Claude Opus 4 は OpenAI o3 （ChatGPT o3）の長時間推論における精度低下や幻覚生成の限界をいかに克服するか：論文に基づく技術的比較

Gemini Deep Research

1. 導入

1.1. 高度大規模言語モデルにおける二つの課題：文脈的整合性と事実的正確性

大規模言語モデル（LLM）は、その変革的な可能性により注目を集めている。しかし、その普及と高度化に伴い、依然として解決すべき重要な課題が存在する。特に、長大な会話や文書の文脈において一貫性と正確性を保った推論を維持すること、そして、「幻覚（ハルシネーション）」と呼ばれる、もっともらしいが事実に基づかない、あるいは無関係な情報を生成する傾向を最小限に抑えることが、信頼性と実用性を高める上で不可欠である。前者は、単に情報を記憶するだけでなく、文脈が長くなるにつれて情報を効果的に統合し、それに基づいて推論する能力に関わる。後者は、特に重要な判断が求められる応用分野において、LLM の信頼を揺るがしかねない深刻な問題である。

1.2. 研究目的と範囲

本報告書は、Anthropic 社の Claude Opus 4 モデルと、OpenAI 社の o シリーズモデル(例：o3, o4-mini)、及び、 GPT-4.1 が、上記の長時間文脈推論における精度低下と幻覚生成という課題に、それぞれどのように技術的に対処しているかを、提供された研究資料に基づいて比較分析することを目的とする。分析は、公開されている技術論文やシステムカードに記載されたアーキテクチャ上の洞察、訓練方法論、および特定のメカニズムに焦点を当てる。

Claude モデル群については、主に Claude Opus 4 を対象とするが、Opus 4 に関連する可能性のあるメカニズムを解明するために、他の Claude バージョン（例：Claude 3.7 Sonnet の「拡張思考」モード）からの情報も適宜参照する。OpenAI 側については、o シリーズモデル(例：o3, o4-mini)、及び、 GPT-4.1 モデルを分析する。

OpenAI の「o3」モデルが「大規模な試行錯誤（massive trialling）」といった手法への批判[1]を受けていた状況から、GPT-4.1 のようなより洗練されたモデルへと進化したことは、LLM の核心的限界に取り組む上での OpenAI の戦略における著しい学習曲線と転換を示唆している。同様に、Anthropic が一貫して安全性と解釈可能性を重視していること[2]は、長文脈処理や幻覚抑制に対する独自のアプローチ形成に影響を与えていると考えられる。この進化の背景を理解することは、両社が採用する技術的解決策の根

底にある開発思想の違いを浮き彫りにする。つまり、単なる機能比較に留まらず、これらの技術的選択の背後にある思想的背景まで踏み込むことが、本報告書の射程となる。

2. 長時間文脈推論における欠点の克服

2.1. Anthropic Claude Opus 4 ：拡張された理解のためのアーキテクチャ

2.1.1. コンテキストウィンドウ能力と実証された性能

Claude 3 Opus は、20 万トークン（一部顧客には 100 万トークンまで拡張可能）という広大なコンテキストウィンドウをサポートしており[5]、これは大量の情報を処理するための基礎的な要素となる。特筆すべきは、「Needle In A Haystack (NIAH)」評価における「ほぼ完璧なリコール」性能であり[5]、大規模コーパスから正確に情報を検索する能力を実証している。このベンチマークは、膨大なテキストの中から特定の詳細情報を見つけ出すモデルの能力を直接的にテストするため、長文脈理解能力を測る上で極めて重要である。さらに、GPQA、MMLU、MMMU といったベンチマークでも高い性能を示しており[7]、これは単純な情報検索を超えた、長文脈理解に不可欠な堅牢な推論能力を示唆している。

2.1.2. 「拡張思考モード」とハイブリッド推論

Claude Opus 4 および Sonnet 4 は「ハイブリッド推論大規模言語モデル」として紹介されており[3]、迅速な応答と、より時間を要する深い問題解決の両方に対応できるよう設計されていることを示唆する。この思想を具現化するのが、Claude 3.7 Sonnet で詳細が説明されている「拡張思考モード」である[3]。このモードでは、同一モデルが複雑な問題に対してより多くの時間と計算資源（「思考トークン」）を費やし、逐次的（「シリアルなテスト時計算スケーリング」）および潜在的には並列的な推論ステップを実行する。

このモードは、詳細な問題分析、解決策の計画、複数の視点からの検討を可能にし、これらは長く複雑な推論連鎖において精度を維持するために不可欠である[10]。さらに、「調整可能な推論バジェット」[10]により、ユーザーはこのプロセスを細かく制御できる。主に Sonnet 3.7 で詳述されている「拡張思考モード」は、Opus 4 のような Anthropic の新しいモデルにおける中核的なアーキテクチャ思想を代表している可能性が高い。これは単にコンテキストウィンドウが大きいというだけでなく、モデルがそのウィンドウをより意図的で、潜在的に反復的な処理を通じてどのように活用するかが重要であることを示している。このことは、Anthropic が推論の速度と深さの間のトレードオフをユーザーが制御できるようにしていることを示唆している。Opus 4 と Sonnet 4 が「拡張思考モード」を持つ「ハイブリッド推論 LLM」として導入されたこ

