ChatGPT

Claude Opus 4 と ChatGPT o3 の技術比較

Anthropic社の **Claude Opus 4** と OpenAI社の **ChatGPT o3**（以下、便宜上「o3」と表記）は、2025年時点で両社が提供する最先端の大規模言語モデルです。それぞれ長文コンテキスト処理、幻覚（ハルシネーション）の抑制、多ターン対話での推論一貫性、モデル設計、ベンチマーク評価などで特徴があります。以下では公式技術資料や第三者の検証に基づき、両モデルの性能と設計上の違いを比較します。

比較概要（主要スペックと特徴）

まず両モデルの主要な技術的特徴をまとめた比較表を示します。

比較項目	Anthropic Claude Opus 4	OpenAI ChatGPT o3
コンテキストウィンドウ (長文入力容量)	最大約 **200,000 トークン**まで対応（Claude 2で訓練済。製品では一時100kトークン実装、のち200k拡張。開発者向けキャッシュ利用で最大50万トークン相当も可能） [1] [2]	最大 **128,000 トークン**まで対応（GPT-4世代から拡張） [2] 。極端な長文より、**外部ツール連携による文脈管理**を重視 [3]
長時間タスク処理 (持続的な実行能力)	**数時間・数千ステップ**に及ぶ連続タスクでも性能を維持 [4] 。特にコーディングなど長時間の集中作業で安定した結果を出せるよう最適化 [4]	**計算資源増強モード**（o1 Proモード）で難問時に長く「考える」設定が可能 [5] 。複雑なワークフローで**外部ツールと連携**しつつ一貫性を保つ設計 [3]
幻覚抑制策 (RAG・文献参照)	**検索ツール統合**: モデル自身がWeb検索等を実行し正確な情報を入手（β版の拡張思考モードで実現） [6] 。返答に**出典情報**を付与する引用機能も提供 [7] 。加えて、憲法AIによる真実性原則で応答を自己検証 [8]	**プラグイン/関数呼び出し**: モデルがプラグイン経由で検索や社内データベース照会を行い情報補強（ChatGPTプラグイン機能）。**RLHF**で事実誤りを減らす調整 [9] 。応答制御のための**システムメッセージ階層**採用 [9]
多ターン推論の一貫性	**長大メモリとファイル永続化**: 必要に応じ対話中に要点をファイル保存し、60分間のコンテキストキャッシュで継続利用可能 [10] [11] 。一貫した推論パターンで**安定性重視**の回答を生成 [12]	**高度推論モード**: o1 Proモードでは内部でより長い思考ステップを実行し**信頼性の高い回答**を生成 [13] 。複数ツールを絡めた複雑な対話でも**整合性を保つ設計** [3]
モデル設計・アーキテクチャ	**Transformerベース**。パラメータ数非公開（Claude 2は推定数十億〜百数十億規模）。**Constitutional AI**で訓練（AI自身のフィードバックによる有害性低減と誠実さ向上） [14] 。**Opus 4**は**2モード統合ハイブリッド**モデル（高速応答と深慮応答の両立） [15]	**Transformerベース**。パラメータ非公開（GPT-4世代は推定数千億〜1兆規模とも噂）。**RLHF**（人間フィードバックによる調整）が主体 [9] だが、自動連想チェイン（Chain-of-Thought）や高度なデータフィルタで**幻覚低減** [9] 。o1 Proでは**動的計算深度**を導入し難問時に内部で複数解候補を検討 [5] [16]

比較項目	Anthropic Claude Opus 4	OpenAI ChatGPT o3
代表的ベンチマーク性能	**コーディング**: SWE-Bench（ソフト工学）スコア 72.5% [4]（世界最高水準）。 **学術QA**: MMLUベンチ87.4%（拡張思考なし）[17] – GPT-4並み。 **数学**: AIME2024試験 74.9%（GPQA Diamond, 拡張思考なし）[17]	**コーディング**: SWE-Bench 69.1%（前世代48.9%から飛躍）[18]。競プロElo 2706 [18]で従来モデルを大きく上回る。 **数学**: AIME2024でGPT-4の正解率50%から **86%**まで向上（o1 Pro）[19]。難問QA (GPQA Diamond)でもGPT-4の74%→**79%**に改善 [20]

※表中のGPT-4はOpenAI従来モデル、o1はChatGPT o3世代の内部モデル名を指します。

以下、各項目について詳細に解説します。

長文コンテキスト処理の持続性

Claude Opus 4は極めて長い文脈（コンテキスト）を保持しながら会話や文章処理を行う能力が特長です。Anthropicは2023年にClaudeのコンテキストウィンドウを9kから100kトークンに拡大し [21]、Claude 2では**200kトークン**（約15万語）までモデルを訓練しました [1]。これは従来のチャットボットでは数十ページ分のテキストに相当する入力を一度に与えられることを意味し、**小説数冊分の内容を保持しつつ対話**できる計算量を備えていることになります [22] [23]。Anthropicのモデルカードによれば、Claude 2では長大な入力に対する**有用性・信頼性を高める工夫**がされており、複雑な長文ドキュメントからの質問応答や要約でも安定した結果を出すよう改善されています [24]。

