GPT-5 vs o3 徹底比較：数学・コーディング・推論から運用までの実力差

Genspark

Aug 08, 2025

比較概要（結論）

- 総合力：最新ベンチマークでは GPT-5 が数学・推論・実務的コーディングで o3 を上回る指標が多く、特に事実性（低幻覚）と効率性（少ない出力トークン・ツール呼び出し）で優位です。一方で一部の実運用型課題（例：ウェブナビゲーション）では o3 が僅差で勝つケースも報告されています。OpenAI1 OpenAI2 TechCrunch3
- 分野別の要点
  - 数学：GPT-5 は AIME 2025 で SOTA（ツール無し 94.6%、Python 使用で 100%）。o3 は AIME 2025 で高得点（88.9%報告、ツール使用では 98.4%/consensus@8=100%）。難問域含め GPT-5 が一段上の到達度を示す構図です。OpenAI1 VentureBeat4 OpenAI5 DataCamp6
  - コーディング：実務寄り指標の SWE-bench Verified で GPT-5 が 74.9%、o3 が 69.1%。編集系（Aider polyglot）でも GPT-5 が 88%と強く、総じて GPT-5 が優勢。ただし o3 も競プロ Elo や各種編集系で強力です。OpenAI2 DataCamp6
  - 論理・推論：大学院レベルの GPQA で GPT-5 が 88.4%（extended reasoning）。o3 は ARC-AGI（高計算）で 88%と強力。総合推論では GPT-5 が低幻覚・低欺瞞で安定性に優れます。OpenAI1 VentureBeat4 DataCamp6
  - 物理・化学・科学：o3 系は STEM 最適化（o3-mini 含む）で博士レベル科学問題に強み。GPT-5 は医療分野評価（HealthBench Hard 等）や MMMU で高スコアを示し、科学的読解・視覚理解も伸長。定量的には課題依存ですが、最新総合では GPT-5 優勢の傾向。OpenAI7 OpenAI1
  - 幻覚・効率：GPT-5 は o3 比で事実誤りが大幅減（~80%減）、出力トークン 22%減・ツール呼び出し 45%減。医療精度や一般プロンプトでの幻覚率低下も顕著。一方、ウェブナビの一部では o3 が僅差で上回る結果も。用途で使い分けが有効です。OpenAI2 TechCrunch3

主要ベンチマークの対比（抜粋）

- 数学（AIME）
  - GPT-5：AIME 2025 ツール無し 94.6%、Python 使用で 100%（Pro）OpenAI1 VentureBeat4
  - o3：AIME 2025 88.9%（報告）、Python 使用で pass@1 98.4%／consensus@8 100% DataCamp6 OpenAI5
- コーディング
  - SWE-bench Verified：GPT-5 74.9%、o3 69.1% OpenAI2 DataCamp6
  - Aider polyglot（コード編集）：GPT-5 88%（o3 は o1 を上回るが公表値は記事内に具体数字なし）OpenAI2 DataCamp6
  - 競プロ（Codeforces Elo）：o3 2706（o1 1891 から大幅向上）DataCamp6
- 推論
  - GPQA（院生レベル）：GPT-5 88.4%（Pro, no tools）OpenAI1 VentureBeat4
  - ARC-AGI：o3 76%（低計算）/ 88%（高計算）DataCamp6
- 科学・視覚理解
  - MMMU：GPT-5 84.2%、o3 は SOTA 達成の主張（詳細表は o3 告知側）OpenAI1 OpenAI5
  - 医療（HealthBench Hard など）：GPT-5 が幻覚率 1.6%（GPT-4o 12.9%、o3 15.8%と比較して大幅改善）TechCrunch3
- 実運用型タスク
  - ウェブナビ（航空会社サイト）：GPT-5 63.5%、o3 64.8%（o3 が僅差で上回る）TechCrunch3

体系差・運用面の違い

- 低幻覚・低欺瞞：GPT-5（thinking）は o3 比で事実誤りが約 80%少ない。欺瞞的応答率も削減（大規模実トラフィックに基づく評価）。難題では「できない」と判断して明確に制約を述べる傾向が強いです。OpenAI1
- 効率とルーティング：GPT-5 は高推論条件で o3 より出力トークン 22%減・ツール呼び出し 45%減。API では推論モデルを直接利用、ChatGPT では推論/非推論/ルーターの統合システムとして動作します。OpenAI2
- コンテキスト長：比較サイトによれば GPT-5（high）40 万トークン、o3 は 20 万トークン（指標上は GPT-5 が 2 倍）。超長文の一括処理・比較・レビューでは GPT-5 が有利です。Artificial Analysis8

