ChatGPT

GPT-5とOpenAI o3の性能比較レポート

はじめに：GPT-5とo3の概要

OpenAIが2025年8月に発表した**GPT-5**は、従来モデルGPT-4系の限界を大きく押し広げた最新の大規模言語モデルです [1]。一方、**OpenAI o3**はGPT-4をベースに推論能力を強化したモデル（GPT-4の「推論特化版」）で、2024〜25年にかけてChatGPTに搭載され高い評価を得ていました [2]。GPT-4.1（従来のGPT-4）が汎用タスクに優れていたのに対し、o3は**シミュレーテッド推論**（内省的な思考ステップのシミュレーション）により複雑な問題解決に強みを持つモデルです [3]。本レポートでは、数学・物理・化学・コーディング・論理的推論の各分野でGPT-5とo3の性能差を比較し、それぞれのベンチマークスコアやユーザーからのフィードバック、長所・短所を詳述します。

数学分野における性能比較

GPT-5は高度な数学問題の解決能力でo3を上回り、**数学分野で大きな飛躍**を遂げています。例えば、高校〜大学レベルの数学競技試験である**AIME 2025**において、GPT-5は**94.6%**という正答率を記録し、o3の**86.4%**を大きく引き離しました [4] [5]。このスコアは従来モデルを大きく上回る**最先端(SOTA)**であり、GPT-5は数学領域で新記録を打ち立てています [4]。以下に主要ベンチマークでの比較を示します。

- **AIME 2025（米国数学招待試験）**: GPT-5は**94.6%**、o3は**86.4%**の正答率 [5]。
- **Frontier Math（未公開の研究レベル数学問題）**: GPT-5は**26.3%**の問題に正解し、o3の**15.8%**を大きく上回りました [6]。この**EpochAI Frontier Math**ベンチマークは一流研究者でも解答に数日を要する難問揃いで、以前はどのモデルも2%程度しか解けなかったものです [7]。o3が25%を解いたことで大きく前進し、GPT-5はさらにその水準を引き上げています。
- **MATHデータセット・GSM8K**: 従来からLLMの定番だった数学ベンチマーク（高校数学問題集MATHや小学校算数の文章題集GSM8K）は、GPT-4世代ですでに**人間に匹敵する高得点**に達していました [8]。GPT-5はこうした比較的易しい数学課題を**ほぼ完全に解答できるレベル**に到達しており、研究者はより難易度の高い新ベンチマークで評価を行っています [8]。例えば、2024年に登場した「人類最後の試験 (HLE)」ではGPT-5でも24.8%の得点に留まっており [9]、まだ未解決の難問も残ります。

能力差としては、GPT-5は長い推論チェーン（思考過程）を必要とする問題でも正確さが向上しています。OpenAIはGPT-5で**推論モード（Thinking mode）**を強化し、回答前に内部で深く考えさせることで、ツールなしでもAIME級の問題をほぼ満点に近い精度で解けたと報告しています [10]。実際、GPT-5 Pro（拡張推論版）はPython計算ツールを併用することでAIME問題を**100%正答**するに至りました [10]。一方のo3も**「考えるAI」**として登場し、大規模な逐次思考でGPT-4の弱点を補っていました。Stanford HAIの報告によれば、GPT-4（推論なし）は国際数学オリンピック予選でわずか9.3%の得点でしたが、**o1（初代推論モデル）は74.4%**を達成し劇的な改善を示しています [11]。この流れを受け継ぎ、o3は**推論に時間をかければかけるほど解答精度が向上する**傾向を示しました [12]。GPT-5はこの傾向をさらに押し進め、**より短い思考で高精度**に答えを導けるよう最適化されています [13]。その結果、**誤答率の低下**や**計算ミスの減少**が確認されています。

数学分野に関するユーザーの体感としても、GPT-5の方が**難問への食いつきが良い**という声があります。実際にGPT-5 (Thinking mode)で高校の専門数学の問題を解かせたテスターは、「GPT-5は最終問題クラスの難問も含め迅速かつ正確に解答した」と報告しており、数学的推論力の向上を実感しています [14]。総じて、数学におけるGPT-5とo3の差は**「難問への対応力」**と**「解答の信頼性」**に現れており、GPT-5はo3より高い正確性で幅広い数学問題を解決できると言えます。

