FrontierScience: AI科学的推論の新たな地平を測る

主要AIモデルの能力と限界を明らかにする新ベンチマークの詳細分析

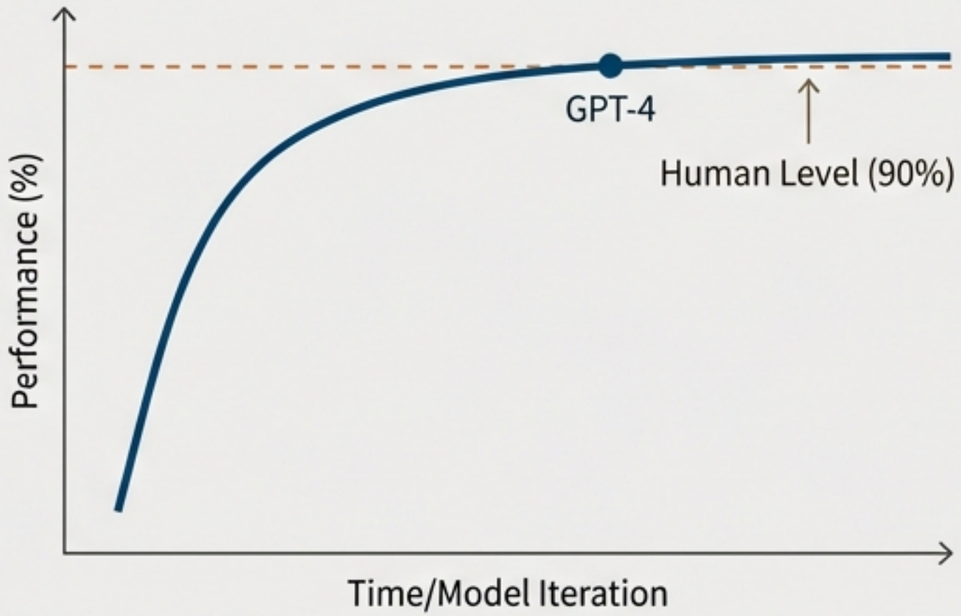
既存ベンチマークの「天井」：なぜ新たな指標が必要なのか？

トップクラスのAIモデルは旧世代のテストで最高性能に達しつつあり、
その真の推論能力の限界が見えにくくなっている。

MMLU（汎用知識）

GPT-4は人間レベルの90%に到達。

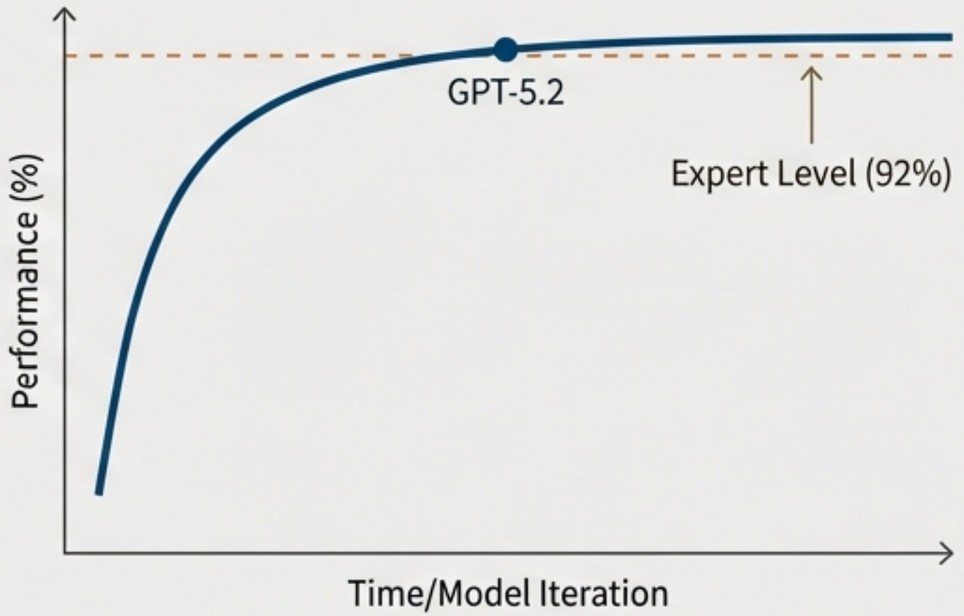
汎用知識の評価は「飽和」しつつある。



GPQA（検索困難な専門知識）

GPT-5.2は専門家レベルの92%を達成。

知識再現型の問題は克服されつつある。



知識の再現から、未知なる問題の解決へ

AI進化の次なるフロンティアは、既知の問いに答えることではなく、
科学者のように複雑で多段階の未知の問題に取り組むことである。

旧来のパラダイム

多肢選択式

既知の知識

正解の検索

新たなパラダイム

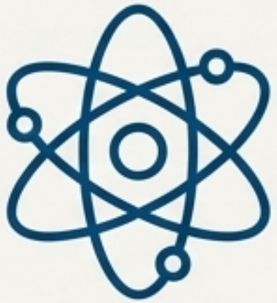
自由記述式

未知の課題

解決プロセスの構築

新たな指標：FrontierScience

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、全く新しいベンチマーク。



物理学

化学

生物学

このベンチマークは、2つのユニークな評価トラックを通じて、
AIの能力を多角的に測定する。