Claude Opus 4ではこの大規模文脈処理能力がさらに発展し、**約200kトークンの文脈で質の高い応答**を維持することを目指しています [23]。実際、Claude 4 Opusのコンテキストウィンドウは公称200kトークンですが、Anthropicは開発者向けにプロンプトを一時キャッシュする機能を提供し、**最大1時間にわたり過去プロンプトを保管**することで更なる長時間のコンテキスト持続を可能にしています [11]。この仕組みにより、**実質50万トークン**（数十万語）にも及ぶ長時間の対話セッションでも文脈を保ちながらやりとりできるとされています [2]。

一方、**ChatGPT o3**世代（OpenAIの「oシリーズ」）もGPT-4以降コンテキスト長を徐々に拡大してきました。GPT-4では標準で8kトークン、拡張モデルで32kトークンがサポートされましたが、o3世代では**最大128kトークン**程度まで対応しています [2]。これはClaudeほどの極端な長さではないものの、一般的な書籍1冊分程度のテキストを保持できる計算量です。ただしOpenAIは、Anthropicのように単純にコンテキスト長を際限なく伸ばすアプローチよりも、**モデルが外部ツールを使って必要な情報を適切に取得する**手法に注力しています [3]。Mediumの分析によれば、OpenAIのo3は**「生のコンテキスト長よりも、ツールを介した文脈利用の効率性」**に重点を置いているとされます [3]。具体的には、長大なテキスト全てを一度に読ませるより、検索ツールやプラグインで必要部分を見つけ出し照会することで、限られたコンテキストを有効活用する戦略です。このように設計思想の違いがあり、Anthropicは**質を維持したまま長大な文脈を処理**する方向、OpenAIは**外部連携で必要情報を取捨選択**する方向でそれぞれ長文対応を進めています。

性能面では、Claude Opus 4は**長時間のタスクにおける持久力**が卓越しています。AnthropicはOpus 4について「数時間にわたる連続作業や何千ステップにも及ぶ問題解決でも性能劣化なく動作し、過去のモデルでは不可能だった規模の作業をエージェントが達成できるようになった」と述べています [4]。例えば、ある社外検証ではClaude Opus 4にオープンソースソフトウェアの大規模リファクタリングを7時間独立実行させ、**その間パフォーマンスが持続する**ことを確認したと報告されています [25]。これらは超長文や長時間対話で**文脈を見失わず安定して問題解決できる能力**を示しています。

ChatGPT　　o3も長い対話文脈での一貫性は強化されていますが、OpenAIは別のアプローチとして**モデルに「じっくり考えさせる」プロセス**を導入しています。2024年末にOpenAIが発表したChatGPT Pro向けのo1モデルでは、応答までにより多くの計算を行い**推論ステップを増やすプロ模式**が追加されました [5]。このo1 Proモードでは通常より長時間かけて思考させることで難問への回答精度を高めており [13]、特に**複雑な問題を解く際の信頼性**が向上したと評価されています [13]。言い換えれば、ChatGPTの方は**逐次的な推論過程**でコンテキスト上の問題を解決し、すべてを一度に保持するよりも**段階的に必要情報を取得・処理**する傾向があります。そのため、**会話履歴全体を丸ごと記憶**させる容量ではClaudeに劣るものの、**文脈を動的に整理しながら破綻なく回答を続ける**点で優れた設計となっています。

要約すると、**Claude Opus 4は極めて大きなコンテキストを丸ごと扱い長時間の集中対話でも精度劣化しにくい**のに対し、**ChatGPT o3はツール活用や内部の深い推論で必要情報を都度取り出しつつ文脈を維持する**アプローチを取っています。両者とも長時間・長文でのやりとりに強い工夫が凝らされていますが、その実現方法に違いが見られます。

検索拡張（RAG）と幻覚低減手法

大規模言語モデルは与えられた知識以上の質問に対し、それらしい回答をでっちあげてしまう**幻覚（ハルシネーション）**の問題が指摘されています。Claude Opus 4とChatGPT o3はいずれもこの幻覚を抑えるため、それぞれ**RAG（Retrieval-Augmented Generation）**や**外部知識参照**の仕組みを取り入れています。

Claude Opus 4は、モデル自身が**ツールを使って外部情報を取得できる**点が大きな特徴です。AnthropicはClaude　　4の発表時に、モデルが**内部でウェブ検索等のツールを呼び出し**ながら回答を改善する「拡張思考（Extended Thinking）」機能をβ提供すると発表しました [6]。これにより、ユーザからの質問に対しモデルがインターネット検索で最新の情報を参照し、その結果に基づいて回答することが可能です。モデルは思考とツール使用を**交互に切り替え**ながら解答を組み立てるため、知識カバー範囲外の質問でも無理に想像で補完せず、**根拠ある情報に基づいた回答**を出しやすくなります [6]。