分野別の実力差（物理・数学・コーディング・化学・論理）

- 数学
  - 定量比較では GPT-5 が AIME 2025 で頂点級。o3 も非常に強いが、最難度域（ツール無し）まで含めた安定 SOTA は GPT-5 に分があります。OpenAI1 OpenAI5 VentureBeat4 DataCamp6

- コーディング
  - SWE-bench Verified など「実在リポジトリ修正」の再現性が高い系で GPT-5 が優位。編集系（Aider）も GPT-5 の高精度が確認され、業務自動化・大規模コードベースの保守において有利。o3 は競プロ・編集系でも強力で、思考時間を増やすことで高 Elo の到達が可能です。OpenAI2 DataCamp6
- 論理（抽象推論・規則学習）
  - GPT-5 は GPQA などで高得点かつ低幻覚・低欺瞞。o3 は ARC-AGI など抽象推論で人間水準を超えるスコアを報告。安定した事実性・説明責任を重視するなら GPT-5、発見的探索や視覚的規則性の発見には o3 も強みがあります。OpenAI1 DataCamp6
- 物理・化学（科学）
  - o3 系は STEM 最適化（o3-mini 含む）で博士レベルの生物・化学・物理問題に強く、思考努力量を上げると性能が伸びる特性。GPT-5 は医療（HealthBench Hard など）や MMMU でも高スコアで、科学図表・視覚理解を伴う課題での堅実さが増しています。OpenAI7 OpenAI1 TechCrunch3
- 実務運用・安全性
  - 幻覚・欺瞞の低減は運用安定性に直結。ヘルスケアのような高信頼領域や顧客対応・社内自動化では GPT-5 が優位。一方、特定の操作系課題（ウェブ UI ナビ）は o3 に軍配が上がるケースもあるため、タスク特性で選択が分かれます。OpenAI2 TechCrunch3

用途別の選び方（実務観点）

- 高信頼・低幻覚が必須（医療、金融、レポーティング、社内自動化）：GPT-5 を第一候補。長文比較や監査追跡が必要なワークフローでも有利（長コンテキスト・少ツール/少トークン）。OpenAI2 Artificial Analysis8
- 競技的/探索的推論や視覚的パターン発見重視（ARC-AGI 的課題、図表推論、競プロ Elo）：o3 が良い選択肢。思考計算量を投じるほど性能が伸びやすい設計思想です。DataCamp6 OpenAI5
- 実務的コード修正・保守（SWE-bench 型の課題、Aider 編集）：現状は GPT-5 が優位。既存リポに対するエージェント型修正で成功率が高い報告が多いです。OpenAI2
- ウェブ操作・UI ナビのピンポイント課題：一部ケースでは o3 が僅差で優位。実デモやパイロットでの比較評価を推奨します。TechCrunch3