2つのトラック、2つのゴール：構造化された推論力と、オープンエンドな研究遂行能力

特徴（Feature）	Olympiad Track	Research Track
目的（Purpose）	競技レベルの科学的推論力	博士レベルの研究遂行能力
形式（Format）	短答式（数値、数式、専門用語）	自由記述式（論述、考察、計算）
問題数（Questions）	100問	60のサブタスク（研究課題）
評価（Evaluation）	自動採点（正誤判定）	ルーブリック採点（10点満点）

Researchトラックでは7点以上を「正解」とみなす。

世界トップクラスの専門家による作問：妥協のない品質と難易度

オリンピックメダリストによる設計

国際科学オリンピックのメダリストやコーチ42名が作成。累計メダル獲得数は109個。

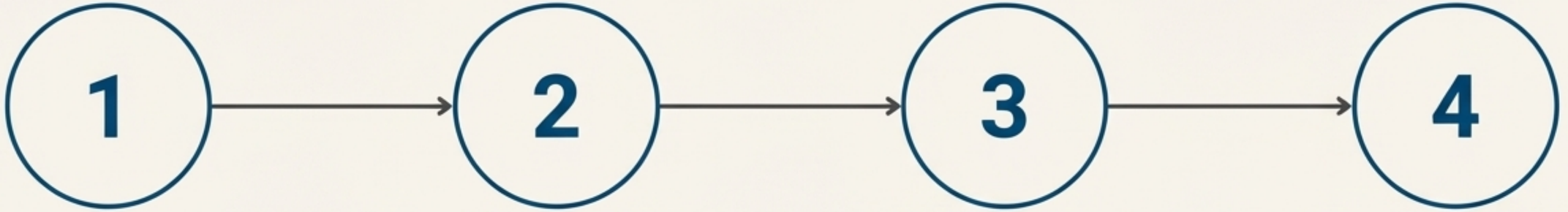
博士号取得者による設計

現役の博士号取得者（博士課程、ポスドク、ク、教授）45名が、実際の研究で直面するレベルの課題を作成。

すべての問題は完全オリジナル。既存の教科書や過去問からの流用は一切なし。

厳格な品質保証プロセス

**すべての問題は、品質、客観性、そして適切な難易度を保証するため、
多段階のレビュープロセスを経ている。**



[作成 (Creation)]

[相互レビュー (Peer Review)]

専門家が客観性、明確さ、
難易度をチェック

[模範解答作成 (Solution Modeling)]

第三者が解答を試み、
問題の不備を検証

[修正 & プレテスト (Revision & Pre-testing)]