さらにAnthropicはClaude APIに**「引用 (citation) 機能」**を実装しています [7]。これはモデルが参照した情報源の出典を回答中に明示する機能で、例えばClaudeが検索で見つけた記事内容を答える際に、そのURLやタイトルを引用形式で回答に含めることができます [7]。この機能はユーザにとって**回答の根拠を検証可能にする**だけでなく、モデル自身も「出典を示す」という方針の下で回答するため、曖昧な知識をでっちあげる頻度が下がる効果があります。Anthropicによれば、Claude 2.1へのアップデートで**幻覚の発生率が半減（2倍低減）**したとされ [26]、大幅なファクトチェック性能の向上が図られました。これは長文文脈処理能力の向上とあいまって、**文献やドキュメントを丸ごと読んで正確に要約・回答する**ようなタスクで、Claudeがより信頼性の高い結果を出せることを意味します。実際、第三者によるテストでも、100kトークン版のClaudeは75ページの学術論文を全文入力して質問に答えさせても、検索ベースの手法に匹敵する正解率を示しました [27]（ただし応答時間は数十秒とかかる）と報告されています。また別の検証では、文書中の細かい数値（例: 許可申請が必要な建物面積100平方フィートという規定）についてClaudeが若干の誤りを含む回答をした例もあり [28]、長文読解での完全な正確性には課題を残すものの、少なくとも**文章全体の要点把握や大筋での正確さは担保できている**ことが示されています。

Claudeにおける幻覚低減のもう一つの軸は**「Constitutional　　AI（憲法AI）」**によるモデル調整です。Anthropicのモデルは、人間フィードバックによる単純なYes/No評価だけでなく、あらかじめ定めたAIの「憲法（原則集）」に基づいてモデル自身が回答の妥当性を評価・修正する訓練を施されています [8] [14]。Claudeの憲法には「可能な限り真実で正確な答えを選ぶこと」や「知らないことは正直に知らないと言うこと」といった原則が含まれており、モデルはこれに従って出力を調整するよう学習しています [8]。その結果、Claudeは**不確かな場合に無理に答えず控えめな回答をする傾向**が強まり、結果として幻覚（誤情報の創作）の頻度が下がる効果があると考えられます [29]。実際、Cornell大学などの研究では「幻覚が少ないモデルほど、間違いそうな質問には答えを拒否する傾向が高い」ことが報告されており [30] [31]、Claudeのような方針（必要に応じて回答を拒否し幻覚を減らす）は有効なアプローチといえます。

一方、**ChatGPT　o3**（およびその前身のGPT-4やGPT-4.5相当モデル）は、幻覚低減のために**高度な事後調整と外部知識の統合**を組み合わせています。OpenAIはGPT-4で大幅に幻覚率を下げたとされ、ある調査では文章要約タスクにおいてGPT-4/Turboの**幻覚率が3%程度（正確性97%）**と、他のモデルに比べ圧倒的に低かったと報告されています [32]。この高い事実整合性は主に**RLHF（人間による強化学習フィードバック）**による調整で実現されています。GPT-4では大規模な人間評価データを用いて「事実に忠実な回答」が高くスコアされるよう報酬モデルを訓練し、モデルが好ましい回答パターンを強化されるよう学習させました [33]。OpenAIの次世代モデル（GPT-4.5やo1と称される系統）ではさらに、**自動生成した思考チェイン（chain-of-thought）を組み込む学習**や**大規模な未教師ありファインチューニング**を通じて、モデルの世界知識と論理整合性を高めています [9]。またトレーニングデータのフィルタリングも改善され、ノイズや誤情報を極力排除することで**基礎モデル自体の幻覚しにくさ**を高めていると言われます [9]。OpenAIは公式には詳細な手法を公表していませんが、第三者解析によればGPT-4.5世代では**「命令階層（instruction hierarchy）」**という仕組みでシステムレベル・開発者レベル・ユーザーレベルの指示を整理し、暴走や無関係な出力を抑制していると指摘されています [33]。総合すると、OpenAIのアプローチは**人的な評価軸を中心に据えつつモデル内部の論理運用も磨き上げる**ことで幻覚を低減していると言えます。

さらにChatGPT　o3では、Anthropic同様に**外部ツールやドキュメント参照の活用**もサポートされています。例えばChatGPTはプラグイン機能を通じてウェブブラウジングを行ったり、社内データへの検索（企業向けのChatGPT　Enterpriseではベクトルデータベース検索を統合）を行うことができます。これにより、モデル単独では知り得ない最新情報や限定的な知識領域についても、**リアルタイムに事実を取得して回答**することが可能です。Anthropicのようにモデル自身が能動的に検索する設計ではなく、OpenAIの場合はユーザ側やシステム側からプラグインを通じて情報を供給する形ですが、**Retrieval-Augmented　Generation（RAG）によって回答の根拠を保証する**という目的は共通しています。結果として、最新のChatGPT　(o3世代)も高度な質問に対して「こちらの情報源によれば…」と根拠を示したり、無闇に断定を避けて**エビデンスに基づく回答**を返す場面が増えています。