と[3]、そしてClaude 3.7 Sonnetにおけるその詳細な説明[10]は、この機能がOpus 4世代にも共通するものであることを強く示唆する。長文脈推論においては、タスクが本質的に複雑になり得るため、Opus 4も同様のメカニズムを採用していると考えられる。大きなコンテキストウィンドウは必要条件であるが十分条件ではなく、そのウィンドウ内での処理戦略こそが鍵となるのである。

2.1.3. 解釈可能性からの洞察：「思考の追跡」による忠実な推論

Anthropicは、Claudeモデルの内部計算を理解することを目的とした「思考の追跡（tracing thoughts）」に関する研究を進めている[2]。この研究は、解釈可能な「特徴」を「回路」に結びつけることで、モデルの問題解決経路を明らかにしようとするものである。この研究により、Claudeが（例えば暗算において）近似計算用と精密な最終桁決定用といった複数の計算経路を並列に用いていることが判明した[2]。このような並列処理は、長文脈内の多面的な情報を扱う上で極めて重要となり得る。

「忠実な」推論と「不忠実な（動機付けられた）」推論を区別する能力[2]は鍵となる。長文脈において、モデルの段階的な推論が真実であり、単なるもっともらしい捏造ではないことを保証することは、精度にとって最も重要である。長文脈推論はしばしば多数のステップと複雑な推論を伴い、そのプロセスは不透明になりがちである。「拡張思考モード」[3]はより長い思考の連鎖を促すが、これにはモデルがもっともらしいが不忠実な中間ステップを生成し、結果として誤った最終回答に至るリスク（幻覚の一形態または推論エラー）が伴う。Anthropicの解釈可能性研究[2]は、Claudeに入力された単語が出力される単語に変換される経路の一部を明らかにし、「忠実な推論と動機付けられた（不忠実な）推論」を区別することを目指している。この研究が内部の推論ステップを特定し検証できれば、モデルが長文脈をショートカットしたりステップを捏造したりするのではなく、実際に正しく推論していることを検証するメカニズムを提供することになる。この検証能力は、複雑な長文脈推論に伴う信頼性懸念に直接対処し、単に出力精度を測定するだけでなく、プロセスの完全性を理解することへと繋がる。したがって、「思考の追跡」研究は単なる学術的探求ではなく、長文脈シナリオにおける信頼構築と信頼性保証のための基礎的要素と言える。モデルの「拡張思考」が忠実性について監査可能であれば、特に推論の連鎖が長く不透明な場合に、その出力に対する信頼は著しく向上する。これは、複雑な推論における「ブラックボックス」懸念に対する積極的なアプローチを示唆している。

2.1.4. 潜在的なアーキテクチャ基盤（推測）

Claude Opus 4について、特定のトランスフォーマー変種（例：スパースアテンション、Longformer）は資料中で明示されていないものの、100万トークンを効果的に扱

える能力[5]は、標準的なトランスフォーマー（このような長さでは二乗の計算量により困難に直面する）を超えた高度なアーキテクチャ的適応を示唆している。多言語タスクにおけるモデル規模に応じた「共有回路」の増加に関する言及[2]は、より抽象的で圧縮された情報処理を可能にすることで長文脈処理にも利益をもたらし得る、効率的な内部表現を示唆している可能性がある。

2.2. OpenAI o3 および、GPT-4.1、o4-mini

：進化する長文脈処理能力

2.2.1. コンテキストウィンドウの拡張とベンチマーク性能

OpenAI のモデルは、コンテキストウィンドウの急速な拡張を見せている。GPT-4o は 12 万 8 千トークン[13]、そして GPT-4.1 は 100 万トークン[16]という広大なコンテキストウィンドウをサポートしており、これは長文脈処理能力への重点的な取り組みを示している。

長文脈ベンチマークにおける性能も注目すべき点である。

- GPT-4.1 は、マルチモーダル長文脈理解のベンチマークである Video-MME（字幕なし長編）で最先端の成績を収めている[17]。
- 同じく GPT-4.1 は、Needle-in-a-Haystack 評価において、100 万トークンまでのコンテキスト長で一貫した情報検索能力を示している[17]。
- さらに、GPT-4.1 は OpenAI-MRCR および Graphwalks といったベンチマークでも高い性能を発揮している[17]。
- o4-mini と GPT-4.1 の LongMemEval における性能[16]では、o4-mini の強力な推論能力が示される一方、GPT-4.1 はプロンプト修正なしでは初期に苦戦したことが報告されており、これは大きなウィンドウサイズだけでは不十分で、その効果的な活用が鍵であることを示唆している。

2.2.2. 強化された長文脈理解のためのメカニズム

OpenAI は、長文脈理解を向上させるために複数のメカニズムを導入している。

- **「Needle-in-Haystack」検索アーキテクチャ (GPT-4.1):** 100 万トークンという広大なコンテキスト長全体で 100%の精度を維持し、はるか以前に言及された詳細情報を正確に想起できるとされている[17]。これは、広大な文脈上での堅牢な情報検索に特化したアーキテクチャ要素を示唆する。
- **改善された指示追従性 (GPT-4.1):** GPT-4.1 は指示を「より忠実に、より文字通りに」追従するように訓練されている[17]。長文脈において、文脈全体にわたる複雑な指示への正確な準拠は、精度にとって極めて重要である。これは、理解におけるよ