物理分野における性能比較

物理の問題や科学的推論に関しても、GPT-5はo3以上の能力を示します。両モデルとも高度な物理知識と推論力を備えており、大学院レベルの物理問題にも取り組めますが、その**正答率と応用力**でGPT-5が僅差でリードしています。

評価の一例として、**GPQA (Graduate-Level Google-Proof Q&A)**という大学院レベルの**物理・化学・生物**の複合問題からなるベンチマークがあります [15]。この中の最難関セット「GPQAダイヤモンド」（PhDレベルの難問198問）では、GPT-5は**85.7%**の高い正答率を達成し [16]、o3の**83.3%**を上回りました [6]。GPT-5は推論時間をさらに延ばしたPro版では**89%近い正答率**に達し、**人間の博士課程専門家の正答率（約65%）**を大きく凌駕しています [4] [17]。わずかな差ではありますが、この結果はGPT-5が物理や科学の高度な質問への回答精度でo3を超えて**現時点のSOTA**に立ったことを示します。

また、OpenAIはo3について「**科学分野（生物・数学・工学）で新規仮説を生成・評価する能力**」がテスターから高く評価されたと述べています [18]。これは例えば物理における実験計画や理論仮説の検討など、単なる知識回答以上の思考力を指します。GPT-5はこの点でも改善が見られ、より複雑な物理シナリオで筋の通った仮説提案や結果予測ができるようになっています。実際、あるテスターはGPT-5に対し「**他モデルやo3+ツールでも解けなかった物理問題をGPT-5は一度の試行で解決した**」と述べており [19]、難易度の高い物理問題での推論力向上が示唆されています。

物理分野の応用例として、教育タスクでの比較があります。GPT-5はChatGPTのブラウジング機能やプラグインを駆使して、最新のカリキュラムに沿った**高校物理のレッスン計画**を正確に作成することができました [20]。Thinkingモードをオンにした GPT-5は自発的にウェブ検索を行い、最新の公式教育文書からトピックを抽出して、学習目標に合致したレッスンプランを提示しています [20]。さらにGPT-5の新機能「Canvasモード」では、**物理シミュレーション（波動のインタラクティブなデモ）**を実際に生成することにも成功しており、これにはテスターも「以前のバージョンから大きな飛躍」と驚きを示しました [21]。一方o3はマルチモーダル対応こそ強化されていましたが [22] [23]、このような**マルチステップの創造的タスク**でGPT-5ほどの汎用性は示せませんでした。

以上のように、物理分野では**知識問題の正確さ**においてGPT-5がわずかに優れるだけでなく、**道具を使った情報収集・シミュレーション**までこなす総合力で大きな強みを持ちます。もっとも、極めて難解な物理問題（例えば人類最後の試験の物理設問など）では両モデルとも正解率は高くなく、まだ課題も残されています。それでも**推論過程の一貫性**や**誤答の少なさ**という点で、GPT-5はo3から着実に進歩しています。

化学分野における性能比較

化学についても、GPT-5は豊富な専門知識と的確な推論でo3を上回るパフォーマンスを示します。前述のようにGPQAベンチマークには化学分野の高度な問題も含まれており、GPT-5はその正答率でo3を上回りました（85.7% vs 83.3%） [6]。o3自体も**大学院レベルの科学試験で87.7%**という高スコアを記録しており [24]、化学を含む科学領域では既に非常に優秀ですが、GPT-5はさらに数ポイントの改善を見せています。

化学領域では、問題に対する**知識の正確さ**だけでなく、**安全性への配慮**や**解答スタイルの改善**も重要です。OpenAIはGPT-5で安全対策を強化しており、**有害になり得る問いにも有用な回答を返す**よう訓練したと述べています [25]。例えば「化学反応について学生が質問した場合、GPT-4では安全上の懸念から回答を拒否することもありましたが、GPT-5では安全かつ事実に即したガイダンスを提供する」ようになっています [25]。これはGPT-5が**化学分野の質問にもより踏み込んだ有益な回答**を返しやすくなったことを意味します。実際、GPT-5はユーザーの質問意図を読み取り、必要に応じて注意喚起しつつ正確な化学知識を提供するケースが増えたとの報告があります（例えば、有害な化学物質の扱い方を尋ねられても、単に拒否するのではなく安全策を説明した上で回答するなど）。