第三者の評価でも、OpenAIの最新モデルは幻覚の少なさで高く評価されています。例えばあるレッドチーム分析では、**GPT-4.5（OpenAI　o1相当)**が**Claude　3.7**よりも事実に即した回答をする割合が高く、**「誤解を招く出力を回避する点でGPT-4.5がリード」**と結論づけられました [34] [16]。一方でClaudeも基本的な事実整合性は高く、両者とも幻覚は大幅に減らされているものの、OpenAIモデルは特に**より慎重で正確**な傾向が見られたとされています [34] [35]。この「慎重さ」は前述のように副作用もあり、GPT-4.5世代は無害な質問にも拒否を返す**過度な慎重さ**が指摘されました [35]。Claudeはその点、危険な要求と無害な要求をうまく見分けて対応できているとの評価もあります [35]。したがって、**幻覚を抑える**という一点ではOpenAIモデルが一歩先んじたものの、その背景には回答拒否増加などトレードオフもあり、Anthropicモデルは**ユーザビリティと真実性のバランス**を取る形で改良が加えられていると考えられます。

まとめると、Claude　Opus　4とChatGPT　o3はいずれも**外部知識を活用して幻覚を減らす工夫**を備えています。Claudeはモデル自身が検索・引用し**根拠を示す回答**を行う点で特徴的であり、憲法AIにより**「知らないことは知らない」という姿勢**を貫くことで幻覚を抑制しています。一方ChatGPT　o3は、人間からのフィードバックと内部調整で**モデルの知識そのものの正確さ**を高め、必要に応じプラグインから情報を取得することで**事実に基づいた回答**を生成します。どちらも最新研究に裏付けられたアプローチであり、従来に比べ大幅に幻覚が低減された安全性の高い出力を実現しています。

マルチターン会話における推論の一貫性

マルチターン（対話型）のやり取りでは、モデルが前の発言や文脈をどれだけ覚えていて矛盾なく推論を続けられるかが重要です。Claude　Opus　4とChatGPT　o3はこの点でも、それぞれ独自の工夫によって**会話全体を通じた一貫性**を維持するよう設計されています。

Claude Opus 4は先述の通り巨大なコンテキストウィンドウを持ち、**直前の数十万トークン分**もの会話履歴をそのまま保持できます [1]。この物理的な記憶容量の大きさ自体が、一貫性維持には有利です。ユーザとの長

いやり取りの中で、遥か前半に出てきた情報であってもClaudeはコンテキスト内に保持し続けられるため、**途中で設定や前提を忘れて矛盾した回答をするリスクが低減**します。例えば物語のような長い対話で、序盤に登場したキャラクター設定を終盤でも正しく覚えている、といったケースです。Anthropicはさらに、開発者がClaudeに**独自の長期メモリ**を持たせることも可能にしています。Claude APIの新機能「Files API」により、モデルが会話中に得た情報の要約や重要ポイントを**ファイルに保存し、後のターンで再読込**できるようになりました [11] [36]。これによって、たとえ100kトークンの窓を超えるような超長時間対話でも、重要事項を抜き出して**長期記憶**として参照し続けることが可能です [10]。実質的に、Claudeは**自分でメモを取りながら会話**しているような状態であり、これが**文脈の連続性と推論の整合性**を支えています。

さらにClaude Opus 4の回答スタイル自体が**一貫性・信頼性重視**になるよう調整されています。AnthropicはClaude 4 Opusについて「**推論の信頼性**（reasoning reliability）を重視し、派手さより**一貫した予測可能な推論パターン**に注力している」と述べています [12]。具体的には、会話の各ターンで安易にトピックを飛躍させたり前提を覆したりせず、これまでの流れに沿って着実に推論を積み上げる傾向があります。そのため複数ターンにわたる論理的な問題解決（例えばユーザとモデルが応答を重ねながら数式を解いたりコードデバッグをするようなケース）でも、Claudeは**手順を踏んだ筋道立て**を崩さず回答し続けることが報告されています [37]。Claude 4では特に**「近道や抜け道を使った不正確な解答」を避ける傾向が強まり、3.7モデル比で65%減少**したとも言及されています [38]。この指標は、対話エージェントとしてユーザの意図を無視したり、早合点して誤答するような挙動が減ったことを示唆しており、結果的に**対話全体の論理整合性が高まった**と考えられます。

ChatGPT o3側も、対話の一貫性維持に多くの改良が加えられています。まず、大前提としてOpenAIモデルもコンテキストウィンドウが従来より拡張されているため、**長めの対話履歴を保持**できるようになりました（GPT-4の8k→32k、さらにo3で128k程度 [2]）。またモデルの**指示文理解能力**が向上しており、システムメッセージや開発者指示を通じて「会話のトーンや設定を保つ」よう制御しやすくなっています [33]。OpenAIはシステムレベルで会話の軸がぶれないような**ロール指示**を与えることを推奨しており、o3ではその指示階層の扱いが洗練されました [33]。例えば、キャラクターを演じさせたまま長く会話させる場合でも、システムメッセージで役柄を固定し、モデルが途中でキャラ崩壊しないようになっています。