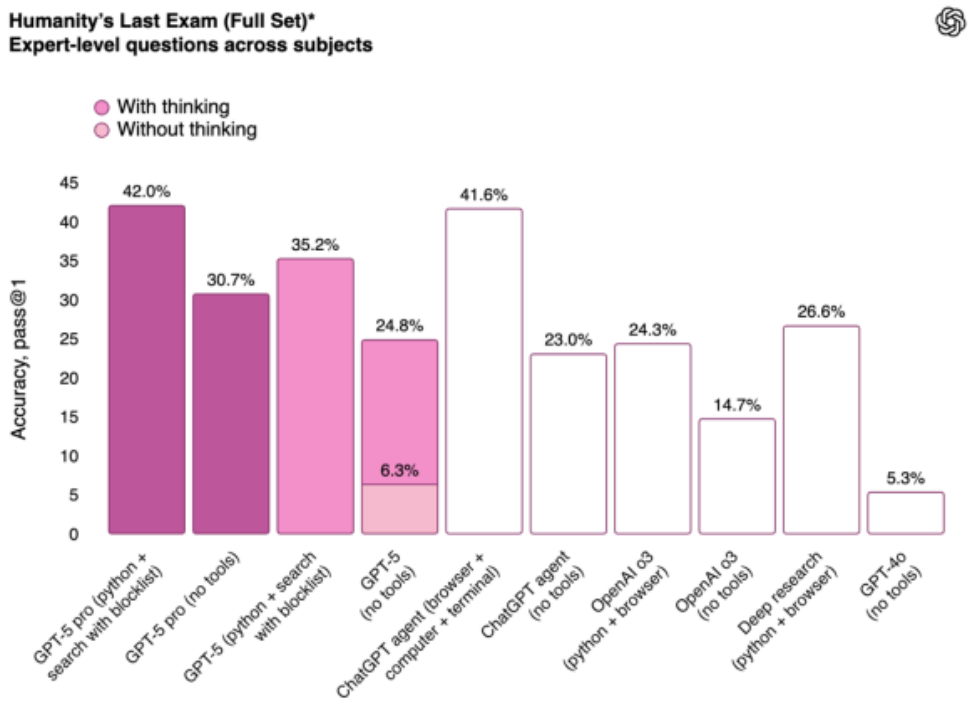
参考となる公式・報道リンク（エンティティ導線）

- GPT-5（OpenAI 公式）1 / GPT-5 for Developers2 / GPT-5 System Card9（PDF 含む）OpenAI1
- OpenAI o3（公式発表）5 / o3-mini（公式）7 OpenAI5

- 報道・第三者まとめ：TechCrunch、VentureBeat、Artificial Analysis、DataCamp など。TechCrunch3 VentureBeat4 Artificial Analysis8 DataCamp6

図版（出典つき）

- o3 は画像・図表・グラフを伴う視覚推論に強み（公式発表より）。OpenAI5



- 医療分野の幻覚率やウェブナビ比較など、GPT-5 と o3 の差分事例がまとまっています。TechCrunch3

- o3 は SWE-bench Verified で 69.1%と大幅向上(o1 比)。GPT-5 は同ベンチで 74.9%に到達しています。DataCamp6 OpenAI2

注意点・読み替え

- ベンチマーク条件（ツール使用可否、思考計算量、ルーティング、consensus サンプリング等）でスコアは変動します。比較時は「同一条件・同一評価セット」かを確認してください。OpenAI1 OpenAI5
- SWE-bench はスキャフォールドや評価環境に敏感です。公表値は各評価設定に依存するため、導入前には自社コードベースでの PoC 計測を推奨します。OpenAI2 www.swebench.com10

簡易まとめ（TL;DR）

- 数学：両者とも超強力。ツール無し到達や安定 SOTA は GPT-5 が一歩先。OpenAI1 DataCamp6
- コーディング：現時点の実務系ベンチ（SWE-bench Verified）では GPT-5 が優位。o3 も強いが総合では GPT-5 に軍配。OpenAI2 DataCamp6
- 論理：GPT-5 は低幻覚・低欺瞞で安定、o3 は抽象パターン学習系で非常に高い潜在力。タスクで使い分け。OpenAI1 DataCamp6
- 物理・化学：o3（と o3-mini）は STEM 最適化で強く、GPT-5 は医療・視覚理解を含む科学系で堅実な高得点。総合では GPT-5 が優勢の傾向。OpenAI7 OpenAI1 TechCrunch3
- 運用：長文・高信頼・省コスト（出力/ツール）なら GPT-5。特定の UI 操作では o3 の方が僅差で良い場面もあり、実データでの A/B が有効です。OpenAI2 TechCrunch3

必要なら、あなたの具体的な課題（数式証明、物理系数値解析、化学反応機構推論、既存リポ修正、自動テスト生成など）に合わせて、評価条件（ツール利用、思考計算量、コンテキスト設計）を最適化した「実データ PoC 計画」をご提案します。どのユースケースから検証しますか？OpenAI2

Appendix: Supplementary Video Resources

GPT-o3/Gemini2.5/Claude3.7/Grok3 横评，o3 竟然不是最强的？

Apr 20, 2025

GPT-4.5 vs o3 mini high：逻辑、情绪、创意三个问题的测评 ...

Mar 15, 2025

【遂にキタ ‼️】超使える推論オープンモデル「GPT-OSS ...

24 hours ago

もっと詳しく

1
openai.com

2

openai.com

3
techcrunch.com

4
venturebeat.com

5
openai.com

6
www.datacamp.com

7
openai.com

8
artificialanalysis.ai

9
openai.com

10
www.swebench.com