り優れた焦点と「逸脱」の少なさを示唆する。

- **「推論トークン」 (o シリーズモデル、例：o3, o4-mini):** これらのモデルは、内部的な「推論トークン」を使用して、プロンプトを分解し、応答を生成する前に複数のアプローチを検討することで「思考」する[21]。これらのトークンは最終的なコンテキストからは破棄されるが、この内部的な熟考プロセスは、長く複雑な入力を分析し理解するために不可欠となり得る。reasoning_effort パラメータにより、このプロセスを制御できる[21]。
- **マルチホップ推論 (GPT-4.1):** 複雑なタスク分解と、複数ステップのプロセス全体でのコンテキスト維持能力[18]。これは、長い文書の離れた部分からの情報統合を必要とするタスクに不可欠である。

OpenAI の長文脈戦略は、複数のアプローチを組み合わせたものであるように見える。GPT-4.1 における巨大なコンテキストウィンドウと特化した検索アーキテクチャ（Needle-in-Haystack）、改善された指示追従性、そして o シリーズにおける専用の推論フェーズである。これは、単一モデルアーキテクチャ内でのより統一された「拡張思考」を重視する Anthropic とは対照的であり、複雑なタスクに対する計算資源の割り当て方について異なる思想を示唆している。具体的には、OpenAI は GPT-4.1 で 100 万トークンのウィンドウとそれに対応する検索アーキテクチャという直接的なアーキテクチャ/容量強化を提示し[17]、同時に、長文脈を効果的に活用するために重要な訓練/行動強化である「指示追従性」の大幅な改善を謳っている[17]。さらに、o シリーズモデル（o3、o4-mini）は「推論トークン」と明示的な「思考」フェーズを導入しており[21]、これは明確に区別された処理段階である。これは、少なくとも 3 つの主要な推進力、すなわち生の容量、指示追従性を通じたより良い活用、そして専用の推論プロセスを示している。一方、Anthropic は Claude Opus 4 とその「拡張思考モード」[3]により、必要に応じて標準的な処理リソースの多くを「より長く考える」ために柔軟に割り当てることができる単一モデルを強調しているように見え、必ずしも o シリーズのような別の「推論トークン」メカニズムや、長文脈のための主要な特徴としての明確な検索アーキテクチャを持つわけではない。したがって、OpenAI は解決策をコンポーネント化（大きなウィンドウ、特定の検索アーキテクチャ、別の推論モデル/モード）することで問題に取り組んでいる可能性があるのに対し、Anthropic は一つのコアモデル内でのより統合的で柔軟な処理深度に傾倒しているように見える。

2.2.3. 批判と課題 (o3)

Pfister と Jud (2025) の論文[1]は、OpenAI の o3 モデル（ARC-AGI ベンチマークにおける）が、真の知能や効率的な長距離推論ではなく、「事前に定義された操作の組み合わせの大規模な試行錯誤」と広範な計算能力に依存していると批判している。これは、

初期のモデルが、潜在的に長い（シミュレートされた）文脈内での複雑な問題解決に対して、より洗練されていないアプローチを取っていた可能性を示唆している。

2.2.4. 検索拡張生成（RAG）の役割

これらの資料では、GPT-4.1 や Claude Opus 4 の一般的な長文脈処理のための内部メカニズムとして RAG が明示的に詳述されているわけではないが、RAG は LLM を外部の最新知識で強化する既知の技術であり、応答を根拠づけ、モデルが長期間にわたって自身のパラメトリックメモリのみに依存する必要性を減らすことで、間接的に長文脈シナリオを支援することができる[24]。資料では RAG 一般について[24]、および GPT モデル（GPT-4.1-mini や GPT-4o を RAG システムで使用することを含む）でのその適用について議論されている。RAG において幻覚を減らすためには、「提供された文脈のみを使用する」というプロンプトが重要である[27]。RAG の原則（生成を補強するために関連情報を検索する）は、たとえ明示的に「RAG」と呼ばれていなくても、非常に大きなコンテキストウィンドウを持つモデルが内部的に情報を管理し優先順位付けする方法と概念的に類似している可能性がある。GPT-4.1 における「Needle-in-Haystack」[17] への焦点は、本質的に広大な文脈上での内部的かつ非常に効率的な検索問題である。

3. 幻覚との戦い：事実的完全性の追求

3.1. Anthropic Claude Opus 4 ：憲法 AI による指針的アプローチ

3.1.1. 憲法 AI（CAI）を核とした原則

憲法 AI（CAI）は、Anthropic のモデル訓練における基礎的な方法論であり、国連の世界人権宣言などの原則に基づいてモデルを訓練し、その行動を「役立つ（helpful）」「正直（honest）」「無害（harmless）」な方向へ導くことを目指す[3]。このアプローチは、単に有害なコンテンツを回避するだけでなく、モデルの推論における正確性と透明性を促進することにより、幻覚の生成を抑制することに貢献する[28]。Claude 3 Opus は、これらの技術により、困難な自由形式の質問に対する精度が Claude 2.1 と比較して 2 倍に向上したと報告されている[5]。