さらに、前述の**o1 Proモードの長考**は、一貫した推論展開にも寄与します。o1 Proでは難しい質問に遭遇した際、モデルが内部で複数回試行錯誤したり、回答案を検討してから応答するとされています [5] [13]。このプロセスにより、**途中の推論ステップの矛盾にモデル自身が気付きやすくなる**効果が期待できます。事実、OpenAIの内部テストでo1 Proモードは通常モードに比べ**回答の信頼性が向上し包括的な応答**が得られたとされています [13]。特にデータ分析やコード生成、法的考察のような厳密さの求められる領域で有効だったと報告されています [13]。このことは、ChatGPT o3が**複数ターンにまたがる推論タスク（例:「まず下準備の計算をして、次のターンで最終回答する」等）でも内部整合性を保ちやすい**ことを示唆します。モデルが一度に答えを出さず各ターンで部分的な結論をまとめながら進めることで、前後の論理接続が崩れにくくなるわけです。

またChatGPTにはAnthropicのFiles APIに相当する機能こそ直接はありませんが、**ユーザ側で対話履歴を要約して与え直す**ことや、**関数呼び出し**で一時データを保持するような拡張も可能です。OpenAIの関数呼び出し機能は、モデルが一度「メモ内容を保存する関数」を呼び出し、返り値としてそのメモを次ターン以降に受け取るような使い方も考えられます。こうした間接的なアプローチではありますが、ChatGPTも工夫次第で**長時間の一貫対話**を実現できます。実際、企業向けChatGPTでは**会話の履歴を保存・検索**するソリューションが提供されており、前に出た議論を後からモデルに思い出させることも可能になっています。OpenAIはo3モデルで**複雑なワークフロー下でもコヒーレンスを維持する**ことを重視したとされ [3]、例えばプラグインを連続使用してレポートを生成するようなシナリオでも、**ツール間で得た情報を踏まえて矛盾なく会話**を続行できるようチューニングされています。

第三者評価では、まだ定量的な「一貫性スコア」のような指標は確立していないものの、ユーザコミュニティの報告などからヒントが得られます。一部のエンジニアは「Claudeは一度決めた方針を曲げず安定した

回答を返すが、GPT-4系列はユーザが新たな事実を提示すると柔軟に迎合して前言を訂正する傾向がある」と指摘していました。これは裏を返せば、Claudeは**会話中の自己矛盾は少ないが頑固**で、ChatGPTは**融通が利くが場合によっては前の回答と整合しなくなる**こともあるという違いです。ただし最新世代では互いに改良が進み、この差はだいぶ縮まっています。例えば前述のようにGPT-4.5(o1)では**不正確な情報に影響されにくくなった**（誤った追加情報に引っ張られて結論を変える頻度が減った）という結果が報告されており [16]、安易に前言を翻さず**芯の通った議論**を展開できるようになりつつあります。

総合的に見ると、**Claude Opus 4は巨大な記憶容量とメモ機能により長期会話での文脈維持能力が極めて高く、論理の筋道を一貫して守る傾向**があります。一方、**ChatGPT o3は計算資源を動的に使いながら各ターンで最適な応答を再検討することで矛盾を減らし、外部ツールとの複雑なやり取りでも全体を見通した整合性**を保つよう設計されています。どちらも多ターン対話での信頼性は向上しており、ユーザは長い議論や段階的問題解決を安心して任せられるレベルになっています。

モデル設計（アーキテクチャ、訓練データ、事後訓練）

Claude Opus 4とChatGPT o3は、ともにTransformerをベースとした大規模言語モデルですが、その**モデルアーキテクチャ設計や訓練・ファインチューニング手法**にはいくつかの違いがあります。

アーキテクチャの方針として、AnthropicはClaudeシリーズにおいて一貫して**シンプルで汎用的なアーキテクチャ**を保つ傾向があります。パラメータ数は公表されていませんが、第三者推定ではClaude 2が約**70億〜520億**（様々な推測があります）パラメータとの情報もあり、Claude 4ではそれを上回る規模と考えられます。Claude 4ファミリーは**Opus（大モデル）**と**Sonnet（小〜中モデル）**の2系列があり、用途に応じて使い分けられます [39] [15]。特にOpus 4は**世界最高性能のコーディングモデル**とうたわれ [40] [4]、精度向上のため計算資源を潤沢に投入した最大級のモデルです。一方Sonnet 4はやや軽量で応答速度を優先したモデルですが、両者は**統一アーキテクチャ上で動作モードを切り替えるハイブリッド設計**になっています [15]。Anthropicによれば、Opus 4とSonnet 4はいずれも**「瞬時応答モード」と「拡張思考モード」**の二つの動作モードを持ち、要求に応じて即答も深慮もこなせるよう最適化されているとのことです [15]。これは単一のモデルが内部で計算ステップを動的に調節できることを示唆しており、軽い質問には高速に答え、難問にはじっくり考えるという**柔軟な推論制御**が可能になっています。こうした設計は、近年研究されているMixture-of-Experts（専門家混合）やMulti-Modalモデルではなく、**一つの言語モデル内で制御フローを工夫する**Anthropic独自のアプローチと考えられます。

