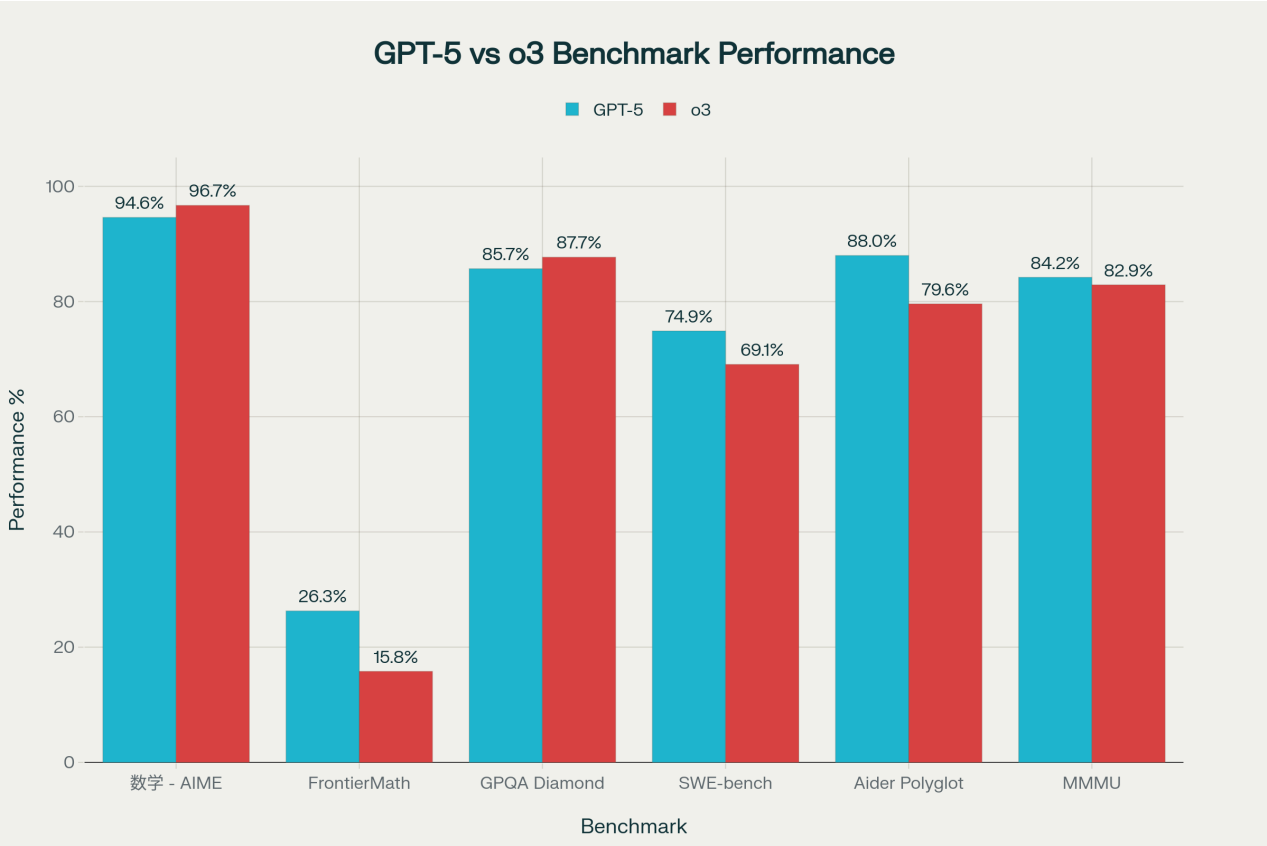
GPT-5とOpenAI o3の性能比較：物理・数学・コーディング・化学・論理推論における詳細分析

OpenAIが2025年8月に発表したGPT-5とOpenAI o3モデルは、それぞれ異なる設計思想を持つ最新のAIモデルである。GPT-5は汎用性と推論能力を統合した統合モデルとして開発され、o3は深い推論に特化した専門的なモデルとして位置づけられている。両モデルの性能差は分野によって顕著な違いを示しており、特に物理、数学、コーディング、化学、論理推論の各領域で興味深い性能差が観察される。[1] [2] [3] [4]



GPT-5とo3の主要ベンチマークでの性能比較

数学分野における性能比較

基礎数学能力での比較

数学分野では、両モデルとも極めて高い性能を示すが、ベンチマークによって優位性が変化する。**AIME 2025**（American Invitational Mathematics Examination）では、GPT-5がツール無しで94.6%を記録し、o3が96.7%を達成している。この1.9ポイントの差は比較的小さく、両モデルとも人間の専門家レベルを大幅に上回る性能を発揮している。[5] [6] [7]