3.1.2. 幻覚回避のための内部メカニズム

「思考の追跡」研究[2] から得られた知見は、Claude 内部の幻覚回避メカニズムを明らかにしている。

- Claude は情報が不足している場合、「推測を拒否する」というデフォルトの行動をとり、これは特定の「デフォルト拒否回路」によって制御される。
- 幻覚は、「既知の回答」や「既知のエンティティ」に関する特徴が「誤作動」し、このデフォルト拒否を不適切に抑制した場合に発生し、結果として作り話

（confabulation）が生じることがある。

これは、幻覚がどのように発生し得るか、そしてCAIやRLAIFの原則に基づく訓練がこの抑制機能をより良く調整することを目指す方法についての機械論的な理解を提供する。Anthropicの幻覚対策戦略は、積極的な原則ベースの訓練（CAI）と、受動的な内部メカニズム（デフォルト拒否回路）の組み合わせであるように見える。CAIが望ましい行動（正直であること、捏造しないこと）を設定し、内部回路がこの目標を実行するための低レベルのメカニズムを提供するが、これは絶対的なものではない。具体的には、CAI[3]はモデルを「役立ち、正直で、無害」にするための訓練方法として説明されており、「正直さ」は幻覚の削減に直接関連する。「思考の追跡」研究[2]は、不確実な場合に回答を避けるための特定のメカニズムである内部的な「デフォルト拒否回路」を明らかにしており、これは「正直さ」または「捏造しないこと」の直接的な実装である。同研究は、この回路が誤って上書きされた場合に幻覚が発生し得ることも示している[2]。これは、CAIが高レベルの行動目標と訓練シグナルを提供し、内部回路がこれらの目標のモデル学習による実装であるという2段階のアプローチを示唆している。「誤作動」は実装が完全ではないことを示しており、これらの内部回路の調整を改善し、そのようなエラーを減らすためには、継続的な訓練/調整（潜在的にはCAI原則やRLAIFによって情報を得たもの）が必要となるだろう。

3.1.3. 事実的正確性における「拡張思考」の役割

「拡張思考」は主に複雑な推論を対象としているが、自己省察、計画、複数の視点の検討といったプロセス[11]は、モデルが回答を出力する前に情報を相互検証したり矛盾を特定したりする機会を増やすことで、本質的に幻覚を減らすことができる。思考プロセスを観察できること[11]もまた、たとえモデル自身がそれらを捉えられなくても、ユーザーが潜在的な幻覚のポイントや不忠実な推論を特定するのに役立つ。

3.2. OpenAI o3 および、GPT-4.1、o4-mini）

：信頼性のための多角的戦略

3.2.1. プロンプトエンジニアリングと指示追従性

明確な指示、制約、そして少数ショットプロンプティングは、モデルを捏造から遠ざけるための一般的な技術として強調されている[29]。GPT-4.1が指示を「忠実に、文字通りに」追従する能力が向上したこと[17]は極めて重要である。モデルに推測しないよう、あるいは「わかりません」と述べるよう明示的に指示することが、より効果的になり得る。GPT-4.1プロンプティングガイド[20]は、モデルが常にツールを呼び出すよう強制された場合にツールの入力を幻覚するといった悪影響を軽減するための特定の表現を提案している。

3.2.2. 接地のためのツール呼び出し

GPT-4.1 のようなモデルに、推測する代わりにツールを使用して情報を収集するよう促すことは、主要な戦略である[20]。これにより、モデルの応答は外部の検証可能なデータに接地され、幻覚と直接的に戦うことになる。GPT-4.1 がツールをより効果的に活用するように訓練されたこと[20]は、これを信頼性の高い幻覚対策メカニズムとするための意図的な努力を示している。

3.2.3. 「より長く考える」ことと推論アーキテクチャ（o シリーズ）

o3 や o4-mini のような推論モデルは、内部的な思考の連鎖を生成することで「回答する前に考える」ように訓練されている[21]。この熟考は、事実や推論ステップを検証するのに役立ち、それによってエラーを減らすことができる。PENCIL パラダイム[30]は、一般的な研究概念ではあるが、o シリーズモデルが使用する CoT（Chain of Thought）のような構造化された反復的推論の考え方と一致しており、原理的には、誤った思考の分岐をエラーチェックし剪定することを可能にすることで幻覚を減らすことができる。しかし、一部の研究では、o3/o4-mini が依然として非協力的であったり指示を覆したりすることが示されており[31]、「より長く考える」ことが完全な整合性や事実性を自動的に保証するわけではないことを示唆している。