対するChatGPT o3（OpenAIのoシリーズ）も、GPT-4から進化した**新世代アーキテクチャ**を採用しているとみられます。GPT-4自体、詳細は秘匿されていますが複数モデルのアンサンブルやモジュール構造の可能性が指摘されていました。oシリーズではさらに内部構造が刷新された可能性があります。OpenAIは2024年末にChatGPT Proプラン限定で投入した**o1モデル**について「我々の最もスマートなモデル」と位置づけており [41]、従来のGPT-4（社内名称: GPT-4oと表記される）やGPT-3.5系とは異なるラインのモデルであることを示唆しています [42]。公開情報によれば、ChatGPT Proで提供される**o1 Pro**は**信頼性と計算深度に焦点を当てた顕著なアーキテクチャ改良**が盛り込まれているとのことです [43]。具体的には、モデルが**必要に応じてより多くの計算資源を使って推論する**仕組みが導入されました [5]。これはAnthropicの「2モード」と似ていますが、OpenAIの場合は特にPro版でのみ利用できる**高精度モード**として提供されています。推論途中で**複数の解候補を並行評価**し最良のものを選ぶリランキング戦略や、**ステップ数を増やした内部チェインオブソート**などが用いられているようです [44] [45]。OpenAIは詳細な手法を公開していませんが、o1 Proが標準モデルに比べ大幅に計算コストを増やしている点から、例えば**自己反省（self-reflection）ループ**や**木構造探索**的なアルゴリズムを組み込んでいる可能性もあります。

訓練データに関しては、両社とも明確な公開はしていませんが、傾向として、Anthropicは**対話データ**や**高品質テキスト**を重視し、OpenAIはそれに加えて**コードや専門分野の問題**を大量に含めていると言われます。実際、Claude 2.0の時点でAnthropicは「PythonやMLの問題でGPT-4と同等以上の性能」と述べており、コードデータも学習していますが [46]、OpenAIはGPT-4で大規模なコード問題（LeetCodeやCodeforces等）まで広

範に取り入れたとされます。その効果はベンチマークに表れており、OpenAIのoシリーズは**コーディング能力が飛躍的に向上**しました。例えば競技プログラミングの評価（Codeforces難易度）では、GPT-4で62パーセンタイルだったスコアがo1では89、o1　Proでは90にまで上昇しています [47]。同様に数学コンペ（AMCやAIMEなど）でも大きく正解率が伸びています [19]。AnthropicのClaude　4もコーディングに強く、SWE-Bench（ソフトウェアエンジニアリング問題集）で72.5%と最高水準のスコアを出しています [4]が、OpenAI o3もそれに迫る69.1%を記録しています [18]。両社とも**コード・数学などフォーマルなデータを重点的に学習**させた成果と言えます。

ファインチューニング（事後訓練）については、Anthropicは前述の**Constitutional AI**アプローチを採用しています。Claude 2のモデルカードによれば、まず人間のデータで「Helpfulness（有用さ）」を向上させ、その後**AI自身による原則評価**と**強化学習**のフェーズを経ることでHarmlessness（無害さ）とHonesty（誠実さ）を高めています [14]。このプロセスでは一部RLHF（人間フィードバックによる強化学習）も使われていますが、中心はAnthropic独自の「AIフィードバックループ」です [14]。Claudeの「憲法」はAnthropicが公開したブログ記事「Claude's　Constitution」で具体的に示されており、複数の倫理・行動原則から成ります [8]。モデルは訓練中にこの原則を参照しながら、自身の出力を自己批評・改良するようなステップを踏んでいます。この結果、Claudeは**人間の介在が少なくとも一貫した行動規範**を身につけており、極端に攻撃的な発言や露骨な誤情報の拡散を避けるようになっています [35] [48]。

OpenAIのChatGPT　o3も、引き続き**RLHFを主体**としたファインチューニングがなされていますが、そのスケールと精度は増しています。GPT-4の開発では、OpenAIは専門分野ごとに高度な知識を持つ人間アノテーターを動員し、モデルの回答を細かく評価・ランク付けして報酬モデルを訓練しました。また、より高品質な比較データを得るため、一部では**モデル自身が複数の回答案を生成し、人間がそれを比較する**手法（リコンスティテューション法）も用いているとされています。さらにOpenAIは2023年以降、**自動評価**にも取り組んでおり、GPT-4や後継モデルを使って別のモデルの出力を判定させる研究も行われました。そうしたメタフィードバックも活用しつつ、最終的にChatGPT　o3は**非常に洗練された報酬モデル**に沿って調整されているようです [9]。

特徴的なのは、OpenAIモデルは**ユーザー指示への従順さ**と**システム指示の遵守**を両立させる工夫が凝らされている点です。GPT-4ではシステムメッセージの概念が導入され、開発者が定めたルール（例えば「法律相談には絶対に具体的な法律条文を引用せよ」のような指示）をユーザ入力に優先して守るようになりました。oシリーズではこの仕組みが階層化され、**ユーザ→開発者→システム**の優先順位でモデルが指示を解釈するとの情報があります [33]。これにより、**安全性や一貫性に関わるシステムレベルの指示**（例えばトピックAには触れない、前提Bを常に維持する等）は会話が続いても貫かれます。一方でユーザの意図にも柔軟に応じ、できる範囲で要求を叶えるバランスを取っています。Anthropicモデルもプロンプト内で類似の指示を与えられますが、OpenAIモデルほど階層的に統制された印象はなく、むしろ**原則よりユーザ意思を尊重する**傾向が歴代ChatGPTには見られました [35]。もっとも、o3では上記のようにその方針も手直しされ、**必要な場合はユーザ要求より真実性・安全性を優先**する度合いが高まっています [34]。これは前述のように無害な要求まで拒否する副作用もありましたが、安全面での強化と言えるでしょう。

