierScienceは**各課題をもう少し細かいサブタスクに分解し、客観的な採点基準を用意する**ことで自動評価を試みています [72] [73]。難易度的にはCritPtが突出して高く、GPT-5.2でさえ0%に等しい成績だったのに対し、FrontierScience-ResearchではGPT-5.2が25%を獲得しています [15] [70]。これはFrontierScienceの課題が**自己完結型でヒントも含むサブプロブレム**であるのに対し、CritPtは**全く未知の問題をゼロから解かせる**という違いによるものです。したがってFrontierScienceはCritPtほど極端ではないものの、**より広い分野での研究課題**を体系的に評価できる点でAIの限界解明に貢献しています。現状トップモデルでも25%程度に留まることから、物理専用のCritPtだけでなく**マルチ分野版の研究ベンチマーク**としてFrontierScienceが位置づけられ、AI研究のボトルネックを浮き彫りにしています [74] [53]。

総括: 従来のベンチマーク（MMLUやGPQA）は**知識問題や定型問題**が多く、最先端モデルには容易すぎるか人間と差がつきにくい状態でした [75]。またOlympiadBenchやCritPtといった新ベンチマークは難易度は高いものの**特定分野だけ**だったり**評価に人的労力**が必要だったりする制約がありました。それに対しFrontierScienceは、**専門家チームによる新作問題**で難易度を引き上げつつ、**3分野にまたがる包括性**と**自動採点可能な枠組み**を両立させた点で大きな貢献をしています [76] [72]。特にResearchトラックの導入により、AIの**推論過程の部分的な正確さ**まで評価可能にしたことは他に類を見ない特徴です。もっとも、このベンチマークで測れるのは科学的思考力の一部であり、真に重要なのはAIが現実の科学に貢献して生み出す**「フロンティア（新発見）」**であることは言うまでもありません [48]。FrontierScienceはその前段階として、モデルの科学推論力をより厳密に追跡するための**新たな尺度**を提供した点で、既存ベンチマーク群の限界を一歩押し広げたと言えるでしょう [61] [1]。

参考文献: FrontierScience公式ブログ [59] [77]、OpenAI (2025)「Evaluating AI's ability to perform scientific research tasks」 [5] [12] [15] [51] ほか.

---

1 4 7 8 9 10 11 13 14 25 30 31 32 34 35 37 41 42 43 44 45 46 48 49 57 58 59 60 61 66 77 Evaluating AI's ability to perform scientific research tasks | OpenAI
https://openai.com/index/frontierscience/

2 3 5 6 12 15 33 36 47 51 53 54 74 75 OpenAI introduces FrontierScience to test AI's expert-level scientific reasoning across physics, chemistry, biology | Mint
https://www.livemint.com/technology/tech-news/openai-introduces-frontierscience-to-test-ai-s-expert-level-scientific-reasoning-across-physics-chemistry-biology-11765909741317.html

16 17 18 19 20 21 22 28 29 50 55 56 OpenAI Benchmark Shows Model Capability with 77% Olympiad Score - AI CERTs News
https://www.aicerts.ai/news/openai-benchmark-shows-model-capability-with-77-olympiad-score/

23 24 26 27 72 73 76 HyperAI
https://hyper.ai/en/news/47756

38 52 OpenAI Is Testing AI's Scientific Ambitions | TIME
https://time.com/7341081/openai-frontierscience-benchmark/

39 40 OpenAI Unleashes FrontierScience for AI-Fueled Scientific Reasoning | eWEEK
https://www.eweek.com/news/openai-frontierscience-launch/

62 63 64 65 GitHub - OpenBMB/OlympiadBench: [ACL 2024]Official GitHub repo for OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems.
https://github.com/OpenBMB/OlympiadBench

67 68 69 70 How Far is AI from the Nobel Prize? Top Models Fail in the CritPt Benchmark for Doctoral-Level Physics with Accuracy Below 10%
https://news.aibase.com/news/23016

71 Gemini 3 Pro and GPT-5 still fail at complex physics tasks designed ...
https://the-decoder.com/gemini-3-pro-and-gpt-5-still-fail-at-complex-physics-tasks-designed-for-real-scientific-research/