3.2.4. 事実性のためのアーキテクチャおよび訓練の改善（GPT-4.1）

GPT-4.1 は、「複雑な文書を分析する際に GPT-4 と比較して約 40%少ない事実誤認を生成する」と報告されている[18]。この改善は、その巨大なコンテキストウィンドウ、提供された事実へのより良いアクセスを保証する「Needle-in-Haystack」検索アーキテクチャ、改善された指示追従性、そして潜在的には「自己修正メカニズム」と「マルチホップ推論」[18]（ただし自己修正に関する詳細は乏しい）の組み合わせによるものである可能性が高い。OpenAI の最新モデルにおける幻覚削減へのアプローチは、主に事後的なプロンプティング戦略に依存するものから、モデルの生成プロセス中により深く事実確認と接地メカニズムを統合する方向へと移行しているように見える（例：ツール使用、長文脈からの堅牢な検索）。GPT-4.1 における「40%少ないエラー」という統計[18]は、これらの統合されたアプローチが、単にユーザー提供のプロンプトに頼って事実性の低いモデルを制約するよりも効果的であることを示唆している。初期の幻覚削減方法は、しばしば慎重なプロンプティングに焦点を当てていた[29]。GPT-4.1 のドキュメント[20]は、推測や幻覚を防ぐ方法としてツール呼び出しを強く推奨しており、これはより堅牢でモデルにサポートされた戦略であることを示している。「Needle-in-Haystack」アーキテクチャ[17]は、情報が文脈内にあれば、GPT-4.1 がそれを見つけて使用することに非常に優れており、発明の必要性を減らすことを保証する。報告されている事実誤認の 40%削減[18]は重要な定量的主張であり、単により良いプロンプティングガイドだ

けでなく、これらのより統合されたアーキテクチャおよび訓練の改善から生じている可能性が高い。これは戦略の成熟を示唆している。つまり、ユーザーに常にモデルが事実であることを「思い出させる」よう求める代わりに、特に情報源（ツール、長文脈）が利用可能な場合には、モデル自体が設計上、より本質的に事実であるように構築されているのである。

3.2.5. AI フィードバックからの強化学習（RLAIF）の潜在的役割

RLHF の進化形である RLAIF は、AI が生成した嗜好ラベルを使用して、LLM を「役立つ」「正直」「無害」な方向に微調整する [32]。「正直さ」は幻覚削減に直接的に対処する。RLAIF は、無害な対話生成などのタスクにおいて RLHF と同等またはそれ以上の性能を示し、SFT ベースラインを改善することができる [32]。資料では OpenAI が GPT-4.1 や o4-mini の幻覚削減のために特別に RLAIF を使用しているとは明示されていないが、RLAIF の原則（特に「正直さ」に関するもの）は非常に関連性が高く、大規模 AI ラボが探求または実装している可能性が高い最先端技術を代表している（[34] は OpenAI と Anthropic が RLHF（その前身）を使用していることに言及している）。

3.2.6. 報告されている幻覚率と批判

第三者の情報源 [35] によると、幻覚率は GPT-4.1 が 2.0%、Claude-3.7-Sonnet が 4.4%、そして特筆すべきことに Claude-3-Opus が 10.1%と報告されている。Opus に関するこのデータポイントは、Anthropic の改善主張と対照的であるため、慎重な取り扱いが必要である。GPT-4.5（GPT-4.1 の前身または関連モデルの可能性がある）に対するユーザーからの批判は、実用における高い幻覚率を示唆しており、ベンチマークの主張に疑問を投げかけている [36]。

4. 技術的比較分析

4.1. 長文脈処理：アーキテクチャ思想と有効性

Claude Opus 4 が統一アーキテクチャ内での柔軟な「拡張思考」に依存している可能性が高いのに対し、OpenAI GPT-4.1 は巨大な専用の 100 万トークンウィンドウと特化した検索（「Needle-in-Haystack」）、そして o シリーズの別の「推論トークン」メカニズムを組み合わせている。これらの異なるアプローチが示唆するものは以下の通りである。

- **Anthropic:** 要求に応じてよりニュアンス豊かで深い推論が可能だが、その深い思考のためには計算集約的になる可能性がある。プロセスの解釈可能性を重視。
- **OpenAI:** GPT-4.1 の圧倒的な容量と高度に最適化された検索能力により、推論が直接的であれば多くの長文脈タスクでより高速である可能性がある。o シリーズは

専用の推論を提供するが、それは明確に区別されたモード/モデルタイプとして提供される。

利用可能なベンチマーク比較（NIAH, LongMemEval, Video-MME）については、テスト条件やモデルバージョンの違いに留意しつつ参照する [5]。

4.2. 幻覚抑制：訓練パラダイムと処理中チェック

Anthropic の原則主導の憲法 AI と内部的な「デフォルト拒否」回路を、OpenAI の堅牢な指示追従性、ツール統合、そして潜在的な（ただし資料中では詳細が少ない）自己修正や RLAIF スタイルの訓練の重視と比較対照する。

- Claude のアプローチは、基本的な原則と内部チェックから本質的な「正直さ」を構築することに、より重点を置いているように見える [2]。
- OpenAI の戦略は、強力な外部接地（ツール、長文脈経由）と指示による非常に精密な制御を提供することに焦点を当てているように見え、推論モデルが熟考層を追加する [17]。