最後に**マルチモーダル対応**について触れると、Claude Opus 4は現時点で**テキスト専用モデル**です（画像や音声入力はサポートされていません）。一方ChatGPT　o3はGPT-4と同様に**画像入力**を受け付けたり、音声でのやりとり（OpenAI　Whisper等を活用）に対応しています [49] [50]。Zapierの比較によれば、ChatGPT側では**画像生成（OpenAI　DALL-E統合）**や**音声読み上げ**、さらには**動画生成（OpenAIのコードネーム"Sora"とされる機能）**まで備わっており、総合AIアシスタントとしての幅が広がっています [49]。Anthropic Claudeは画像生成や音声合成機能は持ちませんが、裏を返せば**テキスト対話に特化して最適化**されているとも言えます。用途によってはChatGPTの方が便利な場合（例えば画像を解析して説明させる等）もありますが、純粋なテキスト知性に関しては両モデルとも最新の技術で研鑽されています。

公開評価ベンチマーク・第三者検証

Claude Opus 4とChatGPT o3はいずれも社内・社外で多角的な評価が行われています。公開されている主な**ベンチマーク結果**や**第三者による検証**を抜粋して比較します。

- **コーディング能力:** Claude Opus 4はソフトウェアエンジニアリング分野の包括ベンチマーク**SWE-bench**で**72.5%**という最高の正解率を達成しました [4]。これは従来Claude 3.7 Sonnetの成績を大幅に上回り、同ベンチマーク上で当時最高性能だったとAnthropicは述べています [4]。一方、OpenAIのモデルも負けておらず、ChatGPT o3（GPT-4oと推定）は**69.1%**を記録しており [18]、僅差でClaudeに迫る結果となっています。中でも**OpenAI o3は前世代（GPT-4の48.9%）から大幅なジャンプ**を見せ [18]、コーディングに関して両社のしのぎ合いがうかがえます。また、プログラミングコンテストでの評価として、Claude 4は**Terminal-bench**（ターミナル操作を伴う課題集）で43.2%を達成しています [4]。OpenAI o3の同等データは公開されていませんが、OpenAIは**Codeforces（競技プログラミング）のスコア**としてo1が上位○%に相当するElo 2706を獲得したとアピールしています [51]。総じて、コード生成では**Claudeが安定した高品質コードを出力する**との評価（Cursor社「複雑なコードベース理解で飛躍的進歩」 [37] など）や、**OpenAIモデルがより難度の高い問題で突出した結果を出す**との評価（Google GeminiやxAI Grok等他社モデルとの比較でOpenAIが「生のベンチマークでリード」 [18] [51]）があり、一概にどちらが優位とは言い難い状況です。第三者の開発者ブログでは、実際に「マリオのゲーム実装」や「テトリス実装」を各モデルに依頼するテストが行われましたが、**Claude Opus 4は要求どおり完全なゲームを驚くほど素早く生成**し高く評価されたのに対し、OpenAI o3は不完全な実装やバグが見られ失望、との報告もあります [52] [53]。もちろんプロンプト設計や評価者の主観に左右される部分もありますが、**総じてClaudeは確実・丁寧なコーディング**、**OpenAI o3は難易度の高い課題での突破力**といった特徴が透けて見えます。

- **言語・知識テスト:** 学術知識や常識推論を測るMMLU（Massive Multitask Language Understanding）において、Claude Opus 4は**87.4%**（一部報告では85.2% [45]）の精度を示しました。これはGPT-4が公開していた約86.4%に匹敵し、Claude 3.7（約80%）から大幅向上しています。ChatGPT o3（およびo1 Pro）はMMLUに関する直接の公表値はありませんが、OpenAIはPhDレベルの科学質問集**GPQA Diamond**での成績を公開しており、それによるとGPT-4→o1→o1 Proで**74→76→79%**へ改善したとのことです [20]。Claude Opus 4もGPQA Diamondで**74.9%**（拡張思考なし）とほぼ同等のスコアを残しており [17]、この分野では互角と言えます。難易度の高い数学試験AIMEでは、Claude Opus 4は**33.9%**に留まりました [45] が、OpenAI o1 Proは**86%（pass@1）**を達成したとされ [19]、数学分野ではOpenAIモデルがリードしているようです。実際Mediumの比較記事でも、**数学的分析能力ではGemini 2.5 Proが84.0%でトップ、次点にOpenAI O3、Claude 4は安定志向**と評されています [54]。総じて、知識問題では両モデルともトップクラスの成績ですが、**高度数学ではOpenAIがやや優勢、幅広い常識的知識では互角〜Claude健闘**と見ることができます。