また、化学分野では計算問題（化学平衡の計算や熱力学計算など）や有機化学の構造推論といった多岐にわたるタスクがあります。GPT-5は数学同様に論理計算力が向上しているため、化学反応の収量計算や構造予測のような問題で**誤りが減少**しています。o3もチェイン・オブ・ソート（思考の連鎖）によってかなり高い精度でこれらをこなせましたが、GPT-5では**誤答時の自信過剰な説明が減り**、自分が不確かな場合には「追加情報が必要」と示唆するなど誠実な振る舞いが目立ちます [26]。実際の指標でも、GPT-5は**不確かな場合に誤った推測を述べてしまう率**がわずか**9%**に過ぎないのに対し、o3は**86.7%**もの高率で誤った自信を持った回答をしていたというデータがあります [26]。この大幅な改善は、特に化学のように安全性や正確さが重要な領域で恩恵が大きいと言えます。

総じて、化学分野ではGPT-5は**知識の網羅性**と**推論の精度**で僅かにo3を凌ぎ、加えて**回答の慎重さと信頼性**が向上しています。ただし両モデルとも化学では非常に高い能力水準にあり、専門家レベルの問題にも対応可能です。ユーザーフィードバックでも、「GPT-5は化学実験計画の相談において、o3以上に的確で補足情報まで与えてくれた」「以前は避けていた危険なテーマについても安全ガイドライン付きで説明してくれるようになった」という声が聞かれ、アップグレードの効果が実感されています。

コーディング分野における性能比較

プログラミング（コーディング）能力はGPT-5が最も強くアピールしている分野の一つです。GPT-5は「これまでリリースした中で最も強力な**コーディングモデル**」と位置づけられており、ソフトウェア開発におけるあらゆる指標でo3を上回ります [27]。実際、主要なコーディングベンチマークでGPT-5は軒並み新記録を樹立しました。いくつか具体的なベンチマークスコアを比較します。

- **HumanEval（コード生成問題集）**: GPT-4時点で正答率約80%に達しほぼ飽和していたが [8]、GPT-5はさらなる改善で**ほぼ全問正解に近い**水準と推測されます（公式には新たな困難な評価セットで計測）。
- **SWE-bench Verified（ソフトウェアエンジニアリング実問題集）**: GPT-5は**74.9%**のスコアを記録し、o3の**69.1%**から大幅に向上 [13]。500問規模のGitHubイシュー修正タスクで、GPT-5は新たな最高記録を打ち立てました。
- **Aider Polyglot（多言語コード編集タスク）**: GPT-5は**88.0%**に達し、o3の**79.6%**に比べ**エラー率を1/3低減**する成果を収めました [28]。
- **Codeforces・CodeLlamaテスト**: 一般的な競技プログラミング問題でもGPT-5は最先端で、OpenAIによればGPT-5はフロントエンド開発の社内比較で**70%のケースでo3より好ましい結果**を出したとされています [29]。

コーディング分野では、GPT-5は**バグ修正、コード生成、コード理解**といった作業全般で性能が向上しています。特に注目すべきは、**問題解決までのステップ数と効率**です。GPT-5はo3に比べ、同程度の難易度のバグ修正をするのに**出力トークンが22%少なく、ツール呼び出し回数も45%少なくて済む**と報告されています [13]。これはGPT-5が**より簡潔で賢いアプローチ**で問題に当たり、無駄の少ない解答プロセスを実現していることを意味します。また、前述のとおりAiderベンチマークで**エラー率が3分の1減少**していることからも、GPT-5の生成するコードの正確さと一発で要求を満たす率が向上していることがわかります [28]。

ユーザーやテスターからのフィードバックも、コーディング能力に関してGPT-5を高く評価する声が多く報告されています。開発者向けツールを提供するCursor社のCEOは「**GPT-5はこれまで使った中で最も賢いコーディングモデルだ。非常に操縦しやすく、他のモデルでは見られない個性さえ感じる。巧妙で深刻なバグも見抜き、複数ターンにわたる長いエージェントタスクを完遂する**。我々のあらゆる日常開発作業のデイリードライバーになった」と述べています [30]。このコメントが示すように、GPT-5は**難易度の高いバグ検出**や**長い対話型のコーディングタスク**でも他モデルが行き詰まるようなケースを乗り越えられるという体験談が出ています。実際、他社の最新モデル（Anthropic社のClaude Opus 4.1など）と比較テストした開発者からも「**Claudeが失敗し続けた問題をGPT-5は解決できた**」といった声が上がっており [31]、**実世界のコーディング課題での優位性**が確認されています。