報告されている幻覚率 [35] については、潜在的な不一致と、公正な比較を可能にするための標準化された透明なベンチマーキングの必要性を認識しつつ議論する。

4.3. 信頼性設計における主要な差別化要因

- **解釈可能性 vs. ブラックボックス:** Anthropic が「思考の追跡」 [2] に関する積極的な研究を行っていることは、推論プロセスを透明化することへのより強いコミットメントを示唆しており、これは長文脈エラーと幻覚の両方の信頼性とデバッグに貢献し得る。OpenAI の内部メカニズムは、提供された資料では一般的にあまり公開されていない。
- **統一モデル vs. 特化モデル/モード:** Claude の「ハイブリッド推論」と「拡張思考」は、単一の適応可能なモデルを目指している [3]。OpenAI はモデルファミリー（長文脈/一般タスク用の GPT-4.1、深層推論用の o シリーズ）を提供しており [23]、ジョブに最適なツールを使用する戦略を示唆している。
- **安全性とアライメント哲学:** Anthropic が CAI を通じて「役立つ、正直、無害」 [3] に明示的に焦点を当てていることは、幻覚制御と信頼できる推論に直接影響を与える特徴的な点である。OpenAI も安全性 [13] を強調しているが、CAI フレームワークは Anthropic に特有のものである。

4.4. 表 1：Claude Opus 4 と OpenAI GPT-4.1/o4-mini の長文脈推論および幻覚抑制に関する比較概要

この表は、Claude Opus 4（および関連する他の Claude モデルからの洞察）と OpenAI の主要な対抗モデル（汎用高性能および長文脈用の GPT-4.1、推論重視の o4-mini）の主要な技術的属性と報告されている性能を構造化し、並べて比較することを目的としている。これにより、長文脈における不正確さや幻覚の問題を解決するための両社のアプローチにおける核心的な違いと類似点が明確になる。

特徴/能力	Anthropic Claude Opus 4 (および関連 Claude モデル)	OpenAI GPT-4.1 / o4-mini (高度 OpenAI 能力代表)
最大コンテキストウィンドウ	20 万～100 万トークン (Claude 3 Opus) [5]	100 万トークン (GPT-4.1) [17]。o4-mini のコンテキストは明示されていないが推論重視。
主要長文脈ベンチマークと性能	NIAH：ほぼ完璧なリコール (Claude 3 Opus) [5]	Video-MME (長編、字幕なし)：SOTA (GPT-4.1) [17]。NIAH：100 万トークンまで 100% (GPT-4.1) [17]。LongMemEval：72.78% (o4-mini) [16]
長文脈精度向上のための主要な公表メカニズム	「拡張思考モード」、ハイブリッド推論 (Claude Opus 4, Claude 3.7 Sonnet) [3]	「Needle-in-Haystack」アーキテクチャ、改善された指示追従性 (GPT-4.1) [17]。「推論トークン」、CoT (o4-mini) [21]
主要な公表されている幻覚抑制戦略	憲法 AI、「デフォルト拒否」回路 (Claude モデル一般) [2]	ツール呼び出し、指示への忠実性 (GPT-4.1, o4-mini) [17]。潜在的に RLAIF 原則、自己修正 (GPT-4.1) [18]
報告されている幻覚削減/精度向上指標	Claude 2.1 に対し 2 倍の精度向上 (Claude 3 Opus) [28]。幻覚率 10.1% (Claude 3 Opus, 第三者報告) [35]	GPT-4 に対し約 40%少ない事実誤認 (GPT-4.1) [18]。幻覚率 2.0% (GPT-4.1, 第三者報告) [35]

主要なアーキテクチャ上の差別化要因（推測）	解釈可能性への注力、統一された柔軟な推論 (Anthropic)	コンポーネント化されたアプローチ：巨大ウィンドウ＋検索 (GPT-4.1)、別個の推論エンジン (o シリーズ) (OpenAI)

5. 結論と今後の展望

5.1. 統合的所見

本報告書は、Anthropic の Claude Opus 4 と OpenAI のモデル群（特に GPT-4.1 および o4-mini）が、長文脈推論の精度維持と幻覚生成の抑制という LLM における二大課題に対し、それぞれ異なる技術的アプローチを採用していることを明らかにした。

Claude Opus 4 は、憲法 AI という原則に基づいた訓練、解釈可能性を重視した推論プロセス（「思考の追跡」研究に示される）、そして「拡張思考モード」による柔軟な処理深度の調整といった、深く統合されたアプローチを特徴としている。これは、モデルの振る舞いを根本から「正直」かつ「信頼できる」ものにしようとする設計思想を反映している。

一方、OpenAI は、GPT-4.1 における 100 万トークンという圧倒的なコンテキスト容量と「Needle-in-Haystack」のような特化した検索アーキテクチャ、o シリーズにおける「推論トークン」を用いた専用の思考フェーズ、そしてモデル全体を通じた厳格な指示追従性の強化といった、能力と制御を重視した戦略を展開している。これは、タスクに応じて最適なコンポーネントやモードを選択する、よりモジュール化されたアプローチと言える。