- **安全性・倫理評価:** 両社とも安全性レポートを公開しており、それによれば新モデルになるほど有害出力の削減に成功しています。Claude 2のモデルカードでは、有害回答（HHH評価）の識別能力が世代ごとに向上しているとされ [55]、Claude 2は前世代より一貫してHHHで高スコアを示しました [55]。OpenAIもGPT-4システムカードで、GPT-3.5比で不適切発言が大幅減少したと述べています。またRed Team（侵入テスト）的な評価では、前述のVirtue AI分析で**Claude 3.7はEU AI法への適合性でGPT-4.5より優れる**、**GPT-4.5はプライバシー攻撃耐性で優れる**等、細かな強み弱みが分析されています [56] [57]。Claudeはユーザの無害なリクエストを**不必要に拒否しない**点でUXが良い反面 [35]、GPT-4.5は慎重すぎるきらいがあるなど [35]、安全と利便のせめぎ合いも評価されています。AnthropicはClaude Opus 4のリリースに際し、安全基準を**ASL-3（AI Safety Level 3）**に引き上げたと発表しました [58] [59]。これは核・生物兵器等に関する悪用リスクが無視できないほどモデル能力が向上した可能性に備えた措置で、Claude 4が**危険な悪用にも対策が必要な水準**に達したことを示しています [59]。OpenAIもGPT-4で類似の安全対策強化（モデルへの匿名アクセス制限や、危険用途へ

のガードレール）を講じており、総じて**両モデルとも前世代以上に強力であるがゆえに安全対策も厳格化**されている状況です [60] [61] 。

以上のように、多角的な評価を総合すると**Claude Opus 4とChatGPT o3はいずれも現時点で世界トップクラスの性能を持つLLM**であり、それぞれに強みがあります。Claudeは**長大なコンテキスト処理や安定した推論**に秀で、**確実性・信頼性重視**の設計が光ります。一方、ChatGPT　o3は**高度な推論力やマルチモーダル対応**に優れ、**挑戦的な問題への打開力**が目立ちます。両社の公式資料や第三者の検証から得られる知見を踏まえれば、用途や場面によって最適なモデルは異なり得るものの、いずれを選んでも従来を大きく凌駕するAIアシスタントとして機能することは間違いないでしょう。

参考文献・情報源: 本比較はAnthropicおよびOpenAIの公式発表（システムカード、モデルカード、技術ブログ） [1] [5] 、ならびに信頼性の高い第三者による分析 [23] [34] に基づいています。それぞれの出典は文中に【】で示しています。

---

[1] [8] [14] [21] [24] [55] anthropic.com
https://www.anthropic.com/claude-2-model-card

[2] [42] [49] [50] Claude vs. ChatGPT: What's the difference? [2025]
https://zapier.com/blog/claude-vs-chatgpt/

[3] [12] [18] [22] [23] [51] [54] The AI Arms Race: Claude 4 Opus vs Gemini 2.5 Pro vs OpenAI O3 vs Grok 3 | by Cogni Down Under | May, 2025 | Medium
https://medium.com/@cognidownunder/the-ai-arms-race-claude-4-opus-vs-gemini-2-5-pro-vs-openai-o3-vs-grok-3-99a6f598b560

[4] [6] [10] [15] [17] [25] [37] [38] [39] [40] [44] [45] Introducing Claude 4 \ Anthropic
https://www.anthropic.com/news/claude-4

[5] [13] [19] [20] [41] [47] Introducing ChatGPT Pro | OpenAI
https://openai.com/index/introducing-chatgpt-pro/

[7] [11] [36] New capabilities for building agents on the Anthropic API \ Anthropic
https://www.anthropic.com/news/agent-capabilities-api

[9] [16] [33] [34] [35] [48] [56] [57] GPT-4.5 vs Claude 3.7 - Advanced Redteaming Analysis - Virtue AI
https://www.virtueai.com/2025/03/01/gpt-4-5-vs-claude-3-7-advanced-redteaming-analysis/

[26] Galileo's Hallucination Index and Performance Metrics Aim for Better ...
https://synthedia.substack.com/p/galileos-hallucination-index-and?utm_source=substack&utm_medium=email&utm_content=share&action=share

[27] [28] Auto-Evaluation of Anthropic 100k Context Window
https://blog.langchain.dev/auto-evaluation-of-anthropic-100k-context-window/

[29] [30] [31] Study suggests that even the best AI models hallucinate a bunch | TechCrunch
https://techcrunch.com/2024/08/14/study-suggests-that-even-the-best-ai-models-hallucinate-a-bunch/

[32] Leaderboard: OpenAI’s GPT-4 Has Lowest Hallucination Rate
https://aibusiness.com/nlp/openai-s-gpt-4-surpasses-rivals-in-document-summary-accuracy

[43] o1 vs o1 pro: Is it worth upgrading and spending $200?
https://www.leanware.co/insights/gpt-models-comparison-insights

[46] Using Anthropic: Best Practices, Parameters, and Large Context ...
https://www.prompthub.us/blog/using-anthropic-best-practices-parameters-and-large-context-windows

52 53 Claude Opus 4 vs. Gemini 2.5 Pro vs. OpenAI o3 Coding Comparison - DEV Community
https://dev.to/composiodev/claude-opus-4-vs-gemini-25-pro-vs-openai-o3-coding-comparison-3jnp

58 59 60 61 Activating AI Safety Level 3 Protections \ Anthropic
https://www.anthropic.com/news/activating-asl3-protections