モデルによる事前テストで
簡単すぎる問題を除外し、
意図的に「モデルに不利な」
セットに調整

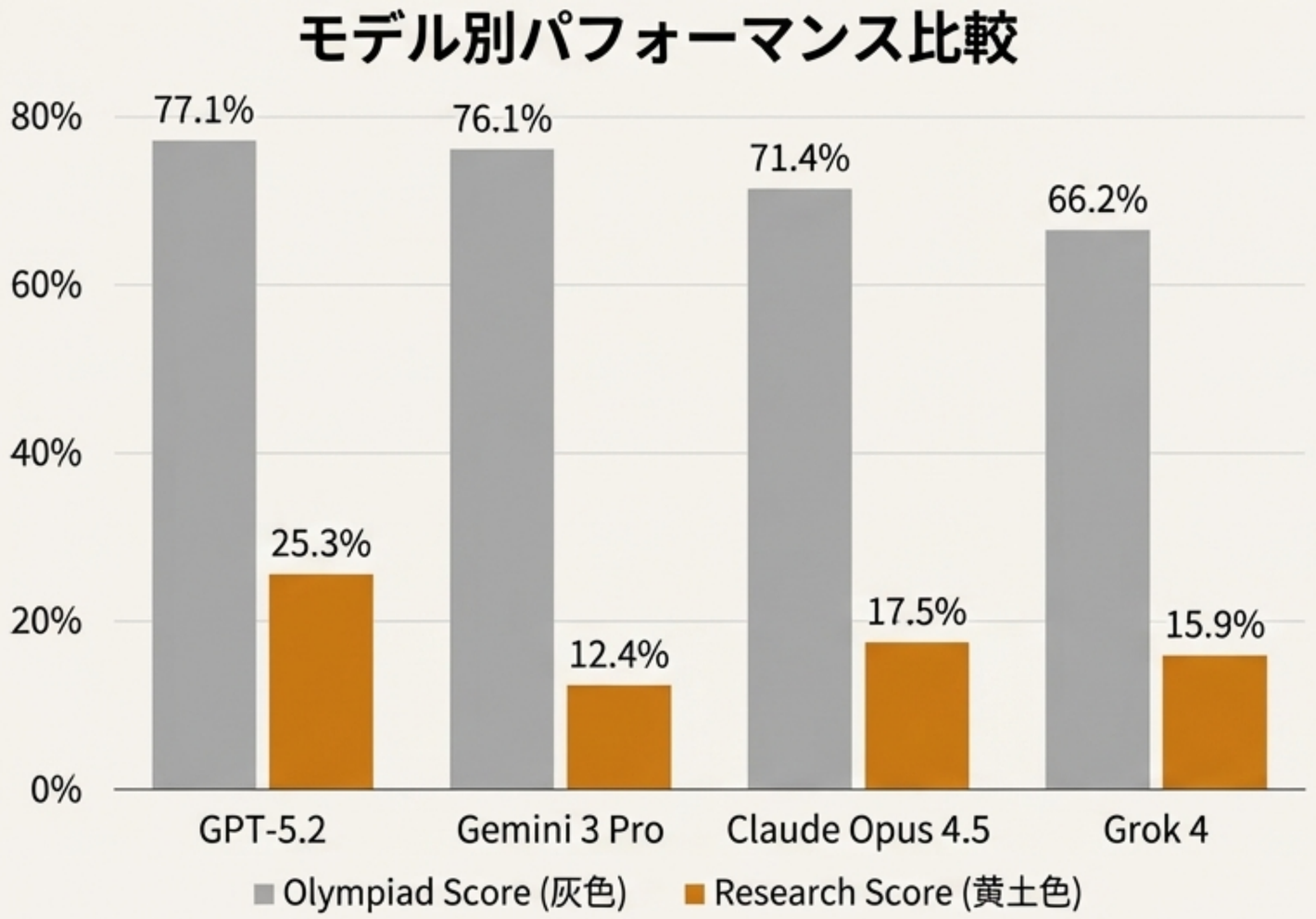
このプロセスにより、ベンチマークがAIの真のフロンティア能力を測定できることを保証する。