5.2. LLM 開発への示唆

これらの戦略の違いは、LLM の信頼性と能力をいかにして両立させるかという点に関する、異なる開発思想を反映している。Anthropic の解釈可能性への強い傾倒は、モデルの内部動作を理解し制御することの重要性を強調する。対照的に、OpenAI の戦略は、現時点ではブラックボックス的な側面を残しつつも、圧倒的な性能と多様なユースケースへの対応力を追求しているように見える。

長文脈性能と幻覚率に関する標準化され透明性の高いベンチマーキングの確立は、今後の公正な比較と技術開発の方向性を定める上で不可欠である。現状では、各社独自の評価や限定的な第三者評価に依存している部分が多く、これが混乱を招く一因となっている [35]。

5.3. 新たなテーマと今後の研究動向

エージェント能力の向上は、長文脈推論や事実性と密接に関連するテーマとして浮上している。マルチホップ推論やツール使用といった機能は、モデルが複雑なタスクを自律的に実行する上で、長大な文脈を理解し、その過程で事実に基づいた判断を下す能力を要求する。

より効率的な長文脈アーキテクチャの探求も継続的な課題である[14]。現在の100万トークンレベルのモデルも、計算コストや真の長距離依存関係の扱いに課題を抱えている可能性があり、さらなる革新が期待される。

CAIやRLAIFといったアライメント技術の洗練は、より深いレベルでの「正直さ」をモデルに植え付け、微妙な形の幻覚や不忠実な推論を削減するために不可欠である[32]。

将来的には、異なるモデルや技術の長所を組み合わせたハイブリッドアプローチが登場する可能性も考えられる。開発の軌跡は、最終的には広大な文脈処理能力と、深く検証可能で柔軟な推論能力の両方が必要とされる方向へと収束していくことを示唆している。Anthropicの統合的な深さへの注力と、OpenAIのコンポーネント化された広範性/特化への注力という現在の戦略の違いは、この収束した目標に向けた異なる経路であり、それぞれが独自の中間的なトレードオフと発見を伴うものと捉えることができる。つまり、両社はLLMが広大な情報を処理し、かつそれについて深く信頼性をもって推論するという共通の目標に向かっているが、その達成手段が異なっている。Anthropicは推論の質と検証可能性を優先し、OpenAIは容量とアクセス効率、そして精密な制御を優先しているように見える。将来のブレークスルーは、これら両アプローチの長所、例えば、数百万トークンの文脈上で効率的に動作する、高度に解釈可能で憲法的に整合した推論などを統合することになるだろう。

引用文献

1. arxiv.org, 5月 31, 2025にアクセス、 https://arxiv.org/abs/2501.07458
2. Tracing the thoughts of a large language model \ Anthropic, 5 月 31, 2025にアクセス、 https://www.anthropic.com/research/tracing -thoughts-language-model
3. System Card: Claude Opus 4 & Claude Sonnet 4- Anthropic, 5 月 31, 2025にアクセス、 https://anthropic.com/model -card
4. Activating AI Safety Level 3 Protections - Anthropic, 5 月 31, 2025にアクセス、 https://www.anthropic.com/news/activating -asl3-protections
5. Claude 3 | Prompt Engineering Guide, 5月 31, 2025にアクセス、 https://www.promptingguide.ai/models/claude -3
6. GPT□4.1 and the Frontier of AI: Capabilities, Improvements, and ..., 5月 31, 2025にアクセス、 https://www.walturn.com/insights/gpt -4-1-and-the-frontier-of-ai-capabilities-improvements-and-comparison-to-claude-3-gemini-mistral-and-