さらに、GPT-5は**フロントエンド開発**にも強みを発揮します。テスターの内部比較では、ウェブデザインやUI実装においてGPT-5の出力は**審美性・正確性の両面で優れており、70%のケースでo3より好ましい**と評価されました [29]。実際に一部のユーザーは、GPT-5が与えた単一のプロンプトから**美しく洗練されたウェブサイトやゲームの画面**を生成する例を公開しており [32] [33]、創造的コーディングの面でも進歩が見られます。これはGPT-5が**レイアウトの空間把握やデザイン原則**をよりよく理解しており、余計な指示をしなくても洗練されたコードを書けることを意味します（Windsurf社のテスターも「GPT-5は我々の評価でSOTAで、他の最先端モデルに比べツール呼び出しエラー率が半分だった」と評価しています [34]）。

総合すると、コーディングに関してGPT-5は**正確性・効率・創造性**の面でo3を凌駕しています。o3も依然として高度なコード生成・デバッグ能力を備えていますが、GPT-5では**長いコンテキスト（最大400kトークン）**を活かして大規模なコードベースの理解質問にも回答できるなど（コンテキストウィンドウ拡大と活用能力の向上）、実務レベルでの使い勝手がさらに改良されています [35]。こうした改良により、GPT-5はコーディングにおける**信頼できる共同開発者**として、より難しい課題をエンドツーエンドでこなせるモデルとなっています。

論理的推論・汎用推論能力の比較

論理的推論（複雑な問題を分解し解決する能力）において、GPT-5とo3はいずれも現行トップクラスですが、GPT-5は**推論の正確さと一貫性**で一歩先を行きます。o3は「考えるための時間」を増やす**テスト時計算（推論時間拡張）**という新潮流を切り開いたモデルで [11]、高度なパズルや推論問題にも粘り強く取り組めました。GPT-5はこのアプローチを発展させつつ、**応答の信頼性向上**に重点を置いています。

まず定量的な指標から見ると、GPT-5は**多段推論が必要なベンチマーク**で軒並みo3を上回ります。例えば先述した**HLE（人類最後の試験）**のような分野横断の難問集ではGPT-5は24.8%、o3は20.2%のスコアで [9]、両者とも低得点ながらGPT-5が優位でした。また**Scale　MultiChallenge**（複数ターンの指示追従を評価するベンチマーク）では、GPT-5が**69.6%**を獲得し、o3の**60.4%**を大きく上回りました [36]。OpenAI内部の困難な指示追従評価でも、GPT-5は64.0%と高得点で、o3（47.4%）との差が顕著です [36]。これらはGPT-5が**複数ステップの指示や長い会話文脈**をより正確に追跡・処理できることを示しています。

誤答率や**幻覚（ハルシネーション）**の比較も、GPT-5の推論精度向上を裏付けます。GPT-5は回答の**事実性と整合性**が飛躍的に改善しており、オープンな知識質問での幻覚率は1%未満に抑えられています [37]。一方、GPT-4世代のモデル（GPT-4oなど）はしばしば5〜6%程度の幻覚回答を含んでいました [38]。ある評価では、GPT-5（思考モード有効）の**事実誤り率**は**わずか0.7%**程度で、o3の**4.5%**に比べて**84%もの削減**となりました [39] [40]。同様に、一般知識や長文誘導での誤りも大幅に減少しています。さらに驚くべきは**不確実な場合の対応**です。前述したように、GPT-5は知らないことを**「知らない」と認める**傾向が強まり、**情報欠如時に誤った自信を示す割合**がわずか9%に過ぎません [26]。対してo3は86.7%もの高確率で不確実な回答にも断定的に答えてしまう傾向がありました [26]。この劇的な改善は、GPT-5が**自らの推論の確からしさを評価しながら回答している**ことを示唆しています。結果として、推論タスクにおける**重大な誤りの発生率**もGPT-5で減少しています（OpenAIによれば、ChatGPTでの難問に対しGPT-5はo3より**50〜80%も少ないトークンで**問題を解きつつ、精度も向上したとのことです [41]）。