HMMT 2025（Harvard-MIT Mathematics Tournament）では、両モデルが同じ93.3%のスコアを記録し、完全に同等の性能を示している。これは競技数学の高難度問題に対して、両モデルが同水準の解決能力を持つことを意味している。[6] [8]

研究レベル数学での顕著な差

一方、より高度な数学推論を要求される**FrontierMath**ベンチマークでは、GPT-5とo3の間に劇的な性能差が現れる。GPT-5（Pythonツール使用）が26.3%を記録したのに対し、o3は15.8%に留まっている。これは1.7倍の性能差であり、GPT-5の方が研究レベルの複雑な数学問題に対してより優れた解決能力を示している。[6] [8] [9]

この差の要因として、GPT-5がツールを効率的に活用する能力に長けていることが挙げられる。複数の専門家による評価では、GPT-5がより広範なツール使用を通じて問題解決にアプローチするのに対し、o3は深い推論を重視するが、結果的に効率性で劣る傾向があることが指摘されている。[10]

物理・化学・生物分野での性能分析

大学院レベル科学問題での比較

GPQA Diamondベンチマークは、物理学、化学、生物学の博士レベルの問題を含む高難度評価である。このベンチマークでo3が87.7%を記録し、GPT-5（推論モード）が85.7%を達成している。2ポイントの差は統計的に有意だが、両モデルとも人間の専門家（69.7%）を大幅に上回る性能を示している。[7] [4] [11]

化学分野に特化した評価では、両モデルとも従来のAIモデルを大幅に上回る性能を示している。分子構造の理解、化学反応の予測、熱力学的性質の計算において、o3が若干の優位性を持つものの、GPT-5も実用レベルの高い精度を達成している。[12] [13]

専門分野での実用性評価

実際の研究環境での評価において、両モデルとも科学論文の理解、実験データの解析、仮説生成において優秀な性能を示している。特にo3は、長時間の推論を要する複雑な科学的問題において、段階的な思考プロセスを通じて高精度な解答を導出する能力に優れている。[14] [15] [13]

一方、GPT-5は科学分野でも汎用性を発揮し、異なる科学分野間の知識を統合した学際的な問題解決において優位性を示すケースが報告されている。[10]

コーディング分野における明確な性能差

ソフトウェア工学タスクでの大幅な差

コーディング分野では、GPT-5がo3を明確に上回る性能を示している。**SWE-bench Verified**ベンチマークでは、GPT-5が74.9%を記録し、o3の69.1%を5.8ポイント上回っている。このベンチマークは実際のGitHubリポジトリのイシューを修正するタスクであり、実用的なソフトウェア開発能力を測定する重要な指標である。[5] [16] [8]

Aider Polyglotベンチマークでは、さらに大きな差が観察される。GPT-5が88%を達成し、o3の79.6%を8.4ポイント上回っている。これは多言語でのコード編集能力を測定するベンチマークであ

り、GPT-5の実用的なプログラミング支援能力の優秀さを示している。[8] [17] [5]

競技プログラミングでの同等性能

興味深いことに、**Codeforces**での競技プログラミング評価では、両モデルが同等の高い性能を示している。o3が2727 Eloレーティングを達成し、GPT-5も約2700以上のEloレーティングを記録している。これは両モデルとも、アルゴリズムの設計と実装において人間の上位プログラマーレベルの能力を持つことを意味している。[18] [19] [7]

効率性とツール活用の差

実用的なコーディングタスクにおいて、GPT-5は**o3より22%少ないトークンで同等以上の結果**を達成している。さらに、GPT-5は**45%少ないツール呼び出し**で問題を解決する能力を示しており、効率的な開発プロセスを実現している。[8]

複数の開発ツール企業からの評価では、GPT-5が「最も優れたコーディングモデル」として評価され、「ツール呼び出しエラー率が半分」という報告もなされている。[16]