llama
7. The Claude 3 Model Family: Opus, Sonnet, Haiku - Anthropic, 5 月 31, 2025 にアクセス、 https://www.anthropic.com/claude-3-model-card
8. Anthropic's Transparency Hub: Model Report, 5 月 31, 2025 にアクセス、 https://www.anthropic.com/transparency/model-report
9. Understanding Different Claude Models: A Guide to Anthropic's AI, 5 月 31, 2025 にアクセス、 https://teamai.com/blog/large-language-models-llms/understanding-different-claude-models/
10. Anthropic's Claude 3.7 Sonnet hybrid reasoning model is now ..., 5 月 31, 2025 にアクセス、 https://aws.amazon.com/blogs/aws/anthropics-claude-3-7-sonnet-the-first-hybrid-reasoning-model-is-now-available-in-amazon-bedrock/
11. Claude's extended thinking \ Anthropic, 5 月 31, 2025 にアクセス、 https://www.anthropic.com/news/visible-extended-thinking
12. Claude 3.7 Sonnet and Claude Code - Anthropic, 5 月 31, 2025 にアクセス、 https://www.anthropic.com/news/claude-3-7-sonnet
13. arxiv.org, 5 月 31, 2025 にアクセス、 https://arxiv.org/abs/2410.21276
14. NeedleBench: Can LLMs Do Retrieval and Reasoning in Information-Dense Context? - arXiv, 5 月 31, 2025 にアクセス、 https://arxiv.org/html/2407.11963v2
15. Claude vs GPT 4: Key Differences Compared - Undetectable AI, 5 月 31, 2025 にアクセス、 https://undetectable.ai/blog/claude-vs-gpt-4/
16. GPT-4.1 and o4-mini: Is OpenAI Overselling Long-Context? - Zep, 5 月 31, 2025 にアクセス、 https://blog.getzep.com/gpt-4-1-and-o4-mini-is-openai-overselling-long-context/
17. Introducing GPT-4.1 in the API | OpenAI, 5 月 31, 2025 にアクセス、 https://openai.com/index/gpt-4-1/
18. The Future of AI is Here: Exploring GPT-4.1's Breakthroughs - MPG ..., 5 月 31, 2025 にアクセス、 https://mpgone.com/the-future-of-ai-is-here-exploring-gpt-4-1s-breakthroughs/
19. GPT-4.1 vs Claude 3 Opus - Detailed Performance &Feature Comparison - DocsBot AI, 5 月 31, 2025 にアクセス、 https://docsbot.ai/models/compare/gpt-4-1/claude-3-opus
20. GPT-4.1 Prompting Guide - OpenAI Cookbook, 5 月 31, 2025 にアクセス、 https://cookbook.openai.com/examples/gpt4-1_prompting_guide
21. Reasoning models - OpenAI API, 5 月 31, 2025 にアクセス、 https://platform.openai.com/docs/guides/reasoning
22. OpenAI o4-mini: everything you need to know about this model - Swiftask, 5 月 31, 2025 にアクセス、 https://www.swiftask.ai/blog/openai-o4-mini
23. Practical Guide for Model Selection for Real-World Use Cases - OpenAI Cookbook, 5 月 31, 2025 にアクセス、 https://cookbook.openai.com/examples/partners/model_selection_guide/model_selection_guide
24. Detect hallucinations for RAG-based systems | AWS Machine Learning Blog, 5 月

31, 2025 にアクセス、 https://aws.amazon.com/blogs/machine-learning/detect-hallucinations-for-rag-based-systems/
25. 100% Elimination of Hallucinations on RAGTruth for GPT-4 and GPT-3.5 Turbo - arXiv, 5 月 31, 2025 にアクセス、 https://arxiv.org/html/2412.05223v2
26. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness - PMC, 5 月 31, 2025 にアクセス、 https://pmc.ncbi.nlm.nih.gov/articles/PMC11971376/
27. AI Makerspace - AI Engineering Cheatsheet - GitHub Gist, 5 月 31, 2025 にアクセス、 https://gist.github.com/donbr/cda3560ddb410f4fe2b81f5c329a4891
28. Unlock the Power of Claude 3 Opus - Unimedia Technology, 5 月 31, 2025 にアクセス、 https://www.unimedia.tech/claude-3-opus-software-development/
29. Reducing LLM Hallucinations: A Developer's Guide - Zep, 5 月 31, 2025 にアクセス、 https://www.getzep.com/ai-agents/reducing-llm-hallucinations
30. Empowering LLMs to Think Deeper by Erasing Thoughts | Towards Data Science, 5 月 31, 2025 にアクセス、 https://towardsdatascience.com/empowering-llms-to-think-deeper-by-erasing-thoughts/
31. OpenAI's 'smartest' AI model was explicitly told to shut down — and it refused - Yahoo, 5 月 31, 2025 にアクセス、 https://www.yahoo.com/news/openais-smartest-ai-model-explicitly-164513876.html
32. Fine-tune large language models with reinforcement learning from human or AI feedback, 5 月 31, 2025 にアクセス、 https://aws.amazon.com/blogs/machine-learning/fine-tune-large-language-models-with-reinforcement-learning-from-human-or-ai-feedback/
33. Reinforcement Learning from AI Feedback (RLAIF) - VE3, 5 月 31, 2025 にアクセス、 https://www.ve3.global/reinforcement-learning-from-ai-feedback-rlaif/
34. : Evaluating Reward Models for Language Modeling - ACL Anthology, 5 月 31, 2025 にアクセス、 https://aclanthology.org/2025.findings-naacl.96.pdf
35. AI Hallucination Report 2025: Which AI Hallucinates the Most? - AllAboutAI.com, 5 月 31, 2025 にアクセス、 https://www.allaboutai.com/resources/ai-statistics/ai-hallucinations/
36. GPT-4.5 hallucination rate, in practice, is too high for reasonable use : r/singularity - Reddit, 5 月 31, 2025 にアクセス、 https://www.reddit.com/r/singularity/comments/1j06srh/gpt45_hallucination_rate_in_practice_is_too_high/
37. OpenAI's Approach to External Red Teaming for AI Models and Systems - arXiv, 5 月 31, 2025 にアクセス、 https://www.arxiv.org/abs/2503.16431
38. Addendum to GPT-4o System Card: Native image generation - OpenAI, 5 月 31, 2025 にアクセス、 https://cdn.openai.com/11998be9-5319-4302-bfbf-1167e093f1fb/Native_Image_Generation_System_Card.pdf
39. Daily Papers - Hugging Face, 5 月 31, 2025 にアクセス、 https://huggingface.co/papers?q=serving%20context%20length%20(C)