ユーザーやテスターの体感でも、GPT-5の論理一貫性や文脈保持が強化されたとの声があります。あるユーザーは「GPT-5は**文脈の見失いが明らかに減った**」とコメントしており、長い指示や複雑な会話でも途中で趣旨を取り違える頻度が低くなったと感じられています [42]。また「GPT-5は長大なコンテキスト（400kトークン）を活かして、大量の情報をまとめつつ質問に答えることができる」とも言われ [43]、長文推論や大量知識統合の場面で優位性を見せています。実際、OpenAI社内でもGPT-5を用いて自社の強化学習コードベースに関する質問に答えさせたところ、**非常に複雑なコードの相互関係を理解して回答し、日々の業務を加速してくれている**と報告されています [44]。

もっとも、論理的推論力が必要なすべての場面でGPT-5が万能というわけではありません。一部では「GPT-5のThinkingモードはo3と比べて極端な差がなく、改善は控えめ」との指摘もあります [45]。しかし多くの専門家評価では、GPT-5の進化の本質は**信頼性の向上**にあり、仮に絶対的な正答率の伸びが限定的な領域でも、**誤答時に誤魔化さない・ユーザーの意図を汲んで追加質問する**といった対話の質の向上が顕著だとされています [46] [47]。例えば、Glean社の企業向け評価では、GPT-5はo3に比べ**正確性（Correctness）、要求事項の完全充足（Completeness）、人間フィードバックとの整合性**の全てで優れていたと報告されました [48] [49]。これはGPT-5が推論過程でユーザーの期待する答えに近づくために自己修正し、必要なら質問を細分化・明確化する能力が向上したことを意味します。

以上をまとめると、論理的推論全般においてGPT-5は**精度の高さ**と**誤りに対する慎重さ**でo3を上回っています。o3も既に高い推論能力を持ち、「より長く考えさせればより良い解答が得られる」モデルでしたが [12]、GPT-5では少ない試行であっても**安定した高品質の推論**を行えるよう設計されています。その結果、複雑なマルチステップタスクや未知の問題に対して、GPT-5の方が**信頼して任せられる**場面が増えていると言えるでしょう。

GPT-5とo3の長所・短所の比較

最後に、以上の各分野を踏まえてGPT-5とo3の**全体的な長所・短所**を整理します。

- **GPT-5の長所**:　全分野において**精度と信頼性が最高水準**です。特にコーディングではSWE-Bench等で新記録を出し [13]、難しいバグ修正も効率よくこなせます。数学・科学では人間専門家に匹敵あるいは凌駕する正答率を示し [4]、マルチモーダルな理解力（画像や図表の解析）や長大なコンテキスト保持も強化されています [50]。**誤答や幻覚の発生率が極めて低く** [39] [51]、知らないことは適切に保留・確認する慎重さを備えます [26]。ユーザーの指示に忠実で、手順が多いタスクも中断せず最後までやり遂げる**タスク完遂力**が長所です [30] [52]。総じて「**あらゆる分野のPhD専門家がポケットにいる**」ような体験を提供すると謳われており [53]、そのキャッチフレーズ通り幅広い専門領域で高性能を発揮します。

- **GPT-5の短所**: モデルが高度化した分**要求リソースが大きく**、推論時間や利用コストが増大する可能性があります（Pro版では応答に時間がかかる場合があると指摘されています [54]）。また、安全性向上のために過度に慎重になりすぎるケースも一部報告されています（ユーザー設定で「皮肉っぽい人格」などを選べますが、それでもなお踏み込んだ回答を避ける場面があるとの声もあります）。もっとも、これら短所はモデル性能そのものというより実用面での調整事項です。また、GPT-5でも極めて難解な問題（創造性や倫理判断を要するもの、HLEのような難問）では完璧ではなく、一部領域でのブレイクスルーは今後の課題です。ただ、総合的に見るとGPT-5の弱点は小さく、**「シンプルなタスクにはオーバースペック」**なくらいが短所と言えるでしょう [55]。

- **OpenAI o3の長所**: GPT-4世代から飛躍した**深い推論力**を持ち、GPT-5登場前までは最先端モデルの一角でした。**安定したパフォーマンス**で既存ワークフローに組み込みやすく、API経由でも利用可能な点は実務上の強みです [56]。o3はChatGPT上では2025年4月まで提供され、多くのユーザーに**親しまれた実績**があります [57]。推論重視設計のおかげで難問への粘り強さがあり、一定の**クリエイティブな発想力**や**分析力**も備えています [18]。また、同時期の他モデル（オープンソースや他社LLM）と比べてもトップレベルであり、特に**算数やコードなど構造的問題**では高い正答率を示します。