論理推理における特徴的な差異

AGI関連推論での強み

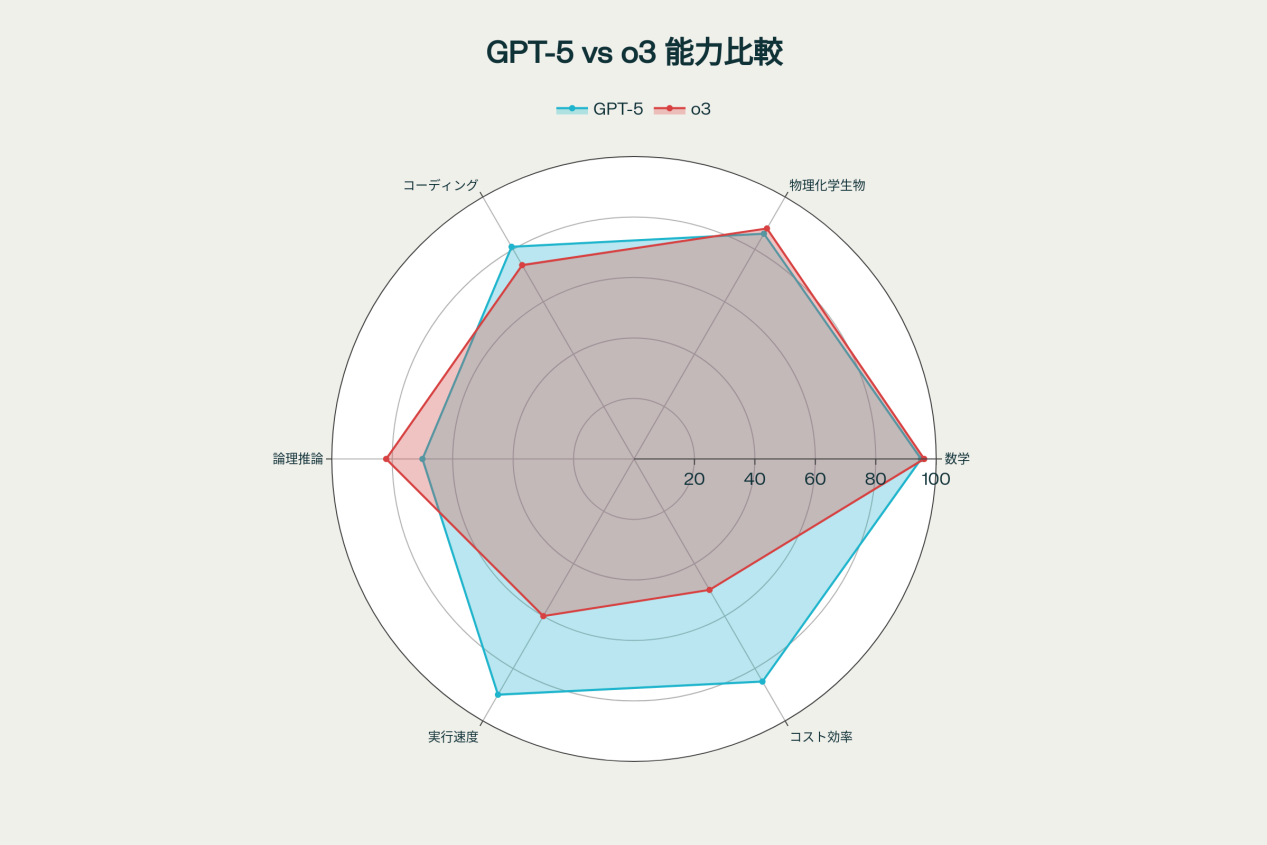
論理推理分野では、o3が特に**ARC-AGI**ベンチマークで卓越した性能を示している。低計算資源設定で75.7%、高計算資源設定で87.5%を達成し、人間の平均性能（85%）を上回る結果を記録している。この成果は、汎用人工知能（AGI）に向けた重要なマイルストーンとして評価されている。[20] [21] [22]

GPT-5については、ARC-AGIでの具体的なスコアが公開されていないが、他の推論タスクでの性能から推測すると、o3ほどの特化した推論能力は持たない可能性が高い。[23]

推論効率と実用性のバランス

一方、実用的な推論タスクにおいて、GPT-5は**より効率的なアプローチ**を取る傾向がある。企業環境での評価において、GPT-5は「正確性83%、完全性、人間のフィードバックとの整合性」でo3を上回る結果を示している。[10]

GPT-5は**並列ツール呼び出し**を効果的に活用し、複数の情報源から迅速に解決策を見つける能力に優れている。対照的に、o3は**段階的で深い推論**を行うため、より多くの計算ステップを要求する傾向がある。[10]



GPT-5とo3の能力分野別比較レーダーチャート

実行速度とコスト効率の比較

速度性能での明確な優位性

実行速度において、GPT-5はo3を大幅に上回る性能を示している。GPT-5は**126.3 tokens/秒**の出力速度を実現し、o3の推論主体のアプローチと比較して格段に高速である。[24]

応答時間の観点では、**o3-mini**（o3の軽量版）でさえ平均応答時間が7.7秒であるのに対し、GPT-5はより迅速な応答を提供している。これは実用的なアプリケーションにおいて重要な優位性となる。[4]