主要モデル評価結果：SOTAモデルのパフォーマンス

モデル名（Model）	Olympiad 正答率	Research スコア
GPT-5.2 (OpenAI)	77.1%	25.3%
Gemini 3 Pro (Google)	76.1%	12.4%
Claude Opus 4.5 (Anthropic)	71.4%	17.5%
Grok 4 (xAI)	66.2%	15.9%

※Researchスコアは10点満点中の得点をパーセンテージ換算した値（例: 25.3% ≒ 2.53/10点）。スコアはAI CERTsの報道に基づく。

「閉じた問題」と「開かれた研究」の間に横たわる深い溝



主要な洞察

- トップモデルでも、オープンエンドなResearch課題ではスコアが大幅に低下。
- GPT-5.2はResearchで他を圧倒するも、自身のOlympiadスコアの約 1/3 に留まる。
- これは、構造化された問題解決能力と、真の研究遂行能力の間に大きなギャップが存在することを示唆する。

性能を左右する要因：分野と「思考時間」の影響

分野別の性能差

物理分野で最高性能を示す傾向。

Olympiad Track (GPT-5.2): 物理 物理 (**81%**) > 化学 ≧ 生物

Research Track (GPT-5.2): 全分野で苦戦。物理ですら**28%**に低下。オープンなL研究課題への脆弱性は共通。

高推論モードの影響

「思考時間」の増加が性能を向上させる。

GPT-5.2は「xhigh」モードで性能が大幅向上。

- Olympiad: 67.5% → **77.1%**
- Research: 18% → **25.3%**

現状のトップモデルでは、精度と計算コスト（時間・トークン消費）の間に明確なトレードオフが存在する。

ベンチマークの生態系におけるFrontierScienceの位置づけ

ベンチマーク	主な目的	FrontierScienceとの主な違い
MMLU	広範な一般知識（多肢選択式）	科学専門、超高難易度、**自由記述式**によるプロセス評価
GPQA	検索困難な専門知識（多肢選択式）	**自由記述式**、よりオープンエンドな問題解決能力を測定
OlympiadBench	既存の五輪問題（数学・物理）	**完全新作**問題、化学・生物を含む、**Researchトラック**の存在
CritPt	博士レベルの未解決物理研究課題	より体系化されたサブタスク、**自動評価**の枠組み、物理以外の分野もカバー

FrontierScienceは、高難易度、専門性、自由記述、研究タスクのシミュレーションを組み合わせることで、既存ベンチマークが測定できなかった「**科学的推論プロセス**」の評価を可能にした。

現状の課題と限界

テキスト以上の能力は評価外

実験データや画像など、マルチモーダルな情報の解析能力は測定できない。

ルーブリック評価の主観性

Researchトラックの自由記述式評価には、特にモデルベースの採点者を利用する場合、主観が入る余地がある。

人間ベースラインの欠如

専門家が解いた場合のスコアが未提示のため、モデルのスコアの絶対的な位置づけが困難。

今後の展望と拡張計画

FrontierScienceは、AIの進化に合わせて継続的に発展していく。

- **他分野への拡大 (Expansion to Other Fields):** 工学、医学など、他の科学分野への展開。
- **マルチモーダル問題の導入 (Introduction of Multimodal Problems):** 画像やシミュレーションデータの解釈を含む課題の追加。
- **ツール利用能力の評価 (Evaluation of Tool Use):** コード実行や数式処理ツールなど、外部ツールを活用する能力の測定。
- **人間ベースラインの確立 (Establishment of Human Baselines):** 専門家による解答データを用いて、モデルの性能を客観的に位置づける。

科学的AGI開発の新たな羅針盤

FrontierScienceは、AIが単に知識を再現するだけでなく、真の科学的発見を生み出すための「フロンティア」への道のりを測る、不可欠なマイルストーンである。

1. **ギャップの可視化 (Revealed the Gap):** 構造化された推論と、オープンエンドな研究遂行能力との間の決定的な差を明確に示した。
2. **基準の引き上げ (Raised the Bar):** 将来のAI開発を導くための、より高く、より実践的な評価基準を提供した。

ご清聴ありがとうございました

Q&A

詳細はこちらの資料をご覧ください：
openai.com/index/frontierscience