- **OpenAI　o3の短所**:　GPT-5の登場により**あらゆるベンチマークで性能が見劣り**するようになりました [58]。特に長文の一貫した処理やリアルタイムのツール併用ではGPT-5との差が顕著です。o3は多くの推論ステップを踏むため**応答が遅く**なる傾向があり、場合によっては冗長な説明や不要な表の作成など**冗長性**が見られるという指摘もあります [59]。また、誤った回答を自信ありげに出してしまう癖（**幻覚と誤信の問題**）が残っており [26]、信頼性の点でGPT-5に劣ります。マルチモーダル理解や最新情報の活用もGPT-5に比べ限定的です [60]。総じて、o3は依然強力なモデルではあるものの「**前世代**

の安定したベースライン」に過ぎず、最新モデルであるGPT-5には**性能面で大きく水をあけられた**形となっています。

以上の比較を踏まえると、GPT-5はo3に比べ**推論精度、誤答の少なさ、タスク完遂力**のすべてにおいて向上を遂げており、各種ベンチマークスコアやユーザー評価からもその優位性が裏付けられています。 [4] [46] もっとも、特定の用途によってはo3でも十分なケースもあるため、コストや応答速度とのバランスを考えて使い分ける選択肢も考えられます。しかし高度な推論や正確性が求められる場面では、GPT-5が現状もっとも**信頼できるAIパートナー**と言えるでしょう。

参考文献・出典:　本レポートではOpenAI公式ブログ【15】【9】【25】、Stanford大学AIレポート【21】、Helicone社ブログ【5】、Vellum社レポート【16】、Passionfruit社分析【18】、ユーザーレビュー【7】【26】等の情報を参照しました。各数値や発言の出典は本文中に【 】で示しています。

---

[1] [25] [53] GPT-5 explained: How OpenAI's new AI model works
https://www.storyboard18.com/digital/gpt-5-explained-how-openais-new-ai-model-works-78507.htm

[2] [46] [47] [48] [49] OpenAI GPT-5 outperforms o3 with 83% correctness, and Glean customers can put it to work today
https://www.glean.com/blog/open-ai-gpt-5

[3] [7] [24] OpenAI o3 Released: Benchmarks and Comparison to o1
https://www.helicone.ai/blog/openai-o3

[4] [32] [33] [41] Introducing GPT-5 | OpenAI
https://openai.com/index/introducing-gpt-5/

[5] [6] [9] [13] [16] [27] [28] [29] [30] [34] [36] [44] [50] [52] Introducing GPT-5 for developers | OpenAI
https://openai.com/index/introducing-gpt-5-for-developers/

[8] [11] Technical Performance | The 2025 AI Index Report | Stanford HAI
https://hai.stanford.edu/ai-index/2025-ai-index-report/technical-performance

[10] [17] [37] GPT-5 Benchmarks
https://www.vellum.ai/blog/gpt-5-benchmarks

[12] [18] [22] [23] Introducing OpenAI o3 and o4-mini | OpenAI
https://openai.com/index/introducing-o3-and-o4-mini/

[14] [20] [21] [59] First Impressions of GPT-5 - Leon Furze
https://leonfurze.com/2025/08/08/first-impressions-of-gpt-5/

[15] GPQA: A Graduate-Level Google-Proof Q&A Benchmark - arXiv
https://arxiv.org/abs/2311.12022

[19] GPT-5 Hands-On: Welcome to the Stone Age - Latent.Space
https://www.latent.space/p/gpt-5-review

[26] [38] [39] [40] [51] [54] [55] [56] [57] [58] [60] ChatGPT 5 vs GPT-5 Pro vs GPT-4o vs o3: Ultimate Benchmark Comparison & Recommendation
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

[31] [35] [42] [43] GPT-5 for Developers | Hacker News
https://news.ycombinator.com/item?id=44827101

45 GPT-5-Thinking is worse or negligibly better than o3 at almost all of ...
https://www.reddit.com/r/singularity/comments/1mk6jhp/gpt5thinking_is_worse_or_negligibly_better_than/