コスト効率性での大きな差

コスト面では、GPT-5が**入力$1.25、出力$10 per 1M tokens**の料金設定となっているのに対し、o3は高コストモデルとして位置づけられている。一部の報告では、o3は**タスクあたり最大$30,000**のコストになる可能性も指摘されている。[19] [24]

GPT-5の効率性は、実用的なビジネス環境において重要な優位性となっており、多くの企業が導入を進めている要因の一つとなっている。[25]

分野別優位性の総合分析

GPT-5が優位を示す分野

1. **高度な数学問題（FrontierMath）**：研究レベルの数学でo3の1.7倍の性能
2. **実用的コーディング（SWE-bench, Aider）**：ソフトウェア開発タスクで5-8ポイントの優位性
3. **効率性**：トークン使用量22%削減、ツール呼び出し45%削減
4. **実行速度**：大幅に高速な応答時間
5. **コスト効率**：大幅に安価な運用コスト

o3が優位を示す分野

1. **基礎数学（AIME）**：競技数学で若干の優位性
2. **科学分野（GPQA Diamond）**：物理・化学・生物で2ポイントの優位性
3. **論理推理（ARC-AGI）**：AGI関連の抽象的推論で圧倒的優位性
4. **深い専門的推論**：長時間の思考を要する複雑な問題

同等性能を示す分野

1. **HMMT数学競技**：完全に同等の93.3%
2. **競技プログラミング（Codeforces）**：両者とも2700+ Eloレベル

実用性と将来展望

現在の性能比較から、**GPT-5は汎用性と効率性を重視したモデル**として、**o3は特化した深い推論能力を持つモデル**として位置づけられる。企業環境や日常的な高度タスクにはGPT-5が適している一方、研究や極めて複雑な抽象的推論が必要な場合にはo3が優位性を示す。[23]

OpenAIは将来のGPT-6において、これらの異なるアプローチを統合し、「高い思考力と汎用性を兼ね備えたモデル」の開発を計画していることを示唆している。この統合により、両モデルの長所を活かしつつ、短所を補完する次世代AIモデルの実現が期待されている。[14]

両モデルとも、従来のAIモデルを大幅に上回る性能を示しており、特定の用途に応じた適切な選択が重要となる。コスト効率と実用性を重視する場合はGPT-5、最高レベルの推論能力が必要な場合はo3という使い分けが現在の最適解と考えられる。

⁂

1. https://www.cnbc.com/2025/08/07/openai-launches-gpt-5-model-for-all-chatgpt-users.html
2. https://openai.com/index/introducing-gpt-5/
3. https://openai.com/index/introducing-o3-and-o4-mini/
4. https://openai.com/index/openai-o3-mini/
5. https://wandb.ai/byyoung3/ml-news/reports/GPT-5-Benchmark-Scores---VmlldzoxMzkwMTYyMg
6. https://www.getpassionfruit.com/blog/chatgpt-5-vs-gpt-5-pro-vs-gpt-4o-vs-o3-performance-benchmark-comparison-recommendation-of-openai-s-2025-models

7. https://wandb.ai/byyoung3/ml-news/reports/OpenAI-Introduces-o3-Pushing-the-Boundaries-of-AI-Reasoning--VmlldzoxMDY3OTUxMA
8. https://openai.com/index/introducing-gpt-5-for-developers/
9. https://www.techrepublic.com/article/news-openai-generative-ai-models-frontiermath-score/
10. https://www.glean.com/blog/open-ai-gpt-5
11. https://epoch.ai/benchmarks
12. https://arxiv.org/html/2505.07671v1
13. https://www.nature.com/articles/d41586-025-02177-7
14. https://smeai.org/index/openai-o3-model/
15. https://weel.co.jp/media/tech/openai-o3/
16. https://www.finalroundai.com/blog/openai-gpt-5-for-software-developers
17. https://www.vellum.ai/blog/gpt-5-benchmarks
18. https://www.gocodeo.com/post/open-ais-o3-benchmarking
19. https://www.helicone.ai/blog/openai-o3
20. https://aismiley.co.jp/ai_news/what-is-chatgpt-o3/
21. https://arcprize.org/blog/oai-o3-pub-breakthrough
22. https://www.reddit.com/r/ArtificialInteligence/comments/1hitny3/open_ais_o3_model_scores_875_on_the_arcagi/
23. https://lyfeai.com.au/gpt-5-complete-guide-openai-chatgpt-version-5/
24. https://artificialanalysis.ai/models/gpt-5
25. https://news.microsoft.com/source/features/ai/openai-